International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *15 August 2023* ***Revised:*** *30 September 2023* ***Accepted:*** *18 October 2023*

Data Mining Technique for Breast Cancer Prediction using Fuzzy Theory

Ms. Parul Bhatnagar[1], Dr. Bhupendra Kumar[2]
[1]Research Scholar, IIMT University, Meerut
Parul0710@gmail.com
[2]Professor, IIMT University, Meerut
singhbhupender231@gmail.com

Abstract— In order to find a reliable approach of breast cancer prediction, Data mining methods are used in the studies provided in this article. This study compares multiple patient clinical data in order to find a reliable model that can predict the occurrence of breast cancer. In this article, the support vector machine (SVM), artificial neural network (ANN), naive bayes classifier, and AdaBoost tree are used as four data mining methods. Furthermore, since it has such a significant impact on the efficacy and efficiency of the learning process, feature space is extensively examined in this work. Combining PCA with other data mining algorithms that use a PCA-like technique to compress the feature space is recommended. This hybrid is intended to assess the effect of feature space reduction. Wisconsin Breast Cancer Database (1991) and Wisconsin Diagnostic Breast Cancer (1995) are two frequently used test data sets that are used to assess the effectiveness of these algorithms. To calculate each model's test error, the method of 10-fold cross-validation is used. The findings of this research show a thorough trade-off between these tactics and also provide a thorough assessment of the models. In practical applications, it is anticipated that feature identification results would help to avoid breast cancer for both doctors and patients.
Keywords- Breast cancer prediction, data mining, k-fold cross-validation

I. INTRODUCTION

The use of machine learning technologies in medical diagnostics is growing as classification and recognition algorithms become more useful in assisting medical professionals in making a breast cancer diagnosis. With this method, many breast cancer patients are spared from invasive surgical biopsies, pointless adjuvant therapies, and high medical costs. Predictive models created using data mining approaches that have sufficient accuracy may be used to assist doctors in making decisions when expert oncologists are not present. One of the most important components of the preventative efforts was identifying key elements that contributed to the occurrence of breast cancer. Geo-computation has made it easier to gather data and provide statistical reports, but sophisticated autonomous approaches for exploratory analysis have not yet gained widespread acceptance. Large amounts of information from many sources are needed before making any choice. Since there are certain uncertainties in the data, using traditional statistical approaches to analyze the data has an impact on the decision-making process. Additionally, it takes a long time to do it for any database. As a result, spatial statistical analysis is no longer appropriate or suited for environments with plenty of data. Recently, a lot of data on individuals and their illnesses has been gathered in medical databases. Depending on the patient, a same illness may present itself extremely differently and at varying intensities. Consequently, treating the imperfect knowledge and information is a common problem in the analysis of medical data. When the diagnosis of an illness includes several layers of uncertainty and imprecision, which is intrinsic to medicine, the present intelligent procedures are unable to draw conclusions. Through the potential for visualization, the concise and understandable prediction models might aid in improving human judgments. Decision trees and neural networks are examples of computational intelligence approaches that are useful for extracting rules from data and comprehending it.

The use of fuzzy logic in bioinformatics and medicine has been well-received. Using the k-means method, fuzzy analysis has been utilized as a data mining tool to create decision tree models. An efficient Decision Support System for binary classification may be hindered by a more realistic aim employing fuzzy logic due to the complexity of biological classification challenges. The most common data mining technique used in rule extraction is the association rule mining algorithm. By connecting the genes in transcriptional time and the temporal dependencies between the genes, temporal association rules play a significant role in biological processes. The occurrence-free survival of breast cancer patients was calculated using Kaplan-Meier analysis and the ID3 (Iterative Dichotomiser: a decision tree approach) methodology. Fuzzy logic generates responses that are more realistic by substituting a topical adjustment in the form of a "more or less" for the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *15 August 2023* ***Revised:*** *30 September 2023* ***Accepted:*** *18 October 2023*

binary "yes/no" and by incorporating linguistic complexity into the decision-making process.

A. Breast Cancer and Fuzzy Logic

Boolean logic is extended by fuzzy logic, which substitutes degrees of truth for binary truth values. At the University of California, Berkeley, Zadeh L. A. provided the first description of it. Very early diagnosis would be the application area for fuzzy logic, according to Zadeh L. A. According to Blechner M. D., fuzzy logic, which allows membership values between 0 and 1, may be able to depict biological image analysis data that is intrinsically noisy and imprecise more realistically. Depending on the patient, the environment, and the severity, there are various degrees of ambiguity involved in the diagnosis of breast cancer. The architecture of medical knowledge information systems allows for the use of fuzzy logic across many medical information sources to help improve the state of the public health. Numerous variables, including age, nutrition, marital status, stage, therapy, inheritance, environmental factors, etc. are linked to the breast cancer illness. There are varying degrees of correlation and ambiguity between the stage of the breast cancer and other variables. As a result, the emphasis of this section is on predicting a patient's vulnerability to developing breast cancer.

B. Breast Cancer

Breast cancer may arise in a variety of breast cell types as well as in fatty tissue and fibrous connective tissue. Breast cancer tumors tend to quickly worsen and lead to death, making it a particularly dangerous disease. Breast cancer is more common in women than in males, although men may get it too. Malignant tumors are dangerous, whereas benign ones pose no health risks. Possible risk factors for breast cancer include being older and having a family history of the disease.

C. Types and Stages of Breast Cancer

Contingent upon the stage and sort of the malignant growth, a few medicines are accessible for it. The order of bosom disease depends on the state of the cell surface receptors.
There are two distinct forms of tumors:

- Benign: This particular tumor type poses little threat to the human body and seldom results in fatalities. This tumor type grows slowly and often appears in one specific location.
- Malignant**:** Breast cancer is a more lethal form of this tumor kind that kills people. At the point when cells in the bosom tissue multiply inappropriately, the threatening cancer structures. among the essential types of bosom disease are:
    1. Ductal carcinoma in situ (DCIS):is treatable and the earliest sort of breast disease.
    2. Invasive Ductal Carcinoma (IDC): the most predominant sort of breast disease and it begins in the milk pipe.
    3. Invasive Lobular Carcinoma (ILC): start in one of the breast's lobules. It has the ability to spread rapidly to the body's lymph hubs and different areas.

D. Primary Breast Cancer Phases

The size of the growth and on the off chance that the illness has spread is portrayed by the breast disease stage. Among the primary stages of breast cancer is:

- Stage I: With regards to essential disease, the growth should be unobtrusive (two cm or less) and not have spread to the lymph hubs.
- Stage II: level IIA and Stage IIB are the two stages that make up this level.
- Stage IIA: Albeit the malignant growth is little — under two centimeters — it has advanced to the axillary lymph hubs, which are situated under the arm. Or on the other hand the malignant growth has gone to the lymph hubs underneath the arm yet is only two to five cm away.
- Stage IIB: The lymph nodes beneath the arm have been affected by the malignancy, which has a diameter of two to five centimeters. OR the disease has gone to the lymph hubs underneath the arm yet is less than five millimeters.
- 'Advanced Breast cancer' is the term used to describe stages III and IV.
- Stage III: The tumors have become larger. They can have expanded to nearby breast tissue and lymph nodes.
- Stage IV: Other bodily components, such the lungs and bones, have been affected by tumors.

E. Breast Cancer Treatment

Patients might seek a solitary treatment or a blend of treatments relying upon the age, kind, and phase of the patient's disease. The essential breast disease treatments are:

1. Surgery: For breast malignant growth, there are fundamentally two careful methodologies. Breast rationing a medical procedure, frequently known as a lumpectomy, is the primary sort of a medical procedure. The motivation behind medical procedure is to eliminate the carcinogenic part of the breast alongside some encompassing sound tissue. The subsequent system is a mastectomy, which includes eliminating the entire breast.
2. Radiotherapy: With gamma radiation, cancer cells are destroyed.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *15 August 2023* ***Revised:*** *30 September 2023* ***Accepted:*** *18 October 2023*

3. Chemotherapy: Cytotoxic medications may be used to kill malignant growth cells all through the body, including the breast.
4. Hormone therapy: It is many times utilized after a medical procedure to assist with bringing down the opportunity of disease repeat or treat malignant growth that has spread to other body regions. The common term is something like five years.
5. Biological therapy: Comprised of novel drugs capability uniquely in contrast to chemotherapy. It decreases the chance of breast malignant growth repeating.

F. Breast Cancer Prevalence and Survival Rates

Breast disease is the most successive malignant growth in ladies, with very nearly 1,000,000 females universally getting a conclusion every year. Moreover, the quantity of ladies who die from breast malignant growth is extensive; in 2011, 211,000 ladies in rich countries and 213,000 ladies in emerging nations died from the sickness. Sadly, more young females are getting new breast cancer diagnoses and dying from it.

Survival from breast cancer is commonly predicted using the '5-year survival rate' after diagnosis, which is the percentage of women who are still alive five years after the beginning of their treatment or diagnosis. Eighty-one percent of women diagnosed with breast cancer at an early stage have a good chance of surviving for five years after diagnosis. However, only 35% of women with advanced or late-stage breast cancer survive for five years after diagnosis.

II. REVIEW OF LITERATURE

Chaurasia et al. (2018) made models for anticipating harmless and forceful breast malignant growth. An information assortment on Wisconsin breast malignant growth was utilized. The dataset included 699 models, two gatherings (threatening and harmless), and nine clinical boundaries with number qualities, for example, cell size consistency. To make the informational index of 683 cases, the specialists erased the 16 examples with missing qualities. 458 (65.5%) were harmless, though 241 (34.5%) were dangerous. The Waikato Climate for Information Examination (WEKA) inspected the trial. The three most popular information mining calculations — Innocent Bayes, RBF Organization, and J48 — were utilized to make the expectation models. For performance comparison, the researchers evaluated the three prediction models' unbiased estimates using 10-fold cross-validation techniques. Based on the efficiency and precision of the methodologies, an assessment of the models' performance was offered. According to experimental findings, Naive Bayes model had the highest performance (97.36% accuracy in classification), followed by RBF Network (96.77%) and J48 (93.41% accuracy in classification). A sensitivity analysis, a specificity analysis, and a combined analysis of all three algorithms were also conducted to better understand the relative contributions of the independent components to survival prediction. The "Class" prognostic factor was the most important predictor, according to the sensitivity data.

Yue et al. (2018) examined studies that looked at how well machine learning (ML) techniques performed in finding and predicting breast cancer. Researchers focused mostly on studies that used artificial neural networks (ANNs), support vector machines (SVMs), decision trees (DTs), and k-nearest neighbors (k-NNs) methods. Use of the Wisconsin breast cancer registry was also included. The extraordinary capacity of machine learning approaches to increase categorization and prediction accuracy has been shown. The material offered by the researchers is organized in a simple and straightforward manner. This data was shown in a table with citations, algorithms, sample methods, and accuracy in categorization. Researchers using the Wisconsin Breast Cancer Database (WBCD) found that a number of algorithms achieved excellent levels of accuracy, but new, more accurate algorithms were still required. In order to create an intelligent Friedreich Ataxia (FRDA) healthcare system, the researchers plan to perform a detailed analysis of the Friedreich Ataxia (FRDA) dataset in the future.

Banu & Ponniah (2018) presented a classifier-based approach to predicting the development of breast cancer. Bayes classifiers such as Tree Expanded Gullible Bayes, Supported Increased Gullible Bayes, and Bayes Conviction Organization were examined to demonstrate how they might be used to provide the most organized and accurate results. The Wisconsin Indicative Breast Cancer dataset includes 569 cases and 32 characteristics, was utilized in this study. All classifiers are integrated with the Gradient Boosting (GB) approach to increase accuracy. The accuracy of the three classifiers was practically identical at 90.1% prior to using the GB method. However, GB improves the outcomes of all classifiers. The precision, particularity, and awareness forecasts were utilized to assess the classifiers' exhibition. The findings demonstrate unequivocally the advantages of using TAN in the categorization of breast cancer.

Akinsola et al. (2017) proposed a technique for predicting breast cancer status that may help clinicians in view of patient clinical information (classes included harmless and threatening growths). More than 1700 cases were remembered for the Home dataset from the Central Government Clinic in Lagos. Eleven rules were utilized, including cell size, cell shape, and expected class. The characterization of breast disease information was finished utilizing three regulated learning calculations. Utilizing the WEKA tool kit, C4.5, Multi-facet Perceptron (MLP), and Nave Bayes were inspected. In view of the precision of the forecasts and the timeframe it took to foster the model, the three algorithms' performance was assessed. The C4.5 had the greatest accuracy of 93.9%, took 0.28 seconds, and was the best

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *15 August 2023* ***Revised:*** *30 September 2023* ***Accepted:*** *18 October 2023*

model. The researchers recommended that the system include another feature to prevent patients who can use the system from doing so.

Mirajkar and Lakshmi (2017) Using data mining's Naive Bayes Classification technique, the kind of cancer was predicted. The proposed approach intended to forecast the risk of certain cancer types. The Nave Bayes algorithm was used to classify cancer symptoms in order to identify cancer risks, such as those for breast and ovarian cancer.

Oskouei et al. (2017) performed a thorough study of all research that used data mining methods for diagnosing, treating, and predicting breast cancer highlighting the primary issues with these findings' further research in this field. Investigated were 45 papers, which were grouped into four groups depending on their primary objectives. The accuracy of using different categorization systems to identify breast cancers was examined across 21 papers. A technique to distinguishing between benign and malignant breast cancers was put out in twelve papers. In one of the publications, breast cancer datasets were used to diagnose early-stage breast cancers using regression data mining techniques. Eleven distributions made a breast malignant growth expectation model (for early recognition or foreseeing breast disease endurance). As indicated by the discoveries of these examinations, most of the investigations analyzed the exactness pace of information mining draws near. Sadly, there isn't an innovation that can immediately distinguish breast disease or suggest the best game-plan for patients. The scientists encouraged trying to make a program that would utilize information mining strategies to consequently recognize breast disease and recommend the best game-plan.

A. Research Objectives

1. To evaluate the efficacy of several data mining techniques (SVM, ANN, naive Bayes, and AdaBoost) for predicting breast cancer.

2. To examine how feature space inspection and reduction methods, such PCA, affect the precision and effectiveness of breast cancer prediction models.

III. RESEARCH METHODOLOGY

The purpose of supervised learning is to construct a classification model from a given data set that comprises certain features and labeled classes. In supervised learning, the training data set and the testing data set are two crucial components. Using the training dataset, a prediction model is built, which also contains attribute and cluster values. To verify the model, test data is often randomly selected from the full database. In this study, SVM, ANN, Naive Bayes classifier, and AdaBoost tree are used for testing. These models were chosen for their performance and appeal in literature. Figure 1 depicts the methodology of this study.

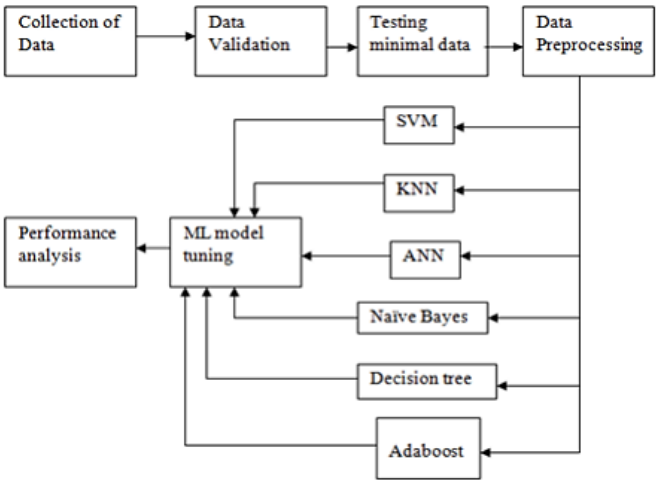

Figure 1: Modeling Method

1) Support Vector Machine

The support vector machine (SVM) was created by Vapnik in the 1970s. It works by optimizing the distance between the hyper planes of two clusters. SVM is widely used in several applications, such as number recognition, handwriting recognition, face detection, cancer classification, time series forecasting, etc. Binary classification is used to divide the training data into categories:

$$T = \{(x_i, y_i), i = 1, \dots, N, x_i \in R^M, y_i \in \{1, -1\}\},$$

Where xi is the class identifier and xi is an M-dimensional feature vector for the i-th instance. The SVM model is presented as the greatest margin of two classifier hyper planes:

$$\min \frac{1}{2} W^T W + C \sum_{i=1}^{N} \varepsilon_i \textbf{(1)}$$

$$s.t. \quad y_i(W^T x_i + b) \geq 1 - \varepsilon_i, i = 1, \dots, N, \textbf{(2)}$$

$$\varepsilon_i > 0, \quad i = 1, \dots, N, \textbf{(3)}$$

Maximizing margins and fulfilling margin requirements for each data points takes precedence is balanced using *C*, where *w* and *b* denote two parameters for separating hyper planes, *i* offers a slack variable to account for outliers and noise.

When the hyper plane is linear, Equation (1) simplifies to a quadratic optimization problem; however, when the hyper plane is nonlinear, constraint (2) also becomes nonlinear, increasing the complexity of Equation (1). Based on Lagrange duality and the KKT condition, it is possible to construct the dual form for Equation (1), which is given as:

$$\max D(\propto) = -\frac{1}{2} \sum_{i,j=1}^{n} \propto_i \propto_j y_i y_j k(x_i, x_j) + \sum_{i=1}^{n} \propto_i \textbf{(4)}$$

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *15 August 2023* ***Revised:*** *30 September 2023* ***Accepted:*** *18 October 2023*

$$s.t. \quad \sum_{i=1}^{n} \propto_i y_i = 0, \qquad i = 1, \dots, N, \textbf{(5)}$$

$$\propto_i \geq 0, \qquad i = 1, \dots, N, \textbf{(6)}$$

To transform nonlinear separable space into linear separable space, where *i* denotes Lagrange multipliers, the kernel matrix *k(xi, xj)* may be used to translate feature space into higher dimension space. Linear kernel functions, quadratic kernel functions, polynomial kernel functions, and Gaussian kernel functions are often used to do the feature mapping.

2) Bayes's Naïve Classifier

A particular kind of probabilistic graphical model is the Naive Bayes classifier. The Bayes Theorem (Equation (7)), which explains the likelihood of associating certain classes with particular examples, is the basis of the classifier. Features are assumed to be conditionally independent by the Naive Bayes classifier model. This may be seen in the work of:

$$P(c_i|x_i) = \frac{P(x_i|c_j)P(c_j)}{\sum_k P(x_i|c_k)P(c_k)}, \textbf{(7)}$$

$$P(x_i|c_j) = \prod_v P(x_{iv}|c_j), \textbf{(8)}$$

The identifier of the *jth*class is *cj*, where xiv denotes the *vth* characteristic of the ith instance. Then, by performing a calculation, instance *x* class is determined:

$$\max P(c_i|x) \textbf{(9)}$$

3) Artificial Neural Network(ANN)

The learning process in the human brain serves as the model for artificial neural networks (ANN). McCulloch and Pitts originally intended the mathematical Artificial Neuron model by modelling the artificial neuron process in 1943. As illustrated in Figure 2, an ANN model generally has three layers: input, hidden, and output. ANN is often used to address non-linear issues.

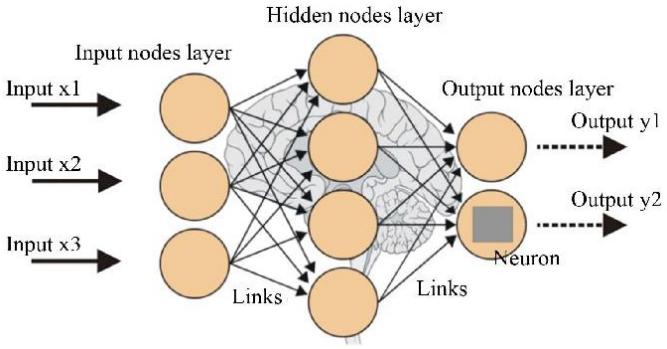

Figure 2: General ANN Process

4) AdaBoost

Boosting is a broad technique for raising the effectiveness of the learning process. The concept of boosting is to form a "committee" out of a number of "weak" learning algorithms. The performance of the learning algorithm as a whole is continuously enhanced by adjusting the weight priority for each weak classifier model in response to the training process. Given training set *T* and a binary issue, the adaptive boosting approach is stated as:

Algorithm 1: AdaBoost algorithm

Step 1: Initialize: $D_1(i) = \frac{1}{m}$ for $i = 1, \dots, m$

For $t = 1, \dots, T$

Step 2: Iteration (Boosting): Train decision tree using distribution D_t

Select h_t to minimalize the weighted error $\in_t$= P ($h_t(x_i \neq y_i)$

Choose $\propto_t = \frac{1}{2} In\ (\frac{1-\varepsilon_t}{\varepsilon_t})$

Update for $i = 1, \dots, m$

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \begin{cases} e^{-\propto t} & if\ h_t(x_i) = y_i \\ e^{\propto t} & if\ h_t(x_i) \neq y_i \end{cases}$$

$$= \frac{D_t\ (i) exp(-\propto_t y_i h_t(x_i))}{Z_t} \begin{cases} e^{-\propto t} & if\ h_t(x_i) = y_i \\ e^{\propto t} & if\ h_t(x_i) \neq y_i \end{cases}$$

Step 3: Get final Classifier

$$H\ (x) = sign\ (\sum_{t=1}^{T} \propto_t h_t(x))$$

An essential factor is allocated to *ht* based on weighted error, which is derived by the classification result, and *Dt (i)* is a distribution for the ith instance on the *t th* iteration in the first algorithm. Decision tree classifier is regarded as the basis classifier in this study.

5) K Fold Cross Validation

Test error is the most crucial metric to assess categorization performance for a given task. However, when a data set is provided, it is assumed that the model will be trained using all of the data, which may improve the model and provide a more trustworthy learning outcome. It is anticipated that after creating the model, a very large specified test set would be used to directly estimate the test error. A variety of statistical strategies are suggested to estimate test blunders when there isn't enough data to go around. This study uses K-fold cross-validation to evaluate the effectiveness of each model. This method randomly splits the data into groups of size k. throughout the method; each fold will act as a testing set, while the remaining groups will function as a training set. The test error, *Err1, Err2... Errk* is computed for each iteration. The technique of moments is used to estimate the model accuracy CV:

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *15 August 2023* ***Revised:*** *30 September 2023* ***Accepted:*** *18 October 2023*

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} Err_i, \quad \textbf{(10)}$$

$$Err_i = I(y_i \neq \hat{y}_i). \textbf{(11)}$$

IV. DATA ANALYSIS AND EXPERIMENTS

a) Data Description

This research makes use of data from the Wisconsin Breast Cancer Database (WBC) and the Wisconsin Diagnostic Breast Cancer (WDBC). These two data sets are used as they are often used in several studies. Test findings from this research may be broadly compared to those from earlier studies.

- Wisconsin Breast Cancer Database (WDC) (1991)

Dr. William H. Walberg obtained the WDC data collection from the University of Wisconsin Hospitals in Madison in 1991. 699 instances and 10 patient characteristics, such as instance identification, tumor information, classifications, etc., are included in the data set. Sixteen files are disregarded because they are lacking only one piece of information. There are 683 occurrences when the missing information data is removed. Table 1 provides a statistical breakdown of the 9 input characteristics.

Table 1: Wisconsin Breast Cancer Database (1991) Summary

Attribute Number	Attribute Description	Range	Mean	Standard deviation
1	Mitoses	1 – 10	3.65	1.23
2	Normal Nucleoli	1 – 10	2.69	2.49
3	Bland Chromatin	1 – 10	2.67	1.65
4	Bare Nuclei	1 – 10	1.25	1.84
5	Single Epithelial Cell Size	1 – 10	1.49	1.94
6	Marginal Adhesion	1 – 10	2.84	2.95
7	Uniformity of Cell Shape	1 – 10	2.31	1.48
8	Uniformity of Cell Size	1 – 10	1.36	1.57
9	Clump Thickness	1 – 10	0.32	0.68

- Wisconsin Diagnostic Breast Cancer (WDBC) (1995)

From the University of Wisconsin Hospitals in Madison in the late 1990s, Dr. William H. Wolberg also obtained the WDBC data collection. There are a total of 569 occurrences in this data set (62.74% benign, 37.26% malignant), 32 patient characteristics, one patient ID number record, thirty tumor diagnostic details, and one tumor diagnosis result record (benign and malignant). As a result of the collection, 30 feature records have been deposited into the data set of the tumor diagnostic information from 10 aspects and the provision of three measure results—the mean, standard error, and biggest value—for each attribution. In Table 2, data structure is briefly described here.

Table 2: Wisconsin Diagnostic Breast Cancer Database (1995) Summary

Attribute Number	Attribute Description	Range		
		Mean	Standard error	Largest value
1	Fractal dimension	7.87 - 36.23	0.11-2.87	7.93 - 36.04
2	Symmetry	9.71 - 39.28	0.36 - 4.89	12.02 - 49.54
3	Concave points	43.79 - 188.50	0.76 - 21.98	50.41 - 251.20
4	Concavity	143.50 - 2501.00	6.80 - 542.20	185.20 - 4254.00
5	Compactness	0.05 - 0.16	0.00 - 0.03	0.07 - 0.22
6	Smoothness	0.02 - 0.35	0.00 - 0.14	0.03 - 1.06
7	Area	0.00 - 0.43	0.00 - 0.40	0.00 - 1.25
8	Perimeter	0.00 - 0.20	0.00 - 0.05	0.00 - 0.29
9	Texture	0.11 - 0.30	0.01 - 0.08	0.16 - 0.66
10	Radius	0.05 - 0.10	0.00 - 0.03	0.06 - 0.21

A) Results and Comparison

Equation (12)'s accuracy measurement is used in order to compare the performance of various models, where True Positive, True Negative, False Positive, and False Negative are denoted by TP, TN, FP, and FN, respectively.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \textbf{(12)}$$

For every model test, 10 iterations are run using estimating test error with ten-fold cross-validation. Table 3 presents the test findings.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 15 August 2023 **Revised:** 30 September 2023 **Accepted:** 18 October 2023*

Table 3: Test Outcome

No.	Model	Data Set			
		WBC		WDBC	
		No. of PCs	Accuracy Mean (Std.)	No. of PCs	Accuracy Mean (Std.)
1	SVM	8	96.32% (0.33%)	31	96.83% (0.52%)
2	PCs-SVM	2	96.24% (0.34%)	9	96.73% (0.49%)
3	PCi-SVM	5	95.84% (0.35%)	11	97.19% (0.44%)
4	ANN	8	96.16% (0.17%)	31	91.72% (0.25%)
5	PCs-ANN	2	96.50% (0.16%)	9	91.79% (0.34%)
6	PCi-ANN	5	96.21% (0.14%)	11	93.32% (0.27%)
7	Naïve	8	94.33% (0.92%)	31	**99.63% (0.27%)**
8	PCs-Naïve	2	95.52% (0.67%)	9	99.61% (0.14%)
9	PCi-Naïve	5	89.88% (1.11%)	11	99.60% (0.20%)
10	AdaBoost	8	96.73% (0.76%)	31	97.90% (0.23%)
11	PCs-AdaBoost	2	**97.47% (0.10%)**	9	98.12% (0.20%)
12	PCi-AdaBoost	5	97.10% (0.09%)	11	97.99% (0.17%)

Table 3 is used as the basis for a paired t-test to assess the prediction accuracy of the two data sources. Since P-value = 0.315, which is the alternative hypothesis that an accuracy difference exists between the two separate data sets, the 5% significant threshold is designed exactly to fail to reject H0. Therefore, statistically speaking, the accuracy of these models is unaffected by diverse data sets.

Table 3 intuitively shows that PCs-SVM provides the most accuracy for WBC data when evaluating the best prediction model for each data set, whereas PCi ANN provides the highest accuracy when evaluating accuracy for WDBC data. To confirm the obvious conclusion, two sample t-tests are used. The following is a list of the null hypothesis and alternative hypothesis:

*H*0: Model *x* and the Best Candidate Model both predict outcomes with equal accuracy;

Ha: Model x exceeds the Best Candidate Model in terms of prediction accuracy.

For WBC data and WDBC data sets, the top potential models are PCs-SVM and PCi-ANN, respectively. Results for P-values for each test are shown in Table 4. For WBC data, PCs-SVM has the maximum accuracy under the 5% significant level requirement, the best accuracy performance for WDBC data is achieved by ANN, PCs ANN, and PCi-ANN.

Table 4: P-value of the hypothesis test (5% threshold of significance)

No.	Model	Data Set	
		WBC	WDBC
		PCs-SVM	PCi-ANN
1	SVM	1.11	1.11
2	PCs-SVM	-	1.11
3	PCi-SVM	1.22	1.11
4	ANN	1.11	1.45
5	PCs-ANN	1.11	1.66
6	PCi-ANN	1.11	-
7	Naïve	1.11	1.11
8	PCs-Naïve	1.11	1.11
9	PCi-Naïve	1.11	1.11
10	AdaBoost	1.11	1.11
11	PCs-AdaBoost	1.11	1.11
12	PCi-AdaBoost	1.11	1.11

Since the primary elements don't accurately represent the full extent of the data set, PCA preprocessing may minimize data noise and provide superior results, which enrich the feature space (elite effect).

CONCLUSION

In this era of big data, the process of data mining is fraught with a great deal of difficulty. The findings of this research provide a perspective on how cancer diagnosis may be improved via the use of data mining tools in the healthcare system. In this investigation, two different data sets are used to evaluate both four different data mining methods and eight different hybrid models. In particular, principal component analysis (PCA), which is a method for dimension reduction, demonstrates certain benefits in terms of the accuracy and efficiency of predictions having said that, there are still a few areas that can be investigated in more depth in the years to come. The principal component analysis (PCA) is a linear approach that transforms feature space into a linear function that is based on uncorrelated variables. Other strategies, such as k-means, may also be evaluated as potential nonlinear feature reduction approaches. Among these strategies is the k-means method. In addition, the data sets that were utilized in this research were all

standard data sets. In the future, more research may be conducted using raw data sets such as SEER.

REFERENCES

[1] Chaurasia V, Pal S, Tiwari BB. "Prediction of benign and malignant breast cancer using data mining techniques". Journal of Algorithms & Computational Technology. 2018; 12(2):119-26.

[2] Yue W, Wang Z, Chen H, Payne A, Liu X. "Machine learning with applications in breast cancer diagnosis and prognosis". Designs. 2018; 2(2):13.

[3] Banu B, Thirumalaikolundusubramanian P." Comparison of Bayes Classifiers for Breast Cancer Classification". Asian Pacific journal of cancer prevention (APJCP). 2018; 19(10):2917-20.

[4] Akinsola AF, Sokunbi MA, Onadokun IO. "Data Mining For Breast Cancer Classification". International Journal of Engineering And Computer Science. 2017; 6(8): 22250-22258.

[5] Mirajkar P, Lakshmi P. "Prediction of Cancer Risk in Perspective of Symptoms using Naïve Bayes Classifier". International Journal of Engineering Research in Computer Science and Engineering. 2017; 4(9):145-149

[6] Oskouei RJ, Kor NM, Maleki SA. "Data mining and medical world: breast cancers‟ diagnosis, treatment, prognosis and challenges". American journal of cancer research. 2017; 7(3):610-27.

[7] 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.

[8] Siegel, R., Ma J., Zou Z., and Jemal A., 2014, "Cancer Statistics 2014," CA: A Cancer Journal for Clinicians, 64(1), 9-29.

[9] Sotiriou, C., Neo, S. Y., McShane, L. M., Korn, E. L., Long, P. M., Jazaeri, A., and Liu, E. T., 2003, "Breast Cancer Classification and Prognosis Based on Gene Expression Profiles From A Population-based Study," Proceedings of the National Academy of Sciences, 100(18), 10393-10398.

[10] Sobin, L. H., Gospodarowicz, M. K., and Wittekind, C., TNM Classification of Malignant Tumors, John Wiley & Sons, 2011.

[11] Rani, K. U., 2010, "Parallel Approach for Diagnosis of Breast Cancer using Neural Network Technique," International Journal of Computer Applications, 10(3), 1-5.

[12] Delen, D., Walker, G. and Kadam, A., 2005, "Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods," Artificial Intelligence in Medicine, 34(2), 113-127.

[13] Choi, J. P., Han, T. H. and Park, R. W, 2009, "A Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis," Journal of Korean Society of Medical Informatics, 15(1), 49-57.

[14] Bellaachia, A. and Guven E., 2006, "Predicting Breast Cancer Survivability Using Data Mining Techniques," Age, 58(13), 10-110.

[15] Psychogios, G., Waldfahrer, F., Bozzato, A., and Iro, H., 2010, "Evaluation of the Revised TNM Classification in Advanced Laryngeal Cancer," European Archives of Oto-Rhino-Laryngology, 267(1), 117-121.

[16] Wang, W., Sun, X. W., Li, C. F., Lv, L., Li, Y. F., Chen, Y. B., and Zhou, Z. W., 2011, "Comparison of the 6th and 7th Editions of the UICC TNM Staging System for Gastric Cancer: Results of a Chinese Single institution Study of 1,503 Patients," Annals of Surgical Oncology, 18(4), 1060-1067.