2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) 978-1-7281-4142-8/20/$31.00 ©2020 IEEE 10.1109/ic-ETITE47903.2020.36

Analysis of Machine Learning Algorithms on Cancer Dataset

B.Prabadevi
School of Information Technology and Engineering
Vellore Institute of Technology
Vellore, India
prabadevi.b@vit.ac.in

N.Deepa
School of Information Technology and Engineering
Vellore Institute of Technology
Vellore, India
deepa.rajesh@vit.ac.in

Krithika L.B
School of Information Technology ad Engineering
Vellore Institute of Technology
Vellore, India
krithika.lb@vit.ac.in

Vani Vinod
School of Information Technology ad Engineering
Vellore Institute of Technology
Vellore, India
vanipranavam@gmail.com

***Abstract*—The most promising of all cancers that are prevailing among and the primary source of women's deaths worldwide is the cancerous breast cells. Accurate discovery of this type of cancer cells is essential in its early stages, which can be attained via. various data mining and machine learning techniques. Therefore, a comparative analysis among different machine learning techniques such as Random Forest, Support Vector Machine, Naive Bayes, Decision Tree, Neural Networks and Logistic Regression is conducted. It is determined using the WEKA tool. Also, the selected machine learning algorithms are evaluated based on accuracy in prediction results and performance comparison of each classifier with a ROC curve on multiple classifiers is performed.**

Keywords— Breast Cancer cells, classification, Logical Regression, Artificial Neural Network, Weka, Support Vector Machine

I. Introduction

Breast cancerous cells are the most typical form of cancer found in female of age greater than 40. Of the two types of breast cancer namely, malignant and benign, the malignant tumour is more vital. Although scientists are unclear about the exact causes for most cancerous cells prevailing today, they cognise the risk factors that increase the tendency of a woman's body developing cancerous breast cells. The attributes such as genetic risk, age and family history were included in the factors. With early diagnosis of these cancerous cells, about 97% of women survive for more than five years. Also, the death toll due to this disease has increased significantly in the last few decades.

The boundless use of computers has led to the development of massive data stores from a wide variety of heterogeneous sources in different application areas. These repositories will be contributing more significantly to current and future decision making by smearing appropriate knowledge discovery mechanisms.

Data mining is used for extracting useful information and knowledge from raw information. The free statistical tools along with the computational technologies embedded in them were applied to find the useful patterns hidden in the heterogenic dataset. To determine these trends and patterns hidden in the heterogenic dataset, Data Mining uses an integration of sophisticated analytical skills, domain knowledge and an explicit knowledge base. This will enable data analysts to generate more new predictions (observations) from the existing data. However, these techniques do not tend to replace existing statistics; instead, it is just an extension of traditional techniques. Several models have been developed using data mining (DM) and machine learning (ML)algorithms in different domains [1-7]. The objective is to analyse the performance of naïve Bayesian, logistic regression, J48 algorithm, random forest, Support Vector Machine (SVM) and Artificial Neural Network (ANN) classifiers.

II. Related Work

Although various research had carried out on the survival prediction using statistical methods and artificial neural networks, only a few studies related to health care issues and its recurrence using decision trees were found. Delen et al., have applied logistic regression (LR), Artificial Neural Network(ANN), decision trees (DT) for developing the prediction models for breast cancer survival by analyzing a huge dataset namely the SEER database[8]. Lundin et al. have applied ANN and logistic regression models for predicting five, ten, and fifteen-year breast cancer survival. They conducted this study on 951 breast cancer patients. The generic parameters of cancerous cells like age, histological type, tumour size, axillary nodal status, nuclear pleomorphism, tubule formation, mitotic count and tumour necrosis were taken as the input variables[9]. An analysis on the prediction of breast cancer survival rate using conventional data mining techniques like C 4.5 decision tree algorithm, Naïve Bayes (NB) and the back-propagated neural network (NN) algorithm using the Weka toolkit was done[10]. Kate et al. had investigated the various ML models, trained the models and evaluated them for predicting breast cancer survival rate at different stages. They have also designed different breast cancer survivability prediction models at each stage separately and have distinguished them with the traditional integrated models designed generically for all stages of breast cancer. They also evaluated the models separately based on their deployment strategy[11]. Sahar A. Mokhtar [12] have analysed three different models viz., ANN, the decision tree classifier and SVM approach for the prediction of the breast masses severity. The statistical measures, gain and ROC charts were used for evaluating their performances. As a result, the SVM model outperformed the others. Malla et al. used WEKA tool on selected ML algorithms and the best-in-breed ML algorithms were used for the selection of the best classifier model for the early diagnosis of breast cancer. Also, they have implemented three different models on the Breast Cancer dataset [13]. Humayun et al., presented the various application of DM approaches in this healthcare sector. They have specified the different challenges posed by DM techniques in the applications used [14]. Lekha, K. Chitra, and S. Prakasam have evaluated the performance of different classifiers namely J48, BayesNet, RandomForest, Logistic classifiers [15]. V.Krishnaiah developed a model for lung cancer disease prediction using DM classifiers. Naïve Bayes, Decision Trees and Neural Network were found to be the most effective models for the prediction. In turn, Lung Cancer Disease was predicted accurately by Naïve Bayes and fared better than Decision Trees [16]. Arutchelvan, K., and R. Periyasamy produced a multi-layered method obtained by a combination of clustering and decision tree technique. Thus resulting in a cancer-risk prediction system. It can also be used for prediction of other diseases like breast, lung, oral, cervix, pelvic, colic, stomach and blood cancers. They applied data mining techniques to identify probable cancer patients[17]. Shiny, K. et al., carried out the prediction of the breast cancer disease through ANN, NB techniques and Logistic regression [18]. The major aim was to evaluate the medical data set based on quality grammatically and evaluating data mining methods concerning their applicability to the data. Kumar et al., after an extensive analysis, suggested that breast cancer at its early stage can be predicted using different data mining techniques and can reduce the risk of death[19]. Jain et al. used some tools which gave impressive results as when RapidMiner was used to build an SVM classifier and achieved 80% accuracy [20]. Sivakami, K., and Nadar Saraswathi built a hybrid classifier (a mix of SVM and Decision Trees) using WEKA resulting in 91% accuracy[21]. Nauck et al. had applied the supervised fuzzy clustering technique and reported the accuracy of 95.57%[22]. Leena Vig had analysed Random Forest classifiers, Naïve Bayes, SVMs and Artificial Neural Networks. The results obtained proved that ANNs, Random Forests (RF) and SVMs can yield models with high accuracy measure, specificity and sensitivity measures whereas NB perform defectly[23]. Kathija and Shajun Nisha performed 10-fold cross-validation to ensure accuracy in breast cancer cell classification as either benign or malignant on minimum features taken from Wisconsin Diagnosis Breast Cancer dataset. For classification, they used SVM and Naïve Bayes classifiers[24]. The NB classifier was applied on the Wisconsin Prognostic Breast Cancer dataset with 198 patients' record and a binary decision class containing recurrent events and non-recurrent events with 47 instances and 151 instances respectively. The accuracy of test diagnosis was 74.24% when compared to other well-known techniques[25].

A. Soltani Sarvestani et al. carried out a comparative analysis of the neural networks such as Radial Basis Function(RBF), Self Organizing Map, Multilayer Perceptron and Probabilistic Neural Network[26]. Gayathri, B. M., and C. P. Sumathi used Naive Bayes classifier and yielded an accuracy of 96.6% [27]. Bin Othman et al., compared different classification techniques like RBF, Nearest Neighbors, Pruned Tree algorithm and Bayes Network through WEKA on the largest breast cancer dataset consisting of about 6000 data with a dimension of 699 x 9 They used 75%-25% of overall data for training and testing respectively. On simulation, the attained the highest accuracy by the Bayes network [28]. Kim et al. constructed a forecasting model based on SVM to predict breast cancer recurrence after surgery (within five years) in the Korean population [29]. Li et al., implemented an enhanced rank-based technique to extract different pairs of CpG sites. They have extracted the pairs with reversal relative DNA methylation levels in the diseased samples and pairs in healthy samples for five different cancer types. The former pairs resulted in the maximum accuracy above 95% for each type of cancer [30]. Shrivastava et al. discuss the various classification techniques such as KNN, SVM, DT and neural network. They have done the classification depending on the feature attribute values in the dataset. As the feature selection and data optimisation is an essential aspect to improve the classification process they have focussed more on that [31]. Mandal and Subrata Kumar aimed to find the minimum subset of features that could guarantee a precise classification of breast cancer. They showed a comparative analysis of different cancer classification models viz., NB, DT and LR classifiers. Logistic Regression (LR) classifier outperforms with the highest accuracy than the other two classifiers [32].

From the above survey, it is evident that data mining and machine learning could be an appropriate approach in identifying patterns in various breast cancer cases. This can be applied for different stages of diagnosing breast cancer, prognosis breast cancer, and its treatment purposes.

III. ANALYSIS METHOD

A large volume of health records was being collected and are made available for research purpose. Therefore, the major aim of the DM process is to transform the data retrieved from the repository into a more meaningful form. It involves the following:

- Data preprocessing: transforms raw data into an understandable form. The data goes through a series of steps during the time of preprocessing:
  - Data cleaning:-Noise removal.
  - Data integration:-Combines heterogeneous data.
  - Data selection:- Appropriate data for analysis is chosen.
  - Data transformation:- appropriate data conversion for the mining procedure.
  - Data mining:- extract patterns suitable for analysis.
  - Pattern evaluation:- Exciting data patterns are evaluated.

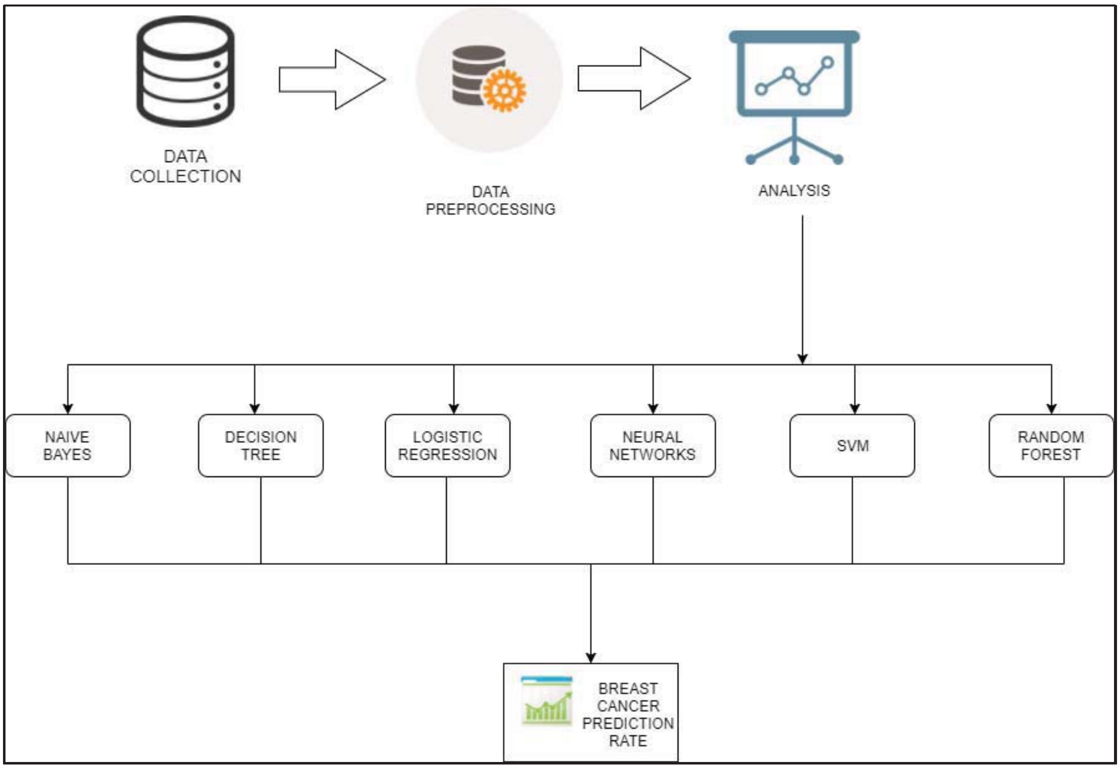


Fig. 1. System Design

Breast Cancer Prediction and Prognosis:
1.The breast_cancer_susceptibility prediction (Diagnosis)
2. The breast_cancer_recurrence prediction (Prognosis)
3. The breast_cancer_survivability rate prediction (Prognosis).

The success of Prognosis prediction is dependent on the quality of the Diagnosis. Six types of classifiers were applied to the Breast Cancer dataset to evaluate the performance of the following ML techniques:

Naïve Bayes: is based on the Bayes theorem. Bayes theorem considers independence assumptions between predictors. It assumes the presence of a specific feature in a class which is not related to the presence of any other feature in the class.

Decision Tree: It determines the target value y of the given test data based on the attribute values of the trained data X available. The internal nodes in the decision tree denote the different attributes, whereas the branches between the nodes denote the possible values the attributes can take and the leaf nodes tell the final value of the target. The *predicted attribute* is thee target since its value depends on the other attributes' values while remaining are the independent variables. *Gini index and information gain* are popular attribute selection measures.

Logistic Regression (LR) is generally preferred for predicting the probability of an outcome with only two values. The predictions are based on one or more numerical and categorical predictors. Logistic regression is preferred for describing records and illuminating the relationship between one or several independent variables that can decide an outcome. It is measured using dichotomous variables, which has only two possible outcomes.

The logit function used is $Z = \log(p/1-p)$

LR model receives real-valued inputs and predicts the probability of the input taken belonging to the default class viz., class 0. If the probability, $p> 0.5$, then output is considered to be class 0 (a prediction for the default class). Otherwise, class 1 (the prediction is for the other class).

Random Forest: A bagging algorithm based on decision trees. It grows multiple trees used for classifying a new entity based on the attributes. Then each tree denotes a particular classification and in turn, that tree "votes" for its class. Then it chooses the class with the highest votes, and in case of regression, the mean outputs by different trees are considered.

ANNs:

ANN is the simulations inspired biologically on the computer to perform specific tasks like clustering, classification, pattern recognition etc. ANN is for detecting complex nonlinear relationships in data. Multilayer perceptron networks (MLPs) classifier is used. It is a kind of multilayer networks learned by the backpropagation algorithm are capable of expressing a wide variety of nonlinear decision surfaces. Like other machine learning methods, a wide variety of tasks have been solved using neural networks.

SVM:

A supervised algorithm commonly preferred for both classification and regression. It plots each data item as a point in n-dimensional space and the value of each feature will be the value of a particular coordinate point. Here n is the total number of features. It constructs a set of hyperplanes that maximize the margin between two classes for classification. The classifer SVM is implemented using a kernel. The learning of the hyperplane is done using some linear algebra.

.

IV. DATASET DESCRIPTION

The dataset was obtained from Kaggle (a platform for predictive modelling). Thirty-two attributes are there in the dataset as specified in Table I. The features are computed from a breast mass. The digitised image of a FNA (fine needle aspirate) is considered. It is a straightforward and fastest procedure to perform. It works by removing some fluid or cells from a breast lesion/cyst with a fine needle similar to a blood sample needle.

TABLE I. DATASET DESCRIPTION

S.No	Parameter
1.	Sample ID
2.	Diagnosis Type (M refers to malignant, B refers to benign)
3-32.	**Attributes of 10 real-valued features that are computed for each cell nucleus:**
	i. perimeter
	ii. Radius
	iii. Area
	iv. Texture
	v. Smoothness
	vi. Symmetry
	vii. Concavity
	viii. Concave points
	ix. Compactness
	x. fractal dimension

V. RESULTS AND DISCUSSION

WEKA, a data mining tool is used for implementing the classifiers. Pre-processing components in WEKA are called “filters”. WEKA contains filters for discretization, normalisation, resampling, attribute selection, transformation and combination of attributes. There are a lot of different filters like supervised and unsupervised. Supervised filters use a class value for their operation. They are not so common as unsupervised filters, which do not use the class value. There are attribute filters and instance filters. Model Performance Chart is a component from the Visualization component of Knowledge flow in WEKA which is used for visualising threshold, i.e. ROC curves for multiple classifiers to ease the comparison. Table II and Table III provide the comparison of various parameters on Breast cancer data set before applying filters and after applying filters. The figures Fig. 2 through Fig. 30 depicts the results obtained stepwise.

A. Applying Naive Bayes, without using filters

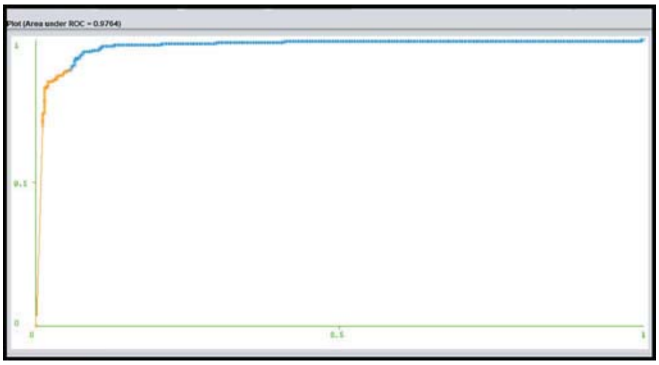


Fig. 2. ROC curve - Naive Bayes Classifier (Area under ROC=0.9764)

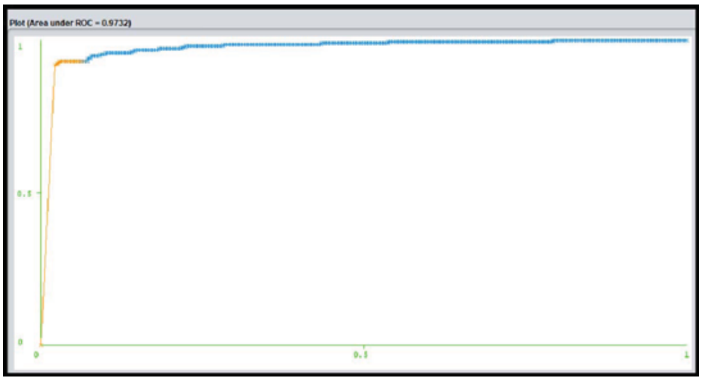


Fig. 3. Fig. 5. ROC curve - Logistic Regression (Area under ROC=0.9732)

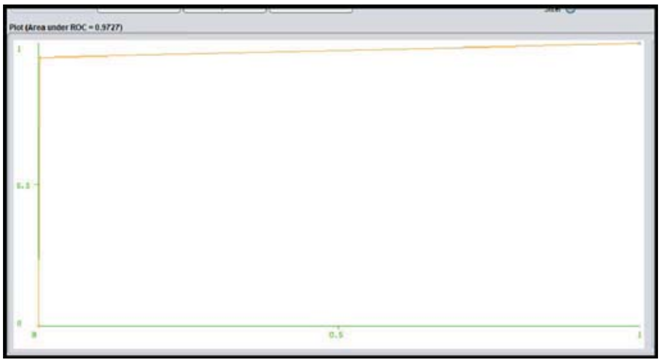


Fig. 4. ROC curve - SVM Classifier (Area under ROC=0.9727)

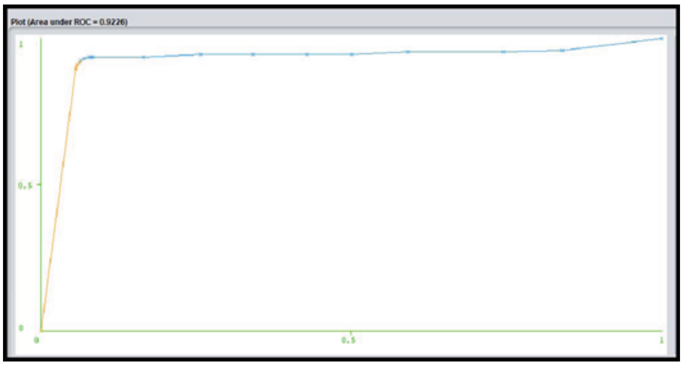


Fig. 5. ROC curve - Decision Tree Classifier (Area under ROC=0.9226)

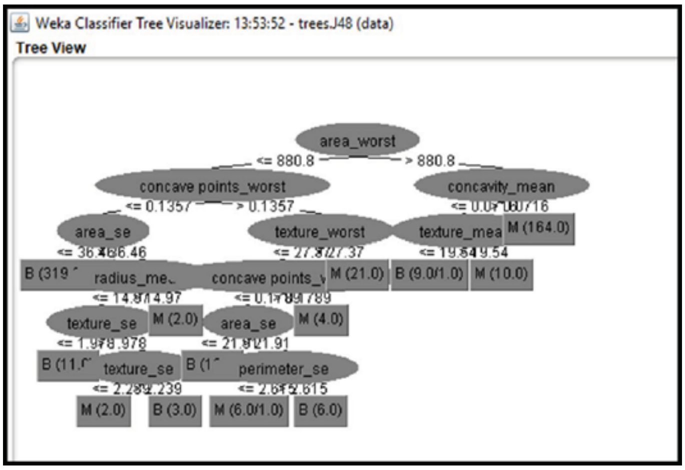


Fig. 6. Decision tree structure

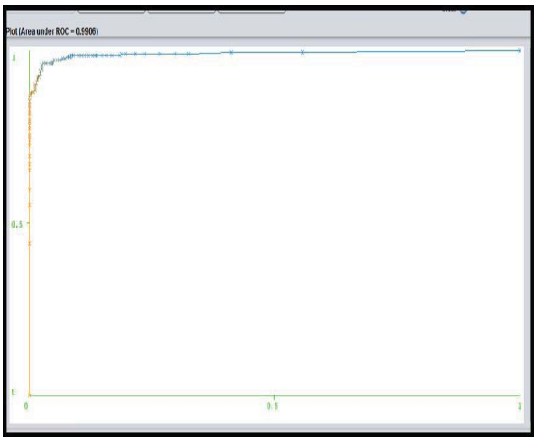


Fig. 7. ROC curve - the Random Forest Classifier (Area under ROC=0.9906)

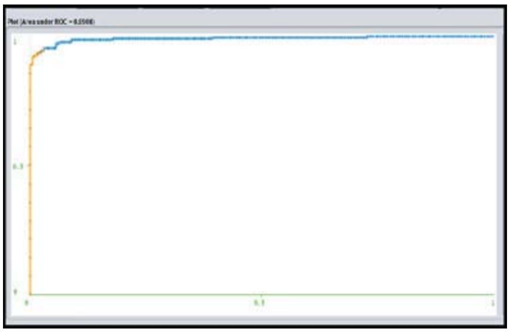


Fig. 8. ROC curve for the ANN Classifier (Area under ROC=0.9908)

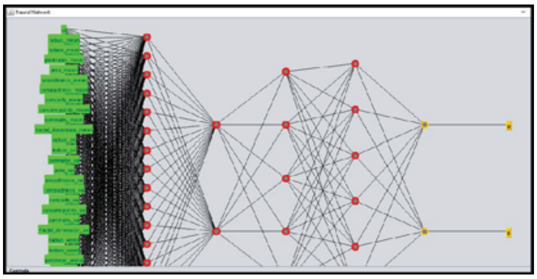


Fig. 9. ANN structure with four hidden layers

TABLE II. COMPARISON OF VARIOUS PARAMETERS ON BREAST CANCER DATA SET BEFORE APPLYING FILTERS

Performance matrix	Naïve Bayes	Decision Tree	Logistic Regression	Random Forest	SVM	ANN
Time	0.01	0.05	0.13	0.14	0.02	3.18
Kappa Statistics	0.8418	0.8502	0.8582	0.9128	0.9545	0.9135
MAE	0.0732	0.0758	0.0663	0.0757	0.0211	0.0387
RMSE	0.2648	0.2608	0.2533	0.1731	0.1452	0.1827
RAE(%)	15.6565	16.2103	14.1764	16.1855	4.5095	8.2798
RRSE(%)	54.7597	53.9333	52.3086	35.8076	30.035	37.7817
Accuracy=(TP+TN)/(TP+FP+TN+FN) (%)	92.6186	92.9701	93.3216	95.9578	97.891	95.9578
Sensitivity=TP/TP+FN (%)	90.476	89.81	89.54	96.55	99.50	94.78
Specificity=TN/TN+FP (%)	93.87	94.9	95.7	95.62	97	96.64

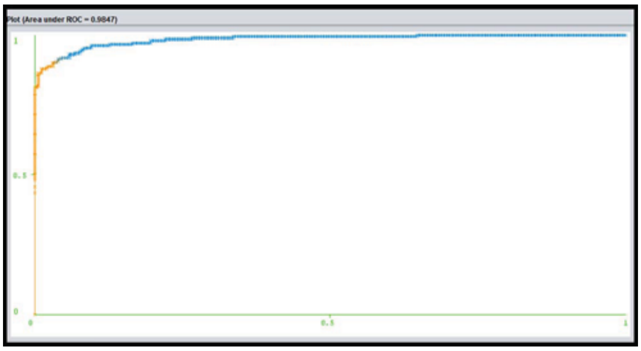


Fig. 10. ROC curve for Naive Bayes Classifier with filters (Area under ROC=0.9847)

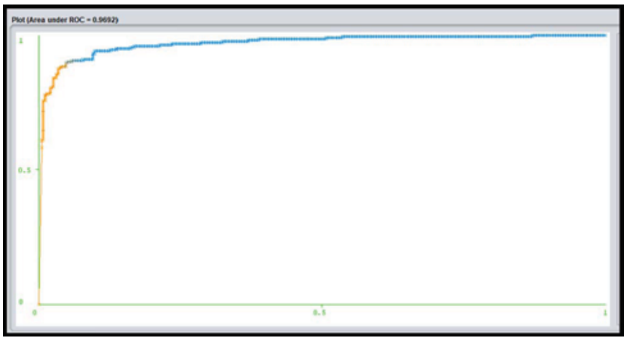


Fig. 11. ROC curve for the Logistic Regression with filters (Area under ROC=0.9692)

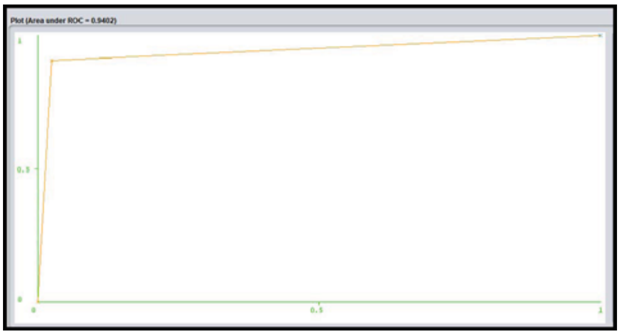


Fig. 12. ROC curve for the SVM Classifier with filters(Area under ROC=0.9402)

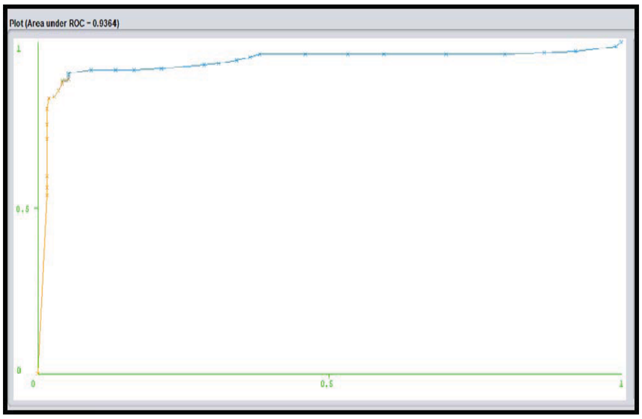
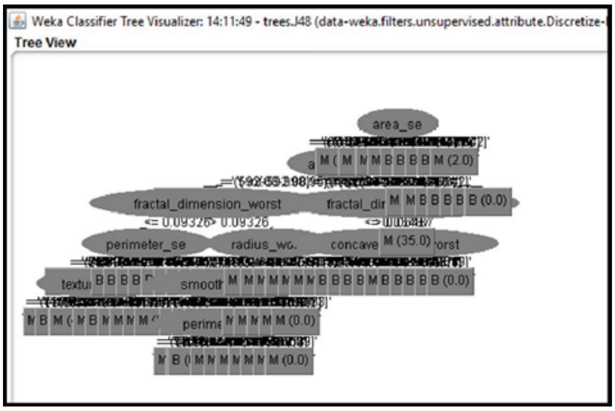


Fig. 13. ROC curve for the Decision Tree Classifier with filters(Area under ROC=0.9364)

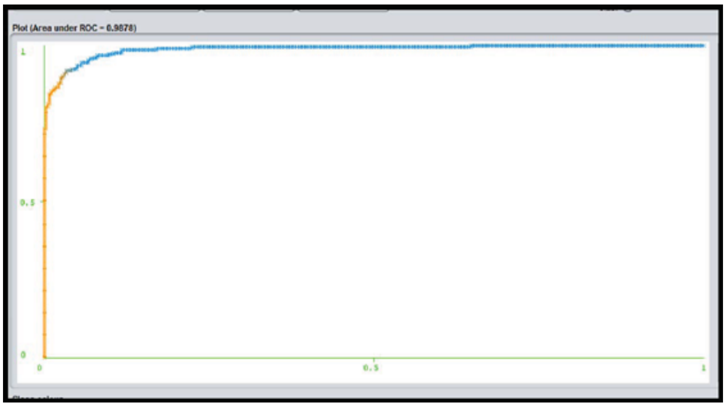


Weka Classifier Tree Visualizer: 14:11:49 - trees.J48 (data-weka.filters.unsupervised.attribute.Discretize-

Tree View

Fig. 14. Decision tree structure

Fig. 15. ROC curve - Random Forest Classifier with filters(Area under ROC=0.9871)



Plot (Area under ROC = 0.9878)

Fig. 16. ROC curve -ANN Classifier with filters(Area under ROC=0.9878)

TABLE III. COMPARISON OF VARIOUS PARAMETERS ON BREAST CANCER DATA SET AFTER APPLYING UNSUPERVISED DISCRETIZATION FILTERS.

Performance metrics	NB	DT	LR	RF	SVM	ANN
Time	0	0.03	0.72	0.32	0.14	271.24
Kappa Statistics	0.8715	0.8376	0.8488	0.9093	0.8898	0.8827
MAE	0.0599	0.0994	0.0738	0.1004	0.051	0.0568
RMSE	0.2323	0.2508	0.2579	0.1963	0.2258	0.2912
RAE(%)	12.810	21.2597	15.7807	21.472	10.898	12.143
RRSE(%)	48.035	51.8789	53.3457	40.6059	46.692	45.3314
Accuracy=(TP+TN)/(TP+FP+TN+FN) (%)	94.0246	92.4429	92.9701	95.7821	94.9033	94.5518
Sensitivity=TP/TP+FN (%)	93.20	90.82	91.74	95.63	95.52	94.14
Specificity=TN/TN+FP (%)	94.49	93.37	93.66	95.867	94.56	94.78

B. COMPARING CLASSIFIERS : KNOWLEDGE FLOW

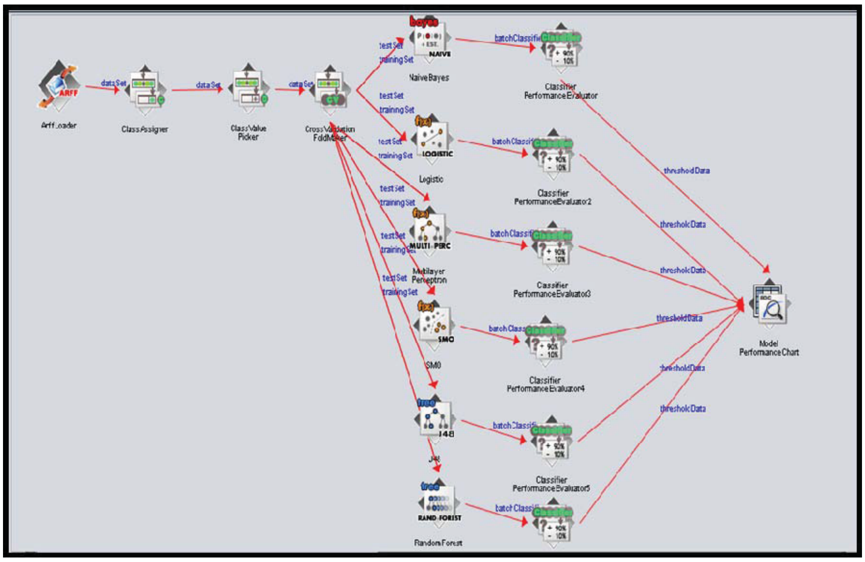


Fig. 17. Performance model chart construction

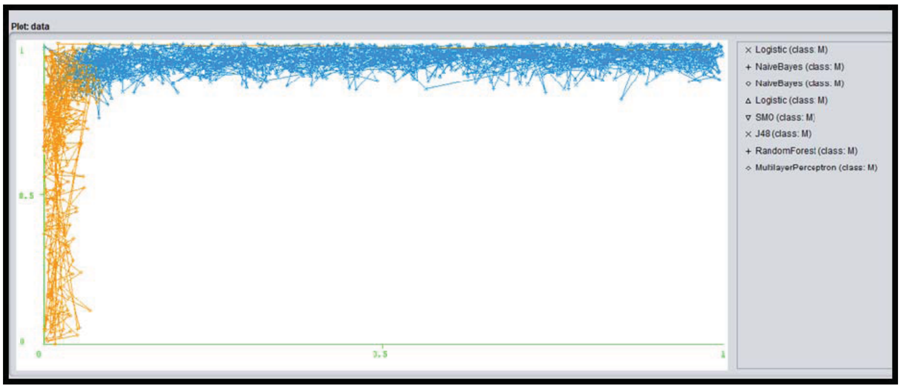


Fig. 18. Model Performance Chart

VI. STATISTICAL ANALYSIS

Breast cancer tops among various diseases in the female population in India. The average age adjusted rate as high as 25.8 and mortality 12.7 per 100,000 women were noted. The death rates due to this cancer type have increased by 0.4% per year from 1975 to 1989, but since have decreased swiftly, for a total deterioration of 39% through 2015 [33]. In turn, 3lakh breast cancer deaths have been stopped. The decrease rate occurred includes women in both younger and older age group. However, it has reduced among women group less than 50 years of age since 2007. The breast cancer death rates were declined by 2.6% annually from 2006 through 2015. This weakening in breast cancer mortality is mainly due to advancement in treatments and early detection/prediction of the disease.

Relative survival rates of cancer patients after diagnosis is determined by considering the normal people of the same age without cancer.

In 2017, around 2 lakh new cases of breast cancer were diagnosed in women, and over 2000 cases were diagnosed in men. Also, 63,410 cases were diagnosed to be Ductal cancer in women. The main issue about its cure is early detection. This requires making a shift from the traditional cure based systems to the prevention-based system. Which is possible only if we belligerently work for the early detection and prediction of the same, using data mining tools in identifying patterns in breast cancer cases, results in an accuracy rate above 90% which can be used for breast cancer prediction in all stages thus helping in combating the mortality rate.

VII. CONCLUSION

Of all type of cancers prevailing among women, breast cancer is one of the noteworthy causes of their death. Therefore, the early detection of this cancerous cells is desirable in declining the mortality rate due to the same. An analytical evaluation of six classifiers –NB, DT, LR, RF, ANN and SVM which were applied on the Breast Cancer dataset by using the open-source tool WEKA, has been produced. Preprocessing is done on the input dataset by applying certain WEKA inbuilt attribute filters, and the effect of cleaned data on the accuracy of the prediction was also noted down. The results shown that the performance of SVM classifier as the best which gave an accuracy rate of 97.89% following ANN and Random Forest which had the second-best accuracy rate of 95.95%, Logistic Regression with an accuracy rate of 93.32%, DT with an accuracy rate of 92.97% and NB with an accuracy rate of 92.61%. Thus SVM classifier proves to be the best classifier among the six others without the addition of an unsupervised filter. With the application of unsupervised discretise filters on the dataset, it resulted in the accuracy rate as follows: Random Forest - 95.78%, SVM -94.90%, ANN-94.55%, Naive Bayes-94.02%, Logistic Regression-92.97% and Decision Tree -92.44%. Thus, Random Forest proves to be the best classifier among the six with the addition of filters to the dataset.

REFERENCES

[1] Ramaiah, M., Baranwal, P., Shastri, S. B., Vanitha, M., & Vanmathi, C. Analytical Comparison of Machine Learning Techniques for Liver Dataset. In *2019 Innovations in Power and Advanced Computing Technologies (i-PACT)* (Vol. 1, pp. 1-5). IEEE, 2019.

[2] Deepa, N. and Ganesan, K., Multi-class classification using hybrid soft decision model for agriculture crop selection. Neural Computing and Applications, Neural Computing and Applications, Vol. 3, Issue 4, pp.1025-1038, 2018

[3] Deepa, N., & Ganesan, K. Decision-making tool for crop selection for agriculture development. Neural Computing and Applications, Vol. 31 Issue 4, pp: 1215-1225, 2019.

[4] Deepa N, Ganesan K, Sethuramasamyraja B. Predictive mathematical model for solving multi-criteria decision-making problems. Neural Computing and Applications, Vol. 31, Issue 10, pp.6733-6746, 2019.

[5] Deepa, N., and Ganesan, K., Hybrid Rough Fuzzy Soft classifier based Multi-Class classification model for Agriculture crop selection. Soft computing, Vol. 23, Issue 21, pp.10793-10809, 2019.

[6] Deepa, N., Ganesan, K., Srinivasan, K., & Chang, C. Y. Realizing Sustainable Development via Modified Integrated Weighting MCDM Model for Ranking Agrarian Dataset. Sustainability, Vol. 11, Issue 21, 6060, 2019.

[7] Deepa, N., Srinivasan, K., Chang, C. Y., & Bashir, A. K. An efficient ensemble vtopes multi-criteria decision-making model for sustainable sugarcane farms. Sustainability, Vol.11, Issue 16, 4288, 2019.

[8] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989 Delen, Dursun, Glenn Walker, and Amit Kadam. "Predicting breast cancer survivability: a comparison of three data mining methods." Artificial intelligence in medicine, Vol. 34, no. 2, pp. 113-127, 2005.

[9] Breast cancer Q&A/facts and statistics (http://www.komen.org/bci/bhealth/QA/q_and_a.asp).

[10] Bellaachia, Abdelghani, and Erhan Guven. "Predicting breast cancer survivability using data mining techniques." Age 58, no. 13 , pp. 10-110, 2006.

[11] Kate, Rohit J., and Ramya Nadig. "Stage-specific predictive models for breast cancer survivability." International Journal of Medical Informatics, Vol. 97, pp: 304-311, 2017.

[12] Mokhtar, Sahar A., and Alaa Elsayad. "Predicting the severity of breast masses with data mining methods." arXiv preprint arXiv:1305.7057 (2013).

[13] Malla, Younus Ahmad, and Mohammad Ubaidullah Bokari. "A machine learning approach for early prediction of breast cancer." International Journal Of Engineering And Computer Science, Vol. 6, no. 7, pp. , 2017.

[14] Humayun, Ahsan, and Adeel Waqar. "A Comparative Study on Usage of Data Mining Techniques in Healthcare Sector." International Journal of Computer Applications, Vol. 162, no. 6, pp.8887, 2017.

[15] Lekha, K. Chitra, and S. Prakasam. "Performance Assessment of Different Classification Techniques." Data Mining and Knowledge Engineering, vol. 9, no. 1 pp. 20-23, 2017.

[16] Krishnaiah, V., Dr G. Narsimha, and Dr N. Subhash Chandra. "Diagnosis of lung cancer prediction system using data mining classification techniques." International Journal of Computer Science and Information Technologies, vol. 4, no. 1, pp. 39-45,2013.

[17] Arutchelvan, K., and R. Periyasamy. "Cancer Prediction System Using Datamining Techniques." International Research Journal of Engineering and Technology, Vol.2, no.8, pp.1179-1183, 2015.

[18] Shiny, K., M. Swaminathan, N. Siva Kumar, L. Thiyagarajan, K. Shiny, M. Swaminathan, N. Siva Kumar, and L. Thiyagarajan. "Implementation of Data Mining Algorithm to Analysis Breast Cancer." International Journal for Innovative Research in Science and Technology, vol.1, no. 9, pp. 207-212, 2015.

[19] Kumar, Murari, Shivkumar Singh Tomar, and Bhupesh Gaur. "Mining based Optimization for Breast Cancer Analysis: A Review." International Journal of Computer Applications, vol.119, no. 13 , 2015.

[20] Jain, Priyanka, and Santosh Kr Vishwakarma. "Collaborative Analysis of Cancer Patient Data using Rapid Miner.", International Journal of Computer Applications, vol. 145, no. 2, 2016.

[21] Sivakami, K., and Nadar Saraswathi. "Mining Big Data: Breast Cancer Prediction using DT-SVM Hybrid Model." International Journal of Scientific Engineering and Applied Science (IJSEAS), vol. 1, no. 5, pp.418-429, 2015.

[22] Nauck, Detlef, and Rudolf Kruse. "Obtaining interpretable fuzzy classification rules from medical data." Artificial intelligence in medicine, vol.16, no. 2, pp.149-169,2015.

[23] Vig, Leena. "Comparative Analysis of Different Classifiers for the Wisconsin Breast Cancer Dataset." Open Access Library Journal vol.1, no. 06,pp.1,2014.

[24] Kathija, Shajun Nisha ,"Breast Cancer Data Classification Using SVM and Naïve Bayes Techniques", International Journal of Innovative Research in Computer and Communication Engineering-Vol. 4, Issue 12, December 2016.

[25] Dumitru, Diana. "Prediction of recurrent events in breast cancer using the Naive Bayesian classification." Annals of the University of Craiova-Mathematics and Computer Science Series 36, no. 2, pp.92-96, 2009.

[26] Sarvestani, A. Soltani, A. A. Safavi, N. M. Parandeh, and M. Salehi. "Predicting Breast Cancer Survivability using data mining techniques." In Software technology and Engineering (ICSTE), 2010 2nd international Conference on, vol. 2, pp. V2-227. IEEE, 2010.

[27] Gayathri, B. M., and C. P. Sumathi. "An Automated Technique using Gaussian Naïve Bayes Classifier to Classify Breast Cancer." International Journal of Computer Applications 148, no. 6, 2016.

[28] Bin Othman, Mohd Fauzi, and Thomas Moh Shan Yau. "Comparison of different classification techniques using WEKA for breast cancer." In 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006, pp. 520-523. Springer, Berlin, Heidelberg, 2007.

[29] Kim, Woojae, Ku Sang Kim, Jeong Eon Lee, Dong-Young Noh, Sung-Won Kim, Yong Sik Jung, Man Young Park, and Rae Woong Park. "Development of novel breast cancer recurrence prediction model using support vector machine." Journal of breast cancer 15, no. 2, pp.230-238, 2012.

[30] Li, Hongdong, Guini Hong, and Zheng Guo. "Reversal DNA methylation patterns for cancer diagnosis." In Systems Biology (ISB), 2014 8th International Conference on, pp. 101-106. IEEE, 2014.

[31] Shrivastava, Shiv Shakti, V. K. Choubey, and Anjali Sant. "Classification Based Pattern Analysis on the Medical Data in Health Care Environment." International Journal of Scientific Research in Science, Engineering and Technology 2, no. 1, 2016.

[32] Mandal, Subrata Kumar. "Performance Analysis Of Data Mining Algorithms For Breast Cancer Cell Detection Using Naïve Bayes, Logistic Regression and Decision Tree." International Journal Of Engineering And Computer Science 6, no. 2, 2017.

[33] American Cancer Society. "Breast Cancer Facts & Figures 2017-2018". Atlanta: American Cancer Society, Inc., 2017.