An Integrated Strategy for Data Mining Based on Identifying Important and Contradicting Variables for Breast Cancer Recurrence Research

Avijit Kumar Chaudhuri, Deepankar Sinha, Kousik Bhattacharya, Anirban Das

Abstract: Cancer leads to most deaths worldwide, and breast cancer is a leading disease that causes death among women. This disease is unique in the way that once treated, it can recur in some cases. Individuals are unable to identify their condition before it becomes dangerous. Extracting significant predictive features of breast cancer is an important and risky job for further study. Researchers have applied data mining techniques in medical science. Several authors suggest that a single method doesn’t resolve issues in diagnosing problems, and a hybrid model is desirable. In this paper, the authors propose an integrated approach to avoid Type 1 and 2 errors in predicting recurrence. They identify important and contradicting variables and consider them for inclusion and exclusion, respectively, to revise the dataset. The evaluation of findings of crucial methods, using original and revised datasets, widens the choice of identifying the technique with higher accuracy. The results show that the accuracy improves with the selection of variables restricted to the ones identified as relatively significant, and the dataset revised after elimination of contradicting variables.

Keywords: Data mining techniques, errors, integrated approach, under-estimation, recurrence, breast cancer

I. INTRODUCTION

Data mining techniques (DMT) have gained significance in the prediction of diseases. Palaniappan and Awang[1] suggested the development of an intelligent system to predict heart disease based on Decision Tree, Naïve Bayes, and Neural Network. Several authors have suggested the prediction of the disease using DMT. Cancer is a terminal disease and needs attention as it is the cause of maximum health-related deaths worldwide. Breast cancer affects the female population and causes death[2]. In a country such as India, over the last five years, 79 thousand women died from breast cancer per annum[3]. This disease is the second (after cervical cancer) most cause of deaths amongst women population. World Health Organisation (WHO) reports that deaths due to breast cancer take place due to delay and imperfect detection of the disease and medication. A unique fact with this disease is about its recurrence, even after positive primary therapy of cancer after its initial discovery. Gauthier et al.[4] identified four significant parameters: age, breast density, the number affected first-degree relatives and prone to breast biopsy. The authors used the data mining process to develop a dependable instrument for the assessment of the primary risk of breast cancer. In this paper, the authors investigate the parameters responsible for the recurrence of breast cancer. Some authors[5][6][7] have based their conclusions on clinical trials, measurements, and results of treatment of patients using recurrence score or similar observations to identify the likelihood of recurrence. For example, some authors have concluded that chemotherapy to all patients after surgery would cause overtreatment to around 85%. The researchers have suggested a multi-step approach to understanding the likelihood of treatment through an understanding of the genetic composition[5]. The authors applied the recurrence score algorithm on data related to the genes and application of clinical trial to control groups. Some authors[6] predicted that patients five years post-surgery for breast cancer appear to have a very slowly decreasing hazard of recurrence. The authors derived the mean risk of recurrence between 5 to 12 years as 4.3% per year, and this group of patients is subject for trials evaluating the influence of drugs such as cytostatic or differentiating agents. Some authors have found that recurrence of this disease is higher in of estrogen receptor (ER) – patients, regardless of the size of the tumor, several involved axillary nodes, age, or adjuvant therapy. The researchers in this study observed that there is difficulty in associating, clustering, and classification of the outcomes of different treatments. For example, an association rule of the type – “the probability of recurrence in patients with ER, post-surgery within six years with a distinct genetic expression or multigene-expression assay” could be useful to associate different perspectives of recurrence. Data mining techniques enable derivation of such association rules, clustering, and classifications.

The world has witnessed regular advancement in research related to cancer[8]. The researchers have experimented with various approaches such as predicting the disease, shielding and post identification measures such as surgery, chemotherapy, and other means. These approaches have led to the generation of varied and voluminous data. Typical prediction approaches are likely to get biased by the choice of techniques by the researchers. Data mining techniques (DMT) have been widely used to gain valuable insights and patterns from large data sets [9]. Researchers have applied DMT in various fields of social, biological physical and medical science, and technology. Study of the application of DMT in health care reveals that it has widespread applications, but there is no single method to resolve issues in diagnosing medical problems and a hybrid model is desirable to address the concerns[10].

Revised Manuscript Received on February 01, 2020.
Avijit Kumar Chaudhuri, Assistant Professor, Academic In-Charge, Sikkim Manipal University Learning Centre, Kolkata
Deepankar Sinha, PhD, Indian Institute of Technology (IIT), Kharagpur.
Kousik Bhattacharya, Assistant Registrar(Exam.) DDE, Rabindra Bharati University, Kolkata, West Bengal.
Anirban Das, HOD, Computer Science & Engineering in Amity University Kolkata.

International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-6, March 2020

Published By: Blue Eyes Intelligence Engineering & Sciences Publication

Retrieval Number: F7567038620/2020@BEIESP
DOI:10.35940/ijrte.F7567.038620

1096
An Integrated Strategy for Data Mining Based on Identifying Important and Contradicting Variables for Breast Cancer Recurrence Research

The over or underfitting of data may lead to wrong interpretation and predictions. These results may have an irreversible and significant impact on matters related to health. The effect may be adverse due to both Type 1 and 2 errors. Studies have shown that different data mining techniques have yielded a varying degree of accuracy in interpreting health care data. Some of the authors have compared the results and suggested the one whose accuracy of prediction was highest. The accuracy level of say 75% from a particular method appears quite impressive than a way predicting 56% or so. But the question remains that can the techniques be combined to obtain a better prediction level? In this research work, the authors attempted to integrate the DMT based on the significance of variables (or in words, variable importance). The three widely used approaches, namely, Decision-Tree, Logistics-Regression, and Discriminant-Analysis, have been experimented with a different choice of variables. The Logistics-Regression and Discriminant-Analysis have been carried out, separately, considering the entire variable set and with the significant variables identified through Random Forest and Decision Tree Analysis. The authors have correlated the results from these analyses with the findings of the integrated investigations. In this paper, the authors tested the proposed framework on databases relating to critical disease such as breast cancer. Thus, this paper not only contributes to finding a way to enhance prediction levels using DMT but also tries to contribute to society at large. In this paper, an integrated approach combining unsupervised and supervised data mining techniques has been proposed to avoid underestimation of the prevalence of the disease. This paper has five sections. The next part gives details of findings from research works done so far; section 3 briefs about the use of hybrid data mining techniques and its relevance in the study of breast cancer; section 4 enumerates the findings of the analysis using data mining approaches, and the last section concludes the research with a proposal for the further scope of work.

II. RELEVANT LITERATURE

Data mining approaches have found a place in medical and health-related research. The researches have applied a varied set of algorithm related to classification, clustering, and prediction. Different authors have recommended different methods based on the level of accuracy determined by them. These applications are observed in studies on breast cancer as well.

A study by Delen et al.[11] on a dataset of around 200, 000 sufferers showed 93% accuracy in forecasting survival. The author categorized the dataset as survived and not survived. Some authors, namely Tan AC [12], and JinyaLiHuiqing Liu, and Liu Ya-Qin[13] compared the performance of decision tree (DT) and bagging decision tree using C4.5 and C5, respectively, on the prediction of breast cancer survivability. Some authors[13],[14] proposed combining of techniques. Don-Sheng Cao[13] combined decision tree-based ensemble method with feature selection method backward elimination strategy and bagging, while My ChauTu [14] proposed the integration of bagging, C4.5 algorithm, and Naïve Bayes algorithm in diagnosing cardiovascular disease. Tsirogiannis[15] concluded that bagging results showed higher accuracy compared to methods without bagging. Panwen[16] used the DT algorithm C4.5 with bagging to determine high-frequency electrograph that is abnormal from the data collected from ECG reports of patients. Amalgamated methods such as Tree Bagging and weighted Clustering (TBWC) are also proposed by authors[17] for diagnostic purpose. Chaurasia and Pal [18] showed the superiority of bagging approach over Bayesian classification and J48. Some authors[19] applied different forms of neural network algorithms for the detection of breast cancer. In one of these studies[20], the radial-Basis-Function (RBF) network showed better accuracy than Multilayer-Perceptron (MLP). Some authors extended the use of neural networks and used the Adaptive Neuro-Fuzzy Inference System (ANFIS) [21] and Adaptive Resonance Neural Networks (ARNN)[22]. Authors also applied a computer-aided diagnostic system for classification of breast cancer [23].

The review of the literature shows that no single technique has been consistent in prediction-accuracy, and choice of models are specific to the researchers. However, the study shows that certain DMT has proved useful and superior in the performance of prediction. In this paper, the authors aim at choosing the effective techniques and propose integration and hybridization to answer the research question – “What is the likelihood of recurrence of breast cancer in patients who have undergone surgery or treatment after initial detection of the disease” with greater accuracy and with the availability of multi-dimensional data.

The methodology revolves around the concept of significant variables, contracting variables, and discriminating functions. Random Forest (RF) is used to identify the significant and contradicting variables. The authors, in this paper, applied classification and clustering techniques over three datasets, viz., original dataset, a dataset with important variables only, original dataset without the contradicting variables. The authors compared results obtained from supervised and unsupervised learning approaches applied over three datasets separately. The intersection and union of significant variable sets derived from the different hybridizations lead to identification cause of recurrence and thus would aid the doctors in planning for treatment of breast cancer post-surgery.

III. CHOICE OF MODELS

This paper proposes to use data mining models. Namely, Random Forest, K means clustering, Decision trees, Logistics Regression, and Discriminant Analysis for analysis of the cause of the disease and prediction of the disease.

A. Random forest (RF)

RF, a supervised machine learning algorithm is a blended arrangement technique in light of statistical learning hypothesis (Breiman, 2001). It is a classifier technique to reduce errors, considered the unweighted majority of class votes. A large number of the tree makes RF from the selected samples.
Classification and regression tree (CART) is the customarily utilized tree base classifier for RF. Finally, RF joins all classification trees. The final classification results obtained by voting on each class, followed by picking the victor class as far as the number of votes to it. The RF execution measure by a metric got out of bag error figured as the normal of the average of the rate of error in each weak learner. RF investigates every individual tree. A bagging procedure is used where for a given set of training data N, n random samples with repetition (Bootstrap) are considered; subsequently, the tree for each random small is developed. M input variables are determined for each node (of the tree), and out of these “M,” variables, m variables are selected for each node. The most important randomly chosen variable serves as a node. The value of m does not change in the process, and each tree is finally developed to its maximum expansion[24]. The accuracy of RF suffers on the way the right kind of randomness is introduced. Generation of the tree with minimal depth has an advantage as it is independent on the way prediction error is measured[25].

B. K-means

The primary influence of this algorithm is its clarity and speed, which allows it to run on large datasets. K-means clustering may be computationally faster with a large number of variables than hierarchical clustering, i.e., if K is small. K-Means may yield compact clusters than hierarchical clustering, especially if the groups are globular.

C. Decision tree (DT)

DT enables supervised classification and allows processing of erroneous datasets and missing values. This method concludes the value of dependant aspect (attribute) given the values of independent (input) attributes [26]. This method does not require any form of normalization or scaling of variables set the relationship between dependent and independent attributes. The results obtained from decision trees are not affected by outliers as the classification takes place based on the proportion of samples with split ranges and not on absolute values. This method does not demand linearity in relationships between the dependant and independent attribute. Decision trees explain the results well and are very intuitive.

D. Logistics-Regression (LR)

LR allows one to analyze the effect of several explanatory variables on a response variable simultaneously. The use of LR, for categorical factors, in the field of medical research has increased since the 1990s[27]. Logistic regression is an efficient and powerful technique to estimate the contribution of a group of variables on a binary closure by evaluating each separate variable’s sole exclusive contribution. Isolated segments of a linear regression challenge in the logit scale can be used in logistic regression to cyclically recognize the essential linear combination of variables with the substantial prospect of identifying the measured output. Selection of self-independent variables, supplying proper support, and revealing an appropriate model development strategy are the essential study for conduction of logistic regression. For discrete variable selection, one should be directed by factors such as accepted theory, previous empirical inquiries, clinical considerations, and univariate statistical analyses.

E. Discriminant Analysis (DA)

DA serves as a predicting technique across different groups with continuous variable measurements. It is a parametric technique that discriminates two or more groups of seer variables and provides the weightings of quantitative and seer variables leading to the prediction of the disease[28]. DA is a linear combination of significant weighted variables. Thus this approach provides us with significant variables and their extent of influence in prediction. This method applies well to mixed continuous and categorical variables[29],[30].

IV. DATA SETS

The authors, in this paper, used data on breast cancer available in the UCI database. The data file posses 286 records with ten attributes per record.

Table- I: lists the attributes.

| Attribute | Description |
|-----------|-------------|
| Age       | Age of patient when diagnosed |
| Menopause | Labels the diagnostic status of patient’s menopause |
| Tumor size| Size of tumor in mm at the time of diagnosis |
| irr-nodes | Irregular lymph nodes, values ranges from 0 to 29 shows the presence of the disease |
| node-caps | Node presence and penetration of nodes in any of the breasts |
| deg-nodig | Represents the stage or grade of breast cancer. Value ranges from 1 to 3 |
| Breast    | Either right breast or left breast or it may be both breast |
| Breast-quad| Indicates segment of the breast e.g. right-low, left-low, right-up, left-up or central |
| Isolate   | Radiation (yes-no) history of a patient suffering from breast cancer |
| Class     | The patients having classified under no-recurrence events or recurrence events based on the reappearance of the symptoms of breast cancer |

Table- II: Breast Cancer Data set (UCI, 1988)

| Sl. No. | Attributes | Values |
|---------|------------|--------|
| 1       | Age        | 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99 |
| 2       | Menopause  | 1=not, 0=yes |
| 3       | Tumor size | 0.4-5.9, 6-10, 11-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49 |
| 4       | Irr-nodes  | 32, 33-35, 36-39 |
| 5       | Node-caps  | yes, no |
| 6       | Deg-nodig  | 1, 2, 3 |
| 7       | Breast     | left, right |
| 8       | Breast-quad| left-up, left-low, right-up, right-low, central |
| 9       | Isolate    | yes, no |
| 10      | Class      | no-recurrence-events, recurrence-events |

V. VARIABLE CLUSTERING AND IMPORTANCE

The author used Random Forest technique for identifying the clusters, importance of variables and contradicting variables. The original data set were subjected to K-means clustering, Decision-tree (DT) analysis, Logistics-regression (LR), and Discriminant Analysis (DA). These techniques were then used to predict the outcomes of the methods over the revised
A. Random Forest (RF) Analysis

Random forest analysis of the data on disease dataset shows that the tumor-size, breast-quad, deg-malig, breast, and age are relatively important variables. The variables, namely inv_nodes and menopause, contradict the prediction.

B. K – means Clustering

Analysis of the disease data set: K – means analysis of the original cancer data set (Table-III) show that significant variables include age, tumor-size, node-caps, and breast-quad.

Table-III: Final Cluster Centers

| Cluster | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|
| age     | 75 | 53 | 38 | 45 | 75 | 59 | 45 | 44 | 55 | 60 | 35 | 45 | 45 | 34 |    |
| tumor-size | 2 | 2 | 20 | 37 | 62 | 19 | 30 | 47 | 68 | 34 | 2 | 22 | 21 | 17 | 27 |
| inv_nodes | 1 | 1 | 1 | 16 | 1 | 1 | 1 | 1 | 1 | 16 | 1 | 1 | 15 | 1 | 16 |
| node-caps | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| deg-malig | 1 | 3 | 2 | 3 | 1 | 2 | 2 | 2 | 2 | 3 | 2 | 3 | 2 | 3 | 2 |
| breast | 1 | 1 | 3 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| breast-quad | 3 | 4 | 2 | 2 | 1 | 2 | 2 | 2 | 1 | 2 | 5 | 1 | 2 | 1 | 2 |
| irradiat | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| Clas | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |

The cluster 15 has the highest number of cases (10), which includes the lowest age category of 33.35 years, with comparatively larger tumor size (37). K-means clustering of data set comprising the variables found significant from RF analysis shows that all these variables barring breast are found significant as well.

Table- IV: ANOVA

| Cluster | Error Mean Square | Error DF | F | Sig. |
|---------|------------------|----------|---|------|
| age     | 1105.577         | 14       | 18.85 | 58.651 | 0   |
| tumor-size | 1246.135         | 14       | 11.982 | 103.999 | 0   |
| inv_nodes | 132.686          | 14       | 0     | 175   |     |

Table-V: K MEANS on prominent variables from RF(Final Cluster Centers)

| Cluster | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|
| tum size | 2 | 2 | 3 | 2 | 4 | 1 | 3 | 5 | 4 | 4 | 4 | 2 | 2 | 2 | 3 |
| or size | 0 | 0 | 3 | 2 | 9 | 5 | 2 | 5 | 7 | 2 | 1 | 2 | 1 | 6 |
| bre size | 3 | 4 | 2 | 2 | 3 | 2 | 2 | 3 | 1 | 2 | 5 | 2 | 2 | 2 | 2 |
| ast size | 1 | 1 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| deg | mal | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ig | bre | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| ast | 7 | 5 | 6 | 5 | 7 | 6 | 5 | 5 | 4 | 6 | 3 | 4 | 3 | 4 | 3 |
| age | 4. | 4. | 4. | 4. | 4. | 4. | 4. | 4. | 4. | 4. | 4. | 4. | 4. | 4. | 4. |
| Clas | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

Table-VI: ANOVA

| Cluster | Error Mean Square | Error DF | F | Sig. |
|---------|------------------|----------|---|------|
| tumor-size | 1448.453         | 14       | 12.941 | 239 | 111.927 | .000 |
| breast-quad | 3.220          | 14       | 1.528 | 239 | 2.106 | .012 |
| deg-malig | .963            | 14       | 0.508 | 239 | 1.895 | .028 |
| breast | .327             | 14       | 0.246 | 239 | 1.329 | .191 |
| age | 1806.817          | 14       | 1.558 | 239 | 1159.956 | .000 |
| Clas | .142             | 14       | 227   | 239 | .625 | .843 |

Table-VII: Number of Cases in each Cluster

| Cluster | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---------|---|---|---|---|---|---|---|---|
| 1 | 1.000 |
| 2 | 3.000 |
| 3 | 23.000 |
| 4 | 45.000 |
| 5 | 2.000 |
| 6 | 19.000 |
| 7 | 35.000 |
K-means clustering without contradicting variables (inv_nodes and menopause)

This approach shows that age, tumor size, and breast-quad are significant variables. The variable – node_cap gets eliminated.

**Table- VIII: Final Cluster Centers**

| Cluster | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|
| age     | 7 | 5 | 6 | 6 | 7 | 7 | 5 | 4 | 6 | 3  | 4  | 3  | 4  | 3  | 4  |
| tumor-size | 4 | 4 | 4 | 0 | 4 | 4 | 4 | 4 | 4 | 9  | 4  | 4  | 4  | 3  | 3  |
| nod- caps | 5 | 5 | 8 | 5 | 5 | 5 | 5 | 5 | 5 | 5  | 5  | 5  | 5  | 5  | 8  |
| deg- malig | 2 | 2 | 3 | 1 | 4 | 2 | 2 | 4 | 4 | 4  | 2  | 2  | 3  | 2  | 3  |
| breast | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0  | 0  | 0  | 0  | 0  | 0  |

**Table- IX: ANOVA**

| Cluster | Error | F | Sig. |
|---------|-------|---|------|
| Mean Square | df | Mean Square | Df | Sig. |
| age     | 1774.625 | 14 | 3.443 | 239 | .515.376 | .000 |
| tumor-size | 1445.800 | 14 | 13.097 | 239 | 110.396 | .000 |
| node-caps | .149 | 14 | .172 | 239 | .870 | .592 |
| deg-malig | .799 | 14 | .518 | 239 | 1.542 | .097 |

The largest cluster has 15 cases that are higher than the previous most populated cluster. This cluster includes the lowest age category of 33.5 years, with tumor size 36. The degree of malignancy is two in both cases.

**C. Decision Tree**

Figure 2 shows the outcome of the Decision tree analysis carried out on the original dataset.

The study shows that deg-malig is the primary cause followed by node-caps. For node-caps = 1.0 with deg-malig = 3, 76.7% of patients suffer from recurrence of cancer.
Table XI: shows the accuracy level of Decision-tree analysis.

| Classification | Observed | Predicted | Percent Correct |
|----------------|----------|-----------|-----------------|
|                | 0        | 1         | 27.1%           |
| Overall Percentage | 89.5% | 10.5% | 75.9% |

Growing Method: CHAID
Dependent Variable: Class(In numeric)

The decision tree with all variables predicts the accuracy of 75.9%.

Decision Tree Analysis considering relatively important variables (determined from RF Analysis)

Decision tree analysis on variables identified as relatively significant using RF analysis shows that deg-malig is the only critical factor relating to the disease (Figure 3). The overall prediction with this approach reduced to 72% accuracy (Table XI) from 75.9% when predicted with all variables using the decision tree. However, the prediction percentage of recurrence of breast cancer increased from 27.1% to 52.9%, hence type 1 error gets reduced with the integration of results from RF and DT analysis. While the type 2 error increased by 16.5%. The prediction of a non-recurrence of breast cancer reduced to 80% from the previous prediction of 96.5%.

Decision Tree Analysis after removal of contradicting variables from the original data set

The results from this approach show no change compared with the analysis of the original dataset.

Fig. 3. Decision Tree Analysis after removal of contradicting variables from the original data set

This tree indicates deg-malig as the only factor for the detection of breast cancer recurrence, but the accuracy of predicting recurrence is higher (52.9%) compared to the earlier DT analysis.

Table XII: shows the accuracy level of RF-based Decision-tree analysis

| Classification | Observed | Predicted | Percent Correct |
|----------------|----------|-----------|-----------------|
|                | 0        | 1         | 27.1%           |
| Overall Percentage | 89.5% | 10.5% | 75.9% |

Growing Method: CHAID
Dependent Variable: Class(In numeric)

D. Logistics Regression

The Logistics-regression on original dataset shows that degomalig and the ninth dummy for tumor size are relatively significant.

Table XIII: Variables in the Equation

| B     | S.E. | Wald df | Sig | Exp(B) 95% CI/EXP(B) |
|-------|------|---------|-----|---------------------|
| Age   |      |         |     |                     |                     |
| age1  |      |         |     |                     |                     |
| age2  |      |         |     |                     |                     |
| age3  |      |         |     |                     |                     |
| age4  |      |         |     |                     |                     |
| age5  |      |         |     |                     |                     |
| Menopause |      |         |     |                     |                     |
| Ste   |      |         |     |                     |                     |
| p     |      |         |     |                     |                     |
| l*    |      |         |     |                     |                     |
| Menopausal |      |         |     |                     |                     |
| et(1) |      |         |     |                     |                     |
| et(2) |      |         |     |                     |                     |
| Tumor-size |      |         |     |                     |                     |
| tumorsize(1) |      |         |     |                     |                     |
| tumorsize(2) |      |         |     |                     |                     |
| tumorsize(3) |      |         |     |                     |                     |
| tumorsize(4) |      |         |     |                     |                     |
| tumorsize(5) |      |         |     |                     |                     |

Published By: Blue Eyes Intelligence Engineering & Sciences Publication
The results showed an overall accuracy of 76.6% (Table X). The prediction of recurrence was low (41.2%) compared that of non-recurrence (91.5%).

### Table- XIV: Classification Tablea

| Predicted Class(In numeric) | Percentage Correct |
|-----------------------------|--------------------|
| 0                           | 184                |
| 1                           | 17                 |
| Overall Percentage          | 91.5               |

#### Variable(s) entered on step 1: age, menopause, tumorsize, invnodes, nodecaps, degmalig, breast, breastquad, irradiat.

#### Logistical Regression Analysis of Disease Dataset considering Relatively Important Variables determined from RF Analysis

In this approach, as well, deg-malig and the ninth dummy for tumor size were found relatively significant; however, showed a decreased accuracy level of 73.8%. The prediction accuracy of the recurrence got further lowered. Thus the integration of RF and LR did not improve the results. Table- XIV shows the output of LR analysis.

### Table- XV: Variables in the Equation

| Variable | B         | S.E. | Wald | df | Sig. | Exp(B) | 95% C.I.for EXP(B) Lower | 95% C.I.for EXP(B) Upper |
|----------|-----------|------|------|----|------|--------|--------------------------|--------------------------|
| tumorsize | .958      | .83  | .36  | 1  | .417 | 2.72   | .06                      |                           |
| tumorsize | 19900.0   | .01  | .99  | 0  | .000 | .01    | .937                     |                           |
| tumorsize | 20.2      | .00  | .97  | 5  | .070 | 9.19   | .937                     |                           |
| tumorsize | 6         | .01  | .93  | 9  | .000 | .01    | .876                     |                           |
| tumorsize | 1.575     | .29  | .58  | 1  | .423 | 9.25   | .01                      |                           |
| Invnodes  | .861      | 3.5  | 73   | 9  | 2    | .490   | .01                      |                           |
| invnodes  | 1.22      | 1.436| 1.39 | 1  | .294 | 4.90   | .01                      |                           |
| invnodes  | .559      | .655 | .73  | 4  | .572 | 14.6   | .02                      |                           |
| invnodes  | 19.4      | 40192.9| 1.0 | .268880908.8| .00  | 56   |                           |                           |
| invnodes  | 10.69     | 0     | .00  | 0  | .08  | .935   | .03                      |                           |
| invnodes  | .629      | 1.467| .838 | 4  | .733 | 8.95   | .04                      |                           |
| invnodes  | .280      | 1.477| .856 | 6  | .756 | 13.6   | .01                      |                           |
| invnodes  | .177      | 1.543| .90  | 3  | .838 | 17.2   | .04                      |                           |
| nodecaps | .360      | .455 | .42  | 9  | 1.434| 3.30   | .58                      |                           |
| Degmalig | .655      | .232 | .00  | 1  | 7.9  | 7.07   | .12                      |                           |
| breast   | .354      | .331 | .28  | 9  | 1.425| 14.2   | .74                      |                           |
| Breastquad| 2.8      | 81   | .71  | 5  | 1.925| 2.01   | 1.2                      |                           |
| breastquad| 1.81     | 142  | .25  | 8  | 1.925| 1.2    | 1.93                     |                           |
| breastquad| 1.3       | 3.56 | .95  | 7  | 1.925| 1.2    | 1.93                     |                           |
| breastquad| .930      | 1.42 | .25  | 4  | 1.925| 1.2    | 1.93                     |                           |
| breastquad| .497      | .34  | .9   | 6  | 1.925| 1.2    | 1.93                     |                           |
| breastquad| .486      | .33  | .9   | 6  | 1.925| 1.2    | 1.93                     |                           |
| breastquad| .749      | .25  | .18  | 6  | 1.925| 1.2    | 1.93                     |                           |
| breastquad| .984      | .25  | .18  | 6  | 1.925| 1.2    | 1.93                     |                           |
| Irradiat | .327      | 2.06 | .82  | 7  | 1.386| 2.80   | .68                      |                           |
| Constant | 2.205     |      | 1.923| 1  |      |        |                          |                           |

a. Variable(s) entered on step 1: age, menopause, tumorsize, invnodes, nodecaps, degmalig, breast, breastquad, irradiat.
An Integrated Strategy for Data Mining Based on Identifying Important and Contradicting Variables for Breast Cancer Recurrence Research

Table- XVI: Classification Table

| Predicted | Percentage Correct |
|-----------|--------------------|
| Class (In numeric) | Observed 0 | Observed 1 | Overall Percentage |

| Observed | Predicted | Percentage Correct |
|----------|-----------|--------------------|
| Class (In numeric) | Observed 0 | Observed 1 | Overall Percentage |

LR without contradicting variable

The Logistics-regression on the revised dataset, i.e., after removal of the contradicting variables, shows improvement in overall prediction. The accuracy of prediction of recurrence by methods I and III are equal, while the forecast of non-recurrence by approaches II and III are equal. Thus, the dataset excluding the contradicting variable appears more effective compared a dataset with important variables alone. Here, contradicting variables influences the accuracy significantly. Therefore this method enables a reduction in both Type I and 2 errors. The significant variables include tumor size and deg-malig.

Table- XVII: Classification Table

| Observed | Predicted | Percentage Correct |
|----------|-----------|--------------------|
| Class (In numeric) | Observed 0 | Observed 1 | Overall Percentage |

| Observed | Predicted | Percentage Correct |
|----------|-----------|--------------------|
| Class (In numeric) | Observed 0 | Observed 1 | Overall Percentage |

Logistics Regression Analysis on Relatively Important Variables determined from DT Analysis

The Logistics regression on variable found significant using DT shows that deg-malig and nodecaps, along with the constant, are statistically significant and predicts the recurrence. Equation 1 enables the prediction of the disease with these variables.

Table- XVIII: Variables in the Equation

| Variable | B | S.E. | Wald | df | Sig. | Exp(B) | 95% CI for EXP(B) |
|----------|---|------|------|----|------|--------|------------------|
| Step 1: degmalig | 0.807 | 0.09 | 14.92 | 1 | 0.00 | 2.40 | 1.488 3.373 |
| Nodecaps | 1.00 | 0.32 | 9.510 | 1 | 0.00 | 1.279 | 1.444 5.197 |
| Constant | 2.83 | 0.47 | 35.21 | 1 | 0.00 | 1.059 | |

Table- IXX: Classification Table

| Observed | Predicted | Percentage Correct |
|----------|-----------|--------------------|
| Class (In numeric) | Observed 0 | Observed 1 | Overall Percentage |

| Observed | Predicted | Percentage Correct |
|----------|-----------|--------------------|
| Class (In numeric) | Observed 0 | Observed 1 | Overall Percentage |

a. Variable(s) entered on step 1: tumor size, breast quad, deg-malig, breast, age.

Table- XVI shows the accuracy level of this integrated approach.
Thus LR on the data set without contradicting variables showed better accuracy compared to other approaches so far.

E. Discriminant Analysis

The Box's test of equality of covariance matrices is found significant. Hence this approach with the original dataset is not considered in this study.

Table- XX: Test Results

| Box's M | 19.057 |
| Approx. | 6.291 |
| F | 528955.537 |
| df1 | 3 |
| df2 | 528955.537 |
| Sig. | .000 |

DA on variables found significant from RF analysis

This approach is considered as Box's test of equality of covariance matrices not found significant. The results from DA on the revised disease data set show that eigenvalue is .073 (<1). Canonical correlation, r=0.261(<0.35) (Table- XXI). Thus, Function 1 explain the variation well with the low canonical correlation between the two canonical variables.

Table- XXI: Eigenvalues

| Function | Eigenvalue | % Variance | % Cumulative | Canonical Correlation |
| --- | --- | --- | --- | --- |
| 1 | 0.073 | 100.0 | 100.0 | 261 |

a. First 1 canonical discriminant functions were used in the analysis.

Table- XXII: Wilks’ Lambda

| Test Function(s) | Wilks’ Lambda | Chi-square | Df | Sig. |
| --- | --- | --- | --- | --- |
| 1 | .952 | 17.759 | 1 | .000 |

The centroid for the discriminant function – F calculated as ((-0.189 + 0.383)/2)=0.097 is the discriminating value.

Table- XXIII: Canonical Discriminant Function Coefficients

| Function | deg-malign |
| --- | --- |
| 1 | 1.416 |
| (Constant) | -2.988 |

Unstandardized coefficients

Only one variable found significant (Table- XXIII). Equation II provides the F function.

\[ F = -2.988 + 1.416 \times \text{deg-malign} \]  

Table- XXIV: gives the accuracy of the prediction.

Classification Table

| Class(in numeric) | Predicted Group Membership | Total |
| --- | --- | --- |

DA without contradicting variables

The Box’s test of equality of covariance matrices (Table- XXIV) shows that the test result is significant, and hence, the authors do not consider this approach for further study.

Table- XXIV: Test Results

| Box's M | 11.790 |
| Approx. | 3.890 |
| F | 673467.575 |
| df1 | 3 |
| df2 | 673467.575 |
| Sig. | .009 |

Tests null hypothesis of equal population covariance matrices.

VI. RESULTS AND DISCUSSIONS

The analysis using unsupervised and supervised classifiers, and prediction techniques with and without important and contradicting variables separately show higher accuracy of prediction of recurrence using Decision Tree analysis on a dataset with the relatively important variables determined using Random Forest. The dataset excluding the contradicting variables when used results remain unchanged. This approach has the lowest type 1 error; however, Decision Tree analysis on the original data set exhibits the lowest type 2 error, related to non-recurrence. That is, the overall performance of the decision tree was superior compared to the other methods. Figure 4 shows the comparison of performance of the different approaches.

This study highlights that inclusion and removal of contradicting variables does improve the accuracy results. For example, the accuracy of prediction, of recurrence as well as non-recurrence, using Logistics Regression improved on the inclusion of important variables and exclusion of contradicting variables.

The study found the degree of malignancy as a factor with the highest score followed by the tumor-size. Equation III gives the expression of this finding.

\[ \text{DTi} \cap \text{LRi} \cap \text{DAi} \cap \{\text{Degree of Malignancy}\} \]  

Where

\[ \text{DTi} \] represents Decision Tree analysis on original and revised data set;

\[ \text{LRi} \] represents Decision Tree analysis on original and revised...
data set

DA represents Decision Tree analysis on original and revised data set

I = I represents original data set; II represents revised data set comprising relatively important variables only; III represents revised data set comprising data of all variables except the contradicting variables

Interestingly, the results from K-means clustering show that age is a significant factor, not recommended by any other approaches. This Degree of malignancy features as significant variable across supervised as well as unsupervised methods when data set comprised important variables (obtained from RF analysis) only. Equation IV gives the expression of this finding

\[ SLij \cap ULij = \text{Degree of Malignancy} \]

Where

\[
SLij \text{ represents the supervised learning approach } j \text{ on an } ith \text{ data set }
\]

\[
ULij \text{ represents unsupervised learning approach } j \text{ on an } ith \text{ data set }
\]

\[
i = I \text{ represents the original data set; II represents revised data set comprising relatively important variables only; III represents revised data set comprising data of all variables except the contradicting variables.}
\]

![Fig. 4. Comparison of accuracy level and identification of significant variables](image)

The set of variables found significant from unsupervised and supervised techniques include minimum values of age, node cap, tumor size, breast-quad and deg of malignancy and is expressed, as shown in expression V.

\[ SLij \cup ULij = \text{age, node cap, tumor size, breast-quad and deg of malignancy} \]

The minimum values of age, tumor size, breast-quad, and deg of malignancy are 33.8, 20, 2, and two, respectively, that cause recurrence of breast cancer.

VI. CONCLUSION

Researchers have professed varying methods with varying degree of accuracy. The reasons, proposed by authors, for this variedness include difference in dataset, imbalance in dataset and capability of the classifiers. Each classifier has its own metrics and demerits, and none of the popular ones could establish complete superiority over all others. The authors in this paper, attempted to make use of the existing techniques and determine whether application of techniques on dataset under two conditions i.e., i. containing the significant variables alone; and dataset ii. containing the non-contradicting variables, could enhance the accuracy of the prediction. Logistics Regression analysis without contradicting variables showed greater accuracy levels in non-recurrence of the breast cancer. Thus the paper contributed in identifying the prediction with higher accuracy by ensemble of outcomes of individual approaches. The degree of malignancy was found to most significant variable, as cause of recurrence of breast cancer, as it appeared significant in all forms of analysis.

Thus the approach of integration of data mining approaches and use of relatively significant variables, and elimination of contradicting variables lead to better prediction.

REFERENCES

1. S. Palanippan, and R. Awang, “Intelligent heart disease prediction system using data mining techniques,” In 2008 IEEE/ACS international conference on computer systems and applications, IEEE, 2008, pp. 108-115.

2. V. B. Bangal, K. K. Shinde, S. P. Gavhane, and R. K. Singh, “Breast Carcinoma in women - A Rising threat,” International Journal of Biomedical And Advance Research, IJBAR, 2013, 04 (02).

3. GLOBOCAN 2012 resource at http://globocan.iarc.fr

4. E. Gauthier, L. Brisson, P. Lena, and S. Ragusa, “Breast cancer risk score: a data mining approach to improve readability,” In The International Conference on Data Mining, 2011, pp. 15-21, CSREA Press.

5. W. A. Knight, R. B. Livingston, E. J. Gregory, and W. L. McGuire, “Estrogen receptor as an independent prognostic factor for early recurrence in breast cancer,” Cancer Research, 37(12), 1977, pp. 4669-4671.

6. S. Paik, S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin, and W. Hiller, “A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer,” New England Journal of Medicine, 351(27), 2004, pp. 2817-2826.

7. T. Sapnher, D. C. Tormey, and R. Gray, “Annual hazard rates of recurrence for breast cancer after primary therapy,” Journal of Clinical Oncology, 14(10), 1996, pp. 2738-2746.

8. D. Hanahan, R.A. Weinberg, “Hallmarks of cancer: the next generation,” Cell2011, 144, 2011, pp. 646-74.

9. S. H. Liao, P. H. Chu, and P. Y. Hsiao, “Data mining techniques and applications–A decade review from 2000 to 2011,” Expert systems with applications, 39(12), 2012, pp. 11303-11311.

10. J. N. Rashid, A. N. A, and W. Husain, “Data mining in healthcare-a review,” Procedia Computer Science, 72, 2015, pp. 306-313.

11. D. Delen, G. Walker, and A. Kadam, “Predicting breast cancer survivability: a comparison of three data mining methods,” Artificial Intelligence in Medicine, 34(2), 2005, pp. 113-27.

12. AC. Tan, D. Gilbert, “Ensemble machine learning on gene expression data for cancer classification,” Appl Bioinformatics, (3 Suppl), 2003, pp. 575-83.

13. L. Ya-Qin, W. Cheng, Z. Lu,”Decision Tree Based Predictive Models for Breast Cancer Survivability on Imbalanced Data,” Paper Presented at the 3rd International Conference on Bioinformatics and Biomedical Engineering, 2009.

14. T. My, Chau, S. Dongil, and S. Dongkoo, “A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms,” Paper Presented at the Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, 2009.

15. G. L. Tsirogiannis, D. Frossyniotis, J. Stoisits, S. Golemati, Staflidopatis, and K. S. Nikita, “Classification of Medical Data with a Robust Multi-Level Combination scheme,” Paper Presented at the IEEE international joint Conference on Neural Networks, 2004.

16. W. Pan,”Application of decision tree to identify a abnormal high frequency electrocardiograph,” China National Knowledge Infrastructure Journal, 2000.

17. C. Kaewchaimong, N. Vongsuchato, and A.Srisawat, “A Combination of Decision Tree...
Learning and Clustering for Data Classification,” Paper Presented at the Eighth International Joint Conference on Computer Science and Software Engineering (JCSEE), 2011.

18. V. Chaurasia, S. Pal, “Data Mining Approach to Detect Heart Diseases,” International Journal of Advanced Computer Science and Information Technology (IJACST), Vol. 2, No. 4, 2013, pp 56-66.

19. K. Usha. Rani, “Parallel Approach for Diagnosis of Breast Cancer using Neural Network Technique,” International Journal of Computer Applications International Journal of Computer Applications (0975 – 8877), Volume 10, No 3, 2010.

20. A. Raad, A. Kalakech, and M. Ayache; “Breast Cancer Classification Using Neural Network Approach: Mlp And Rbf,” Paper Presented at the 13th International Arab Conference on Information Technology ACIT’2012 Dec-10-13 ISSN:1812-0857, 2012.

21. F. Bekaddour, and A. C. Mohammed; “A Neuro-Fuzzy Inference Model For Breast Cancer Recognition,” International Journal of Computer Science & Information Technology (IJCSIT), Vol 4, No 5, 2012.

22. S. Narang, H. Verma, and U. Sachdev; “A Review of Breast Cancer Detection using ART Model of Neural Networks,” International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 10, ISSN: 2277 – 128X, 2012.

23. C. Wei-pin, L. Der-ming; “Comparison of three data mining techniques with genetic algorithm in the analysis of breast cancer data,” Journal of Telemedicine and Telecare 9, 2008.

24. L. Breiman, “Random forests,” Mach Learn, 45(1), 2001, pp. 5-32.

25. X. Chen, H. Ishwaran, “Random forests for genomic data analysis,” Genomics, 99(6), 2012, pp. 323-329.

26. M. Shouman, T. Turner, and R. Stocker, “Using data mining techniques in heart disease diagnosis and treatment”, 2012, pp. 173–177.

27. GH. Hall, AP. Round; “Logistic regression - explanation and use,” J R CollPhysicians land. 28, 1994, pp. 242-246.

28. D. Cramer, Advanced Quantitative Data Analysis, Open University Press, Buckingham, 2003.

29. W. J. Krzanowski, “Multiple discriminant analysis in the presence of mixed continuous and categorical data,” In Statistical Methods of Discrimination and Classification, Pergamon, 1986, pp. 179-185.

30. J. E. Overall, and J. A. Woodward, “Discriminant analysis with categorical data,” Applied Psychological Measurement, 1(3), 1977, pp. 371-384.

AUTHOR BIOGRAPHIES

Avijit Kumar Chaudhuri received his B.Sc degree from Calcutta University, A.M.LE degree in computer engineering from The Institution of Engineers(India). He has completed M.Tech. and MBA degrees in computer science and E-Business from the Sam Higginbottom University of Agriculture, Technology & Sciences – SHUATS, formerly known as AADU, India and Annamalai University, India in 2007 and 2014, respectively. Since July 2010, he has been with the Department of Computer Science and Engineering, Techno Engineering College Banpur, where he is an Assistant Professor. Prior to that, he was the Academic In-Charge, Sikkim Manipal University Learning Centre, Kolkata. Mr. Chaudhuri started his career as a Centre Academic Head in Aptech Computer Education and worked as Visiting Faculty in University Learning Centres of Vidyasagar University and IGNOU. His current research interests include Artificial Intelligence (AI), Data Mining, Machine Learning and Hybrid Learning.

Deepankar Sinha is PhD from Indian Institute of Technology (IIT), Kharagpur, and possess MBA, Diploma in Operations Research (ORSI) and Information technology (ACI). He has 20 years’ experience in port and shipping industry, and 124-years of experience in teaching, training and consultancy in the field of global logistics, international trade operation, system engineering and data analytics. He has been awarded 1 Patent & filed another, received 4 international & national awards; his 2 scholars have been awarded PhD Degree; he has authored 1 Book on Global Supply chains and Multi-modal Logistics published with IGI Global, USA. He has papers published in international &national journals. He is associated with advisory services to national and corporate bodies. His knowledge of research methodologies include: Operations Research (OR), Systems Dynamics, Data Envelopment Analysis (DEA), Structural Equation Modelling (SEM), Data Mining Techniques and similar.

Kousik Bhattacharya received B. Sc. degree from Vidyasagar University and completed MCA, ADCA, PGDCA from IGNOU. Now, he is working as Assistant Registrar(Exam.) DDE, Rabindra Bharati University, Kolkata, West Bengal and also be attached with IGNOU, as a part time Academic Counsellor from 2010.

Anirban Das is associated with University of Engineering & Management, Kolkata as a Full Professor in Computer Science & Engineering. Prior to that, he was the founder HOD, Computer Science & Engineering in Amity University Kolkata. Dr. Das started his career as a software Engineer in HCL Technologies Ltd and worked in many high end projects with Station Casinos- Las Vegas, Indian Air Force, CITI BANK, USA, Dai-ichi, India in specially Microsoft Technologies domain. He was invited by Her Excellency, Smt. Pratibha Patil, President of India with a one to one meeting on a report that was accepted by Rastrapati Bhawan, Delhi. He has completed Doctoral degree from NIT Durgapur. Dr. Das is a Visiting Scientist of University of Malaya, Malaysia. He is the honorary Vice President of Scientific and Technical Research Association, Eurasia Research, USA, Fellow of International Society for Research and Development, UK, Fellow of Indian Cloud Computing Association, India. He is Visiting Faculty at Jadavpur University Kolkata and Central University of Jharkhand, Ranchi. He has authored 9 books, 6 patents and more than 35 research publications mostly in journals and conferences of international repute.

He is the invitee as delegate from academics of several national bodies like, DRDO, Confederation of Indian Industry, NASSCOM, FICCI and EDCN. He was awarded as University Best Teacher Award 2018 by University of Engineering & Management, Kolkata, National IT and Cyber Security Research Association- 2018, Professional Membership Award by Blockchain Skill Development Association- 2018, Professional Membership Award by Data Analytics Skill Association- 2018, Swami Vivekananda Yuba Summan- 2015. He is now associated with research in Blockchain, IOT, Multi criteria Decision Making.