ISSN: 0974-5823 Vol. 6 No. 3 October-December, 2021

International Journal of Mechanical Engineering

Breast Cancer Prediction Using Machine Learning Algorithms

Kavitha Chinniyan
Department of Computer Science and Engineering, PSG College of Technology, India

Roobini Subramani
Department of Computer Science and Engineering, PSG College of Technology, India

***Abstract*- Breast Cancer is one of the major causes of death in women. Treatment against malicious cancer in tissue has lead to numerous visual examinations for Breast cancer. In cancer research, classification of tissue samples is necessary for cancer diagnosis. This can be done with the help of quantitative inference and extraction models to prevent tissue from further growth. The classification of tissues surrounding the malicious cancer cells into benign and malignant categories is extremely challenging task to predict using data mining algorithms. In this paper, a novel quantitative inference of breast cancer tissues is examined using semi-supervised mechanism. In semi-supervised approach, the records in dataset are clustered using farthest first clustering. The clustered records are classified using Artificial Neural Network and Naive Bayesian classifiers. The proposed method is evaluated using k-fold cross validation. The Wisconsin Original breast cancer dataset from UCI Repository is used to test the efficiency of the proposed model. The experimental results demonstrate that, the proposed technique produces the best accuracy of 98.0% for Artificial Neural Network and 95.2% for Naïve Bayesian Classifier. Among them, Artificial Neural Networks has proved to be one of the best classifier with 2.8% of improved accuracy**

Index Terms— Artificial Neural Network, Prediction, Accuracy, Decision Tree, Simulated Annealing.

I. INTRODUCTION

Machine learning methods are extensively used in medical application which includes identification and classification of tumours. Machine learning is improving in diagnostics, to predict outcomes, and begins to scratch the surface of personalized care. It is mainly used as an aid for cancer diagnosis and prediction. Researchers have recently involved in pertaining the machine learning techniques towards cancer prediction and prognosis. Machine learning is a more powerful arena because it allows decisions to be made which could not be possibly made using conventional methodologies [15]. Predictive analytics is one of the areas in data mining which deals with extracting information from data and used to predict the trends and behaviour patterns. Predictive analytics is a high level statistical method which has an ability to build predictive models [16]. Breast cancer is an important research topic in medical science.Breast cancer is the most common invasive cancer among women, with more than one million cases and deaths occurring worldwide annually. The most effective way to reduce breast cancer death is to detect at an earlier stage [28]. The main objective is to predict the breast cancer in advance that ensures a long survival of patients. A complicated test for the main diagnosis of breast cancer makes it difficult to obtain the results as cancer or non-cancer [18]. In predictive analytics, predicting the outcome of a disease is one of the most fascinating and challenging tasks. The machine learning algorithms could be used directly to find the final result as cancer or non-cancer by exploiting various data mining techniques. There are several possible solutions for early diagnosis with accurate prediction of breast cancer such as supervised and unsupervised learning [23]. Supervised Learning includes Decision tree, a popular classification approaches in knowledge discovery and data mining, which classifies the labeled trained data into a tree or rules [21], Artificial Neural Network (ANN) is a mathematical model or computational model based on biological neural networks, K- nearest neighbor (Knn) is used to classify the building model [26], SVM built optimal separating boundary between datasets to solve optimization problem and the association rule discovery techniques to construct classification systems [22]. Unsupervised learning includes clustering which discovers useful patterns within the data. Semi –Supervised learning is also called as inductive learning which is used to infer the correct label for unlabeled dataset [17].

II. RELATED WORK

MeghaRathi and VikasPareek [4] suggest a framework which has Data Pre-processing, Feature selection, Feature Subset and different classifiers for making prediction. The data pre-processing is used to remove data-inconsistency and irrelevant data. The feature selection is used to extract subset of features using MRMR (Maximum Relevance and Minimum Redundancy) algorithm. It is used for selection of features and to improve the accuracy of classifiers by selecting subset of features. Four classifiers such as End Meta, Naïve Bayes, SVM (Support Vector Machine), and FT

(Function Tree) is used. After extracting the relevant features from the data set, classifier is applied to check the performance in terms of accuracy. The classifiers are trained using training set, and then classification algorithm is applied one by one and SVM with MRMR produces better results. The disadvantage is selected features are correlated strongly and it is mutually far away from each other due to which it has high correlation and redundant features.

Hamid Mohamadi, JafarHabibi, Mohammad SanieeAbadehand HamidSaadi [7] proposes a simulated annealing based Fuzzy Classification System (SAFCS). SAFCS generates a fuzzy if-then rules and temperature initialization. The procedure is repeated k times. Then temperature is decremented using cooling parameter and iterates until the stopping criterion is reached. SAFCS is compared with C4.5 which is based on entropy criteria and pruning techniques, inorder to discard the parts of a tree. These classification methods are applied to different datasets, among them SAFCS achieves better results in terms of accuracy for both training set and testing set. The disadvantage of SAFCS is difficult to fix the cooling parameter and time consuming.

Animesh Hazra, Subrata Kumar Mandal, Amit Gupta [8] the main objective is to find the smallest subset of features that can ensure a best accurate classification of breast cancer. It compares three classifiers Naïve bayes, SVM and ensemble classifier. After pre-processing, the comparison is carried out in three ways such as Feature selection using Pearson correlation coefficient, Pearson correlation with binning and principal component analysis. The Pearson correlation is used to reveal about how much the class attributes and attributes of data set are related based on which the features are ranked. In second methodology after data cleaning and Pearson correlation coefficient, discretization of binning technique was applied. In the next method, feature extraction was performed with principal component analysis (PCA) after pre-processing. Then classification technique was applied, among all classifiers naïve bayes provides better accuracy percentage and time complexity. The disadvantage is execution time for prediction model and PCA maps the data into lower dimensional space due to which performance of classifier is degraded.

Mohamed Junaid.K.A, [10] discusses about two layer neural network back propagation algorithm. During learning phase, the network learns by adjusting the weights and takes longer learning time. As a result, parallelization techniques are used to speed up the computation process and outperforms in terms of accuracy. As a result of two layer neural networks, parallelization techniques are used to speed up the computation process and outperforms in terms of accuracy. The disadvantage of this approach is too many hidden neurons and difficult to train the network.

Deepshree A.Vadeyar,Yogish H.K [11] focuses on farthest first clustering for reorganization of website structure to improve user navigation. The websites are considered as a graph, each node as a web page, redirecting URL between pages as edges and links are represented as 1/0. The clusters are formed in which URL acts as objects for cluster and a threshold criterion is set for clusters [5]. The website structure depends on the threshold criterion of the clustering. It is difficult to assign the threshold criterion is the disadvantage.

SunitaSoni, O.P.Vyas [12] discusses about association rule mining, associative classifiers and advanced classifier rule mining. Classification using association rule mining is a predictive analytics technique which discovers a subset of rules to predict an accurate classifier called Classifier Association Rule (CAR). The advanced CAR provides better accuracy which depends on set of rules before applying classifiers. The associative classifiers using Fuzzy association rule, deals with sharp boundary problems. The fuzzy and weighted associative classifier provides a greater accuracy when compared to other associative classifiers. The disadvantage is difficulty in handling temporal data which is being dynamic. The rules predicted are not static and it has additionally related attributes.

Ibrahim M. El-Hasnony, Hazem M. El-Bakry, Ahmed A. Saleh [13] discusses a hybrid methodology of K-means, Feature reduction with feature selection (FRFS) and discrenibility K-nearest neighbor (D-Knn) classifier. The hybrid feature selection and data reduction method is used for combining the rough set features and calculating the attribute dependencies. The classifier makes decisions at each point and classifies the data by contrasting test set is called Instance based learning (IBL). Knn assumes that class label for each record is same for nearest neighbor which is simple that helps to enhance its predictive accuracy. The disadvantage is to determine value of parameter computation cost is quite high.

III. METHODOLOGIES

The following are the research methods employed in this paper.

A. Data collection and Pre- Processing

The dataset is collected from UCI Machine learning data repository of Wisconsin (Original) Breast cancer dataset (WBC). WBC has 699 instances, 2 class labels (2 for Benign/4 for Malignant) and 11 attributes. The attributes are integer valued. The WBC dataset is provided in Table I.

TABLE I. DATASET DESCRIPTION

S.NO	**ATTRIBUTE**	**DOMAIN**
1.	Sample Code Number	Id number
2.	Clump Thickness	1-10
3.	Uniformity of Cell Size	1-10
4.	Uniformity of Cell Shape	1-10
5.	Marginal Adhesion	1-10
6.	Single Epithelial Call Size	1-10

7.	Bare Nuclei	1-10
8.	Bland Chromatin	1-10
9.	Normal Nucleoli	1-10
10.	Mitoses	1-10
11.	Class	2-benign, 4-malignant

The dataset contains missing values '?' is pre-processed by single imputation method, i.e., the mean value of a variable. The advantage is sample mean remains unchanged [25].

B. Fuzzy C-Means Clustering

The clustering has been classified as soft clustering and hard Clustering. In hard clustering, the data point belongs to exactly one cluster. In Soft Clustering, the data point can belong to more than one cluster [14]. The Fuzzy C-Means Clustering (FCM) algorithm is an unsupervised and soft clustering algorithm, which introduces the fuzziness for an object. The main limitations are sensitive to noises, minimizes an objective function and difficult to select appropriate parameters. The algorithm is provided in Table II.

In algorithm, where

K= Number of Iterations
V_j= j^{th} cluster center
m= Fuzziness Index [1,∞]
C= Number of Cluster Center
μ_{ij} = membership of i^{th} data to j^{th} cluster center
d_{ij} = Eucledian distance between i^{th} data and j^{th} cluster
center
x_i = Original data point
n = Number of Data Points

TABLE II.FUZZY C-MEANS CLUSTERING ALGORITHM

C. Simulated Annealing

Simulated Annealing is an iterative method which is mainly used as an optimization search paradigm to escape from local minima and to achieve global optima. This optimization can be done by accepting moves which degrades the probability on a parameter called temperature. The temperature is gradually decreased by using cooling schedule [1]. The algorithm behavior ends when the temperature reaches to zero. The parameter required for simulated annealing are starting temperature, final temperature and temperature decrement. The final cluster structure depends on how the cooling is performed [9]. The ideal cooling rate is difficult to compute. The temperature decrement is given by,

$$f(t) = t\,\alpha \qquad (1)$$

Where,
t = time in minutes, which lies between [1, ∞]

α = Cooling rate which lies between 0.5 - 0.99

D. Decision Tree [C4.5]

Decision Tree is a supervised algorithm which can be used for classification problems. C4.5 is also a type of decision tree which is improved from ID3 algorithm by dealing with both continuous and discrete attributes, missing values and pruning trees [2].C4.5 builds decision trees from a set of training data by calculating the information gain for each attribute. The complexity of a decision tree is, the tree becomes unstable even when there is a small change in entropy value [21]. It will result in a generation of different tree. The sub trees replicated several times.

$$Entropy(S) = -\frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right) \qquad (2)$$

The attribute with highest information gain is taken as a root for decision tree [27]. In the equation 2, where

p =Number of positive classes
n=Number of negative classes.

The algorithm steps are provided in Table III.

TABLE III. DECISION TREE [C4.5] ALGORITHM

Algorithm: Fuzzy C-means Clustering Algorithm
Input : Pre-processed dataset
Output : Clustered cancer dataset

1. Randomly 'C' Cluster centres are selected.
2. Calculate fuzzy membership matrix.

$$\mu_{ij} = 1/\sum_{k=1}^{c}(d_{ij}/d_{ik})\wedge(2/m-1)$$

3. Compute Fuzzy centers 'V_j' using

$$V_j = (\sum_{i=1}^{n}((\mu_{ij})\wedge m)x_i)/\sum_{i=1}^{n}((\mu_{ij})\wedge m), \forall j = 1,2..c$$

4. Repeat 2 & 3 until minimum J value is achieved

$$J(u,v) = \sum_{i=1}^{n}\sum_{j=1}^{c}((\mu_{ij})\wedge m))(\| x_i - v_j \|\wedge 2)$$

```
Algorithm: Decision Tree (C4.5)
Input : Training dataset (D)
Output : Decision Tree (T)
Tree{}
if D is "pure" OR other stopping criteria met then terminate
end if
for all attribute a € D do
    Compute information gain to split on 'a'
end for
$a_{best =}$ Best attribute according to gain calculated
Tree = Create a decision node that tests $a_{best}$ in the root
$D_v$ = Induced sub – datasets from D based on $a_{best}$
for all $D_v$do
   $Tree_v$ =C4.5($D_v$)
   Attach $Tree_v$ to the corresponding branch of tree
end for
return Tree
```

E. Farthest First Clustering algorithm

Farthest First clustering is a type of Hard Clustering algorithm in which one data point can belong to only one cluster [20]. It is similar to K-Means clustering algorithm it chooses centroids and assigns objects in cluster with max distance. Farthest first clustering solves k-center problem and it is very efficient for large set of data. In farthest first algorithm, the mean for calculating centroid is not required which takes centroid arbitrary and distance of one centroid from other is maximum [24]. After calculating centroid, the points with minimum distance are assigned to clusters. The algorithm is provided in Table IV.

TABLE IV. FARTHEST FIRST CLUSTERING ALGORITHM

```
Algorithm: Farthest First Clustering
Input :  Pre-processed dataset
Output :  Clustered cancer dataset
Steps :
     1.  Pick any data point and label it as point 1
     2.  For i=2,3,…,n
     3.  Find the unlabelled point furthest from {1,2,…,i-1} andlabel it as i.
//Use d(x,S) = minyϵS d(x,y) to identify the distance of apoint from a set
     4.  π(i) = argmin j<i, d(i,j)
     5.  Ri=d(i,π(i))
```

F. Naïve Bayesian

In machine learning, naïve bayes algorithm is a simple probabilistic classifiers based on Bayes theorem with strong independence assumptions between the features [6]. Naive Bayes classifiers are highly scalable which requires a number of parameters to be linear and the number of variables (features/predictors) for training the model. Naive Bayesian classifier is based on Bayes' theorem and the theorem of total probability. The Advantages of Naïve Bayesian Model are:

- It is a simple classifier because it has an underlying probabilistic model.
- Requires only small amount of data to train the model.
- If the independence assumption holds well, NB Classifier performs well and provides better results.

The probability with vector x = < x1... xn> belongs to hypothesis h is,

$$P(Y \mid X_1,..X_n) = P(X_1,..X_n \mid Y) / P(X_1, X_2...Xn)$$
(3)

Where, X = Number of predictors.
Y = Class of probability.

TABLE V. BACK PROPAGATION ALGORITHM

G. Artificial Neural Network

The artificial neural network (ANN) was implemented using back propagation architecture. Each of the layers has certain elements called as neurons, which has input layer, hidden layer and output layer. The input layer is selected based on the features and output layer has two neurons either as benign or malignant [15]. The transfer function is computed as a weighted sum of input signals. The learning capability of an artificial neuron is achieved by adjusting the weights of neurons. Initially, the process is carried out as forward pass, then back propagation is applied to achieve target node. Back Propagation learns by iteratively processing a set of clustered samples. For each sample, weights are modified to minimize the error between network's classification and actual classification Process [29]. The algorithm of ANN is provided in Table V. The advantages of ANN are [30]:

- ANN is used to solve complex problems.
- Less effort is required to train the network model.
- Implicitly detects Complex non – linear relationship between dependant and independent variables.
- Ability to detect all possible interactions between predictor (attributes) variables.

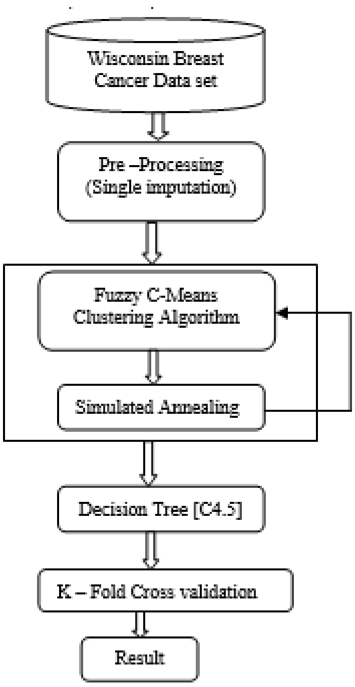
Wisconsin Breast Cancer Data set
Pre –Processing (Single imputation)
Fuzzy C-Means Clustering Algorithm
Simulated Annealing
Decision Tree [C4.5]
K – Fold Cross validation
Result

Figure. 1. Existing System Flow Chart

IV. SYSTEM DESIGN

In existing method, the combined approach such as Fuzzy C-means clustering with Simulated annealing and Decision tree (C4.5) classifier is used for diagnosis of breast cancer. The detailed working of the existing module is described as follows :

A. Existing System

The dataset from UCI repository is used for prediction, in which single imputation pre-processing technique was used. Then inorder to predict labels, FCM clustering is processed [1] on dataset which is then optimized using Simulated Annealing. In FCM, 'm' is a fuzziness index which measures the tolerance for clustering and its value lies between [1, ∞]. If the value of 'm' is larger, it has larger overlapping between clusters. m=1 for crisp and 2 for fuzzy clustering. In this research [19], m=1.4 is chosen as fuzzy index. Then, the clustered data is annealed for which the cooling schedule is chosen as f(t)=4. The starting and final temperature is chosen as minimum and maximum of a feature in a random manner. After clustering with annealing, C4.5 classifiers are used to classify the clustered dataset and labels are predicted either as Benign or Malignant. The model is then cross validated using K-fold cross validation, here K=10. The existing system flows as it is being provided in figure 1.

B. Disadvantages of existing model

The drawbacks of existing model are:

- FCM clustering takes O (ndc^2i) as a computation time to cluster the dataset and optimized by simulated annealing
  where,
  n = Number of Data Points.
  d = Number of dimensions.
  c = Number of Clusters.
  i = Number of Iterations.
- FCM provides K-center problem ie. Distances between cluster centers should satisfy the theory, triangle of inequality.
- Computationally expensive, difficult to fix thresholds and cooling rate.
- In decision tree, normally over fitting occurs, then pruning is required due to which accuracy gets decreased.

C. Proposed System

In proposed model, farthest first clustering and Naïve Bayesian model is used for diagnosis of breast cancer. Farthest first clustering is used to cluster the pre-processed dataset as given in table 3. In existing work, simulated annealing was used to optimize the clusters. The farthest first clustering does not provide local optimum even in worst case so it does not require optimization. Farthest first clustering always provides global optimum so, it guarantees to provide optimal solution. In our proposed model, two clusters are considered because it has only two labels such as benign and malignant. After clustering, Naïve Bayesian model and Artificial Neural Network is used to classify the clustered data and labels are predicted either as Benign or Malignant. Then, it is cross validated using K-fold cross validation. In this k=10, the dataset is divided into 10 equal subsets. Each subset acts as testing set, whereas the rest of the partition acts as training set. This procedure is repeated ten times so that each partition is used to test only once. The proposed system flows as it is being provided in figure 2.

D. Advantages of Proposed model

The advantages of proposed model are:

- FFC Clustering computation time is O(nk)
  Where, n = Number of data points.
  k = number of clusters.
- FFC achieves global optimum because it does not require iteration cluster formation takes place in a single pass.
- Farthest first solves K-center problem ie.,it achieves triangle of inequality property.
- Artificial Neural Network (ANN) and Naive Bayesian provide optimal solution because it performs well does not over fit.
- It is a simple classifier with high accuracy.

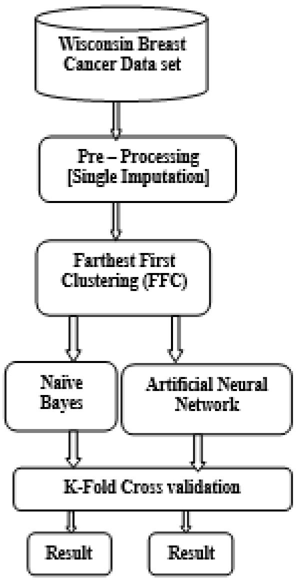
Wisconsin Breast Cancer Data set
Pre – Processing [Single Imputation]
Farthest First Clustering (FFC)
Naïve Bayes
Artificial Neural Network
K-Fold Cross validation
Result
Result

Figure 2. Proposed System Flow Chart

V. EXPERIMENTS AND RESULTS

A. *Performance Metrics*

In this experiment, the medical data related to breast cancer classification is initiated with the preprocessing using mean imputation which is followed by the farthest fast clustering. The clustered data is classified using artificial neural network and naive Bayes, to predict the tumors. The performance of classifier is validated based on error rate and mean accuracy. The performance metrics are precision, recall, F-Measure, error rate and accuracy [20].

B. *Precision*

Precision is a ratio of true positive tuple and all positive tuple in a dataset. It is used to measure the exactness or quality. Precision is given by,

$$\text{Precision} = TP / TP + FP \quad (4)$$

C. *Recall*

Recall is a ratio of true positive tuple against positive and negative tuple. Recall is the measure of completeness or quality. It is also called as sensitivity. Recall is given by,

$$\text{Recall} = TP / TP + FN \quad (5)$$

D. *F-Measure*

F – Measure is also called as F – Score. F –Measure is a mean of precision and recall. F- Measure value varies from 0 to 1. If the value of F-Measure is higher, then it is said to be a better classifier. It is given by,

$$F - \text{Measure} = 2 * ((\text{precision} * \text{Recall}) / (\text{Precision} + \text{Recall})) \quad (6)$$

E. *Accuracy*

The classifiers accuracy is an important metric for evaluation. It is a ratio of positive tuples and negative tuples against all the tuple. It is given by,

$$\text{Accuracy} = TP + TN / TP + TN + FP + FN \quad (7)$$

F. *Error Rate*

The error rate is an essential measure for evaluation. Lower error rate is said to be a better classifier. Error rate determines the error between the prediction and actual. It is given by,

$$\text{Error rate} = FP + FN / TP + TN + FP + FN \quad (8)$$

TABLE VI: CONFUSION MATRIX FOR DECISION TREE [C4.5]

Testing & Training (%)	True positive (TP)	True Negative (TN)	False positive (FP)	False Negative (FN)
50 -50 %	219	0	130	0
60 – 40 %	172	0	108	0
75 - 25 %	110	0	65	0
80 – 20%	118	0	22	0
90 – 10 %	43	0	27	0

TABLE VII: CONFUSION MATRIX FOR NAÏVE BAYES MODEL

Testing & Training (%)	True positive (TP)	True Negative (TN)	False positive (FP)	False Negative (FN)
50 -50 %	282	50	0	17
60 – 40 %	212	56	1	11
75 - 25 %	143	24	0	8
80 – 20%	110	22	0	8
90 – 10 %	57	10	0	3

G. *Confusion Matrix*

The confusion matrix is used to describe the performance of a classifier. True positive (TP) refers to positive tuple and True Negative (TN) refers to negative tuple classified by the basic classifiers. False Positive (FP) refers to positive tuple but predicted as negative and False Negative (FN) refers to negative tuple but predicted as positive which is incorrectly classified by the classifiers. The training dataset is used to train the classifiers. The correctly classified instances and incorrectly classified instances for existing model ie., Decision Tree [C4.5] is provided in the table VI. The Proposed model ie., Naïve Bayes model in table VII and Artificial Neural Network is provided in table VIII.

TABLE VIII: CONFUSION MATRIX FOR ARTIFICIAL NEURAL NETWORK [ANN]

Testing & Training (%)	True positive (TP)	True Negative (TN)	False positive (FP)	False Negative (FN)
50 -50 %	291	49	1	8
60 – 40 %	218	56	1	5
75 - 25 %	145	24	0	6
80 – 20%	114	22	0	4
90 – 10 %	58	10	0	2

TABLE IX. PERFORMANCE METRICS EVALUATED FOR FUZZY C-MEANS CLUSTERING (FCM) WITH

Testing & Training (%)	Precision	Recall	F-measure	Accuracy (%)	Error-rate
50 -50 %	0.6275	1.000	0.771	62.7%	0.3725
60-40%	0.614	1.000	0.760	61.4%	0.386
75-25%	0.6285	1.000	0.771	62.8%	0.3715
80-20%	0.84	1.000	0.91	84%	0.16
90-10%	0.614	1.000	0.760	61.4%	0.386
Balanced Average	0.6648	1.000	0.7944	**66.4%**	**0.3352**

SIMULATED ANNEALING (SA) AND DECISION TREE C4.5

TABLE X. EVALUATION METRICS OBTAINED

Testing & Training (%)	Precision	Recall	F-measure	Accuracy (%)	Error-rate
50 -50 %	1.000	0.943	0.970	95.1%	0.049
60-40%	0.9953	0.9506	0.9723	95.7%	0.043
75-25%	1.000	0.947	0.972	95.4%	0.046
80-20%	1.000	0.9322	0.964	94.2%	0.06
90-10%	1.000	0.95	0.974	95.7%	0.043
Balanced Average	0.9990	0.9445	0.9704	**95.2%**	**0.048**

FOR FARTHEST FIRST CLUSTERING AND NAÏVE BAYESIAN MODEL

TABLE XI. EVALUATION METRICS OBTAINED FOR FARTHEST FIRST CLUSTERING AND ARTIFICIAL NEURAL NETWORK MODEL

Testing & Training (%)	Precision	Recall	F-measure	Accuracy (%)	Error-rate
50 -50 %	0.9965	0.973	0.9845	97.4%	0.026
60-40%	0.9954	0.977	0.986	97.8%	0.022
75-25%	1.000	0.96	0.979	97.7%	0.023
80-20%	1.000	0.94	0.9690	97.1%	0.029
90-10%	1.000	0.967	0.983	97.1%	0.029
Balanced average	0.9983	0.9634	0.9803	**98.0%**	**0.02**

H. Results

The experiments have been done on R-studio of version 3.0.3.The results for existing and proposed model are shown in Table IX, X and X1. The table describes the performance of the model in terms of accuracy, precision, recall, f-measure and error rate. The result shows that the classifier with Farthest First Clustering (FFC) provides better accuracy with lower error rate. The Farthest First Clustering clusters the data with maximum distance. The quality of clusters depends on distance between the cluster centers should be farther and the cluster is said to be well formed. Farthest first clustering satisfies the criteria when compared to Fuzzy C-Means Clustering. The Farthest First Clustering with Artificial Neural Network provides a better balanced accuracy of 98.0%, lower error rate of 2% with all testing and training phases because ANN has a capability to solve a complex and mimic problem. Due to that, ANN can be used for many real world applications which give better results.

VI. **CONCLUSION**

In this paper, the prediction model for breast cancer is constructed by using classifiers with clustering. The performance of a model is analyzed by comparing the different classifiers with clustering. The Farthest First Clustering provides a better prediction compared to fuzzy clusters. The farthest first clustering with Naïve Bayes and Artificial Neural Network classifier provides balanced accuracy of 95.2% and 98.0%. Therefore, Farthest First Clustering [FFC] with Artificial Neural Network achieves an improved accuracy of 2.8% because ANN is a powerful machine learning algorithm. ANN is suggested as a better prediction model for diagnosis of breast cancer to ensure the long survival of patients. In future, ANN with genetic algorithm can be made as a hybrid model to predict the disease which evolves a huge remark in medical science [31].

REFERENCES

[1] SudiptaAcharya,SriparnaSaha and YaminiThadisina, “Multultiobjective Simulated Annealing-Based Clustering of Tissue Samples for Cancer Diagnosis”, IEEE Journal of Biomedical and Health Informatics, vol 20, no.2, March 2016.

[2] Sonia Singh and Priyanka Gupta, “Comparative Study Id3, Cart and C4.5 Decision Tree Algorithm: A Survey”, International Journal of Advanced Information Science and Technology (IJAIST), vol.27, no.27, July 2014.

[3] K.Vembandasamy, T.Karthikeyan, “Noval outlier detection in diabetics Classification Using Datamining techniques” , International Journal of Applied Engineering Research, ISSN 0973 – 4562, Vol 11, 2016.

[4] MeghaRathi, VikasPareek, “Hybrid Approach to predict Breast Cancer using Machine Learning Techniques”, International Journal of Computer Science and engineering, May 2016.

[5] Digambar A Kulkarni, Vijaylaxmi K Kochari, "Detection of Breast Cancer Using K Means Algorithm" in International Journal of Emerging Technology and Advanced Engineering Volume 6, Issue 4, April 2016.

[6] Murat Karabatak , "A new classifier for breast cancer detection based on Naïve Bayesian in Measurement”, Elsevier Science Direct , 2015.

[7] Hamid Mohamadi, JafarHabibi, Mohammad SanieeAbadehand HamidSaadi,” Datamining with a simulated annealing based fuzzy classification system”, Elsevier Science Direct, pp.1824 – 1833, Nov 2013.

[8] AnimeshHazra, Subrata Kumar Mandal, Amit Gupta, “Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms”, International Journal of Computer Applications
(0975 – 8887) Volume 145 – No.2, July 2016.

[9] Wei Yang, Luis Rueda, AliouneNgom, “A Simulated Annealing
Approach to Find the Optimal Parameters for Fuzzy Clustering Microarray Data”, Elsevier Science Direct, December 2014.

[10] Mohamed Junaid.K.A, “Classification Using Two Layer Neural
Network Back Propagation Algorithm”, Scientific research
publishing, June 2016.

[11] Deepshree A.Vadeyar,Yogish H.K, “Farthest First Clustering in Links
Reorganization”, International Journal of Web & Semantic
Technology (IJWesT) Vol.5, No.3, July 2014.

[12] SunitaSoni, O.P.Vyas,“Using Associative Classifiers for Predictive
Analysis in Health Care Data Mining”, International Journal of
Computer Applications (0975 – 8887) Volume 4 – No.5, July 2010.

[13] Ibrahim M. El-Hasnony, Hazem M. El-Bakry, Ahmed A. Saleh, “Classification of Breast Cancer Using Softcomputing Techniques”, International Journal of Electronics and Information Engineering, Vol.4, No.1, PP.45-54, Mar. 2016.

[14] Tejwant Singh and Mr. Manish Mahajan, “Performance Comparison of Fuzzy C Means with Respect to Other Clustering Algorithm”, International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), vol 4,Issue 5, pp.89-93,May 2015.

[15] M. Sheha, M. S. Mabrouk, and A. Sharawy, “Automatic detection of melanoma skin cancer using texture analysis,” Int. J. Comput. Appl., vol. 42, no. 20, pp. 22–26, Mar. 2012.

[16] Knox H.Todd, “Cancer facts and figures 2012,” American Cancer Society, Atlanta, GA, USA, 2012.

[17] U. Maulik, A. Mukhopadhyay, and D. Chakraborty, “Gene-expression based cancer subtypes prediction through feature selection and transductive SVM,” IEEE Trans. Biomed. Eng., vol. 60, no. 4, pp. 1111–1117, 2013.

[18] S. C. Dinger, M. A. Van Wyk, S. Carmona, and D. M. Rubin, “Clustering gene expression data using a diffraction inspired framework”, Biomed. Eng. Online, vol. 11, no. 1, p. 85, 2012.

[19] S. Saha, A. Ekbal, K. Gupta, and S. Bandyopadhyay, “Gene expression data clustering using a multiobjective symmetry based clustering technique,” Comput. Biol. Med. vol. 43.11, pp. 1965–1977, 2013.

[20] DelshiHowsalya Devi R and Dr. M Indra Devi, “Outlier Detection Algorithm Combined With Decision Tree Classifier for Early Diagnosis of Breast Cancer”, International Journal of advanced engineering and technology(IJAE), vol. VII, Issue II, pp.93-98,April-June 2016.

[21] K.Sivakami, “Mining Big Data: Breast Cancer Prediction using DT - SVM Hybrid Model”, International Journal of Scientific Engineering and Applied Science (IJSEAS), vol.1, Issue5, pp.418-429, August 2015.

[22] SmrutiRekha Das, Pradeepta Kumar Panigrahi, Kaberi DaS and Debahuti Mishra,”Improving RBF Kernel Function of Support Vector Machine using Particle Swarm Optimization”, Internaional Journal of Advanced Computer Research, vol-2, no-4, Issue-7, ISSN : 2249-7277 December 2012.

[23] Cuong Nguyen, Yong Wang and Ha Nam Nguyen, “Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic”, J. Biomedical Science and Engineering, pp.552-560, May 2013.

[24] S.Kharya, D.Dubey, and S.Soni, “Predictive Machine Learning Techniques forBreast Cancer Detection” International Journal of
Computer Science and Information Technologies, Vol. 4, Issue 6, pp.1023-1028, Nov – Dec 2013.

[25] K.Arutchelvan and Dr.R.Periyasamy, “Cancer Prediction System Using Data mining Techniques”, International Research Journal of Engineering and Technology (IRJET), ISSN: 2395-0056, Vol.02, Issue 08, Nov a2015.

[26] J.S.Raikwal and KanakSaxena, “Performance Evaluation of SVM and K-Nearest Neighbour Algorithm

Over Medical Data Set" , International Journal of Computer Applications, vol.50, no.14, pp.35-39,July 2012.

[27] Jahanavi Joshi, RinalDoshi, "Diagnosis and prognosis of Breast Cancer using Classification rules", International Journal of Engineering research and General Science, Vol 2, Issue 6, October – November, 2014.

[28] Souad Demigha,"Data Mining for Breast Cancer Screening", IEEE Spectrum, Sep 2015.

[29] R. R. Janghel, Anupam Shukla, Ritu Tiwari "Breast Cancer Diagnosis using Artificial Neural Network Models ", IEEE Spectrum, Aug 2010.

[30] Mihir Borkar, Prof. Khushali Deulkar, Abhinav Garg, "Prediction of Breast Cancer Using Artificial Neural Netwroks", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 4 Issue 09, Sep 2015.

[31] M.Deepika*,L.Mary Gladence, R.Madhu Keerthana, "A review on prediction of Breast Cancer using various Datamining Techniques", Research Journal Of Pharmaceutical, Biological and Clinical Sciences (RJPBCS), ISSN : 0975-8585, Jan- Feb 2016