Breast Cancer Analysis Using Random Forest Classifier

K.B.Nawveen[1], D.Mikeakshay[2], L.Vignesh[3], N.Jeyaganesh[4]
[1, 2 ,3 ,4] Dept of EEE
[1, 2, 3, 4] Anna University Regional Campus Coimbatore

Abstract- *Breast cancer is a mortiferous malady that develops from breast tissue. Signs of breast cancer may include a lump in the breast, a change in breast shape, dimpling of the skin, fluid coming from the nipple, or a red scaly patch of skin. Cells are the building blocks for the organs and tissues in the body. When the growth of new cells is uncontrolled then they build-up mass of tissue called a tumor. The tumors are categorized into benign and malignant tumors. In the proposed work ,we compared the predictive accuracy of the classification of tumor cells using Random forest classifier, Support vector machine, and Decision tree classifier and affirming that Random forest Classifier algorithm is the best predictor for the classification of tumor cells. Random forest classifier is an ensemble learning method for classification, regression, and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class. Random forest classifier correct for decision trees' habit of overfitting to their training set.*

I. INTRODUCTION

Breast cancer develops from breast tissue. Signs of breast cancer may include a lump in the breast, a change in breast shape, dimpling of the skin, fluid coming from the9999 nipple, or a red scaly patch of skin

In those with distant spread of the disease, there may be bone pain, swollen lymph nodes, shortness of breath, or yellow skin Risk factors for developing breast cancer include being female, obesity, lack of physical exercise, drinking alcohol, hormone replacement therapy during menopause, ionizing radiation, early age at first menstruation, having children late or not at all, older age, and family history. About 5–10% of cases are due to genes inherited from a person's parents, including BRCA1 and BRCA2 among others. Breast cancer most commonly develops in cells from the lining of milk ducts and the lobules that supply the ducts with milk. Cancers developing from the ducts are known as ductal carcinomas, while those developing from lobules are known as lobular carcinomas. In addition, there are more than 18 other sub-types ofbreast cancer. Some cancers, such as ductal carcinoma in situ, develop from pre-invasive lesions.

The diagnosis of breast cancer is confirmed by taking a biopsyof the concerning lump. Once the diagnosis is made, further tests are done to determine if the cancer has spread beyond the breast and which treatments it may respond to.
Nowadays data mining application has been increased in medical field. There are a few arguments that can support the use of datamining in health sector for breast cancer like early classification like data mining, artificial intelligence, neural networks for the better accuracy in the diagnosis of breast cancer. Data mining classification algorithms are used on large set of breast cancer data to classify whether the cell is benign or malignant. Even though some of the papers included several individual algorithms for the correct classification of the benign or malignant tumor.

We have taken the three classification algorithms such as Random forest classifier, Support vector machine, and Decision tree classifier trained with dataset individually and affirming that Random forest Classifier algorithm is the best predictor for the classification of tumor cells.

Random forest classifier is an ensemble learning method for classification, regression, and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class. Random forest classifier correct for decision trees' habit of overfitting to their training set. Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance. This comes at the expense of a small increase in the bias and some loss of interpretability, but generally greatly boosts the performance in the final model.

We have make use of correlation graph so that we can remove multi colinearity it means the attributes are dependenig on each other. So we should avoid it because what is the use of using same attributes twice .Therefore , attributes that do not support in increasing the accuracy can be eliminated .In addition n-fold cross vali--dation is used which is used for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set). To reduce variability, in most methods multiple rounds of cross -validation are performed using different partitions, and the validation results are combined (e.g. averaged) over the rounds to estimate a final predictive model.

II. EXISTING SYSTEM

The existing system compares the performance of three classification algorithms and their combination using ensemble approach that are suitable for direct interpretability of their results. This system inspected the generalization performance of J48, Naïve Bayes, and SVM in order to boost the prediction models for decision-making system in the prediction of breast cancer survivability. They are using a Voting classifier approach where all three classification algorithms are combined for the prediction of breast cancer.

In the existing system the best accuracy of prediction isAchieved by combining three clasification algorithms together into single classifer called VOTING CLASSIFIER. Voting algorithm cannot process large datasets with higher dimensionality and it also suffers overfitting.

Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. Voting algorithm cannot process large datasets with higher dimensionality.

III. LITERATURE SURVEY

1)Salama Gouda I., M. B. Abdelhalim, and MagdyAbd-elghanyZeid. "Experimental comparison of classifiers for breast cancer diagnosis." Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on. IEEE, 2012.

Salama, Gouda I., M. B. Abdelhalim, and MagdyAbd-elghanyZeid compare different classifiers decision tree, Multi-Layer Perception, Naive Bayes, Sequential Minimal Optimization, and Instance-Based for K-Nearest neighbor on three different databases of breast cancer (Wisconsin Breast Cancer (WBC), Wisconsin Diagnosis Breast Cancer (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC)) by using classification accuracy and confusion matrix based on 10-fold cross validation method. And also combine classifiers to get better accuracy, the experimental results show that in the classification using fusion of MLP and J48 with the PCA is superior to the other classifiers using WBC data set.

2.GhoshSoumadip, SujoyMondal, and Bhaskar Ghosh. "A comparative study of breast cancer detection based on SVM and MLP BPN classifier." Automation, Control, Energy and Systems (ACES), 2014 First International Conference on. IEEE, 2014.

Ghosh, Soumadip, SujoyMondal, and Bhaskar Ghosh applied different classification techniques namely, MLP using Backpropagation Neural Network and Support Vector Machine on Breast Cancer Wisconsin dataset from the UCI machine language repository for detection of breast cancer. The author concluded that SVM classifier has the potential to significantly improve the conventional classification methods for use in medical or in general, Bioinformatics field.

3.Elouedi Hind, et al. "A hybrid approach based on decision trees and clustering for breast cancer classification." Soft Computing and Pattern Recognition (SoCPaR), 2014 6th International Conference of. IEEE, 2014

Elouedi, Hind, et al proposed a hybrid diagnosis approach of breast cancer based on decision trees and clustering. First they have done Extraction of the malignant instances and apply K-means algorithm to split the malignant instances. After that they apply decision tree algorithm (C4.5) to every cluster and compare the different accuracies calculated in each case. Combined results of benign and malignant are applied to the C4.5 algorithm to find out the results of classification based on the confusion matrix and the global and detailed accuracy values. They have gotten better results than those obtained with the original dataset.

4)MittalDishant, Dev Gaurav, and SanjibanSekhar Roy. "An effective hybridized classifier for breast cancer diagnosis." 2015 IEEE International Conference on Advanced Intelligent Mechatronics (AIM). IEEE, 2015

Mittal, Dishant, Dev Gaurav, and SanjibanSekhar Roy proposed an effectively hybridized classifier which is made by combining an unsupervised artificial neural network (ANN) method named self-organizing maps (SOM) with a supervised classifier called stochastic gradient descent (SGD) for breast cancer diagnosis. Also they compare there results with three supervised machine learning techniques decision tree (DTs), random forests (RF) and support vector machine (SVM). The hybrid model builds up by integrating stochastic gradient descent with self-organizing maps gave an accuracy of 99.521% over training set and 99.686% over the testing set.

5)Phetlasy, Sornxayya, et al. "Sequential Combination of Two Classifier Algorithms for Binary Classification to Improve the Accuracy." 2015 Third International Symposium on Computing and Networking (CANDAR). IEEE, 2015.

Phetlasy, Sornxayya, et al proposed a0020hybrid method for data classification with two different classifier algorithms. The first classifier responds to reduce FN, and the second classifier is in charge of reducing FP. they implement their experiment with five popular algorithms, two algorithms for one combination so that they obtain 20 cases in total. The five algorithms are Sequential Minimal Optimization (SMO) for SVM, Na茂ve Bayes (NB), decision tree J48, Instancebased learning IBK algorithm for K-Nearest Neighbor, and Multilayer Perceptron (MLP) for Neural Network. The author obtained the best result of 99.13% accuracy.

6)Lavanya D., and K. Usha Rani. "Ensemble decision making system for breast cancer data." International Journal of Computer Applications 51.17 (2012)

Lavanya D., and K. Usha Rani studied a hybrid approach: CART decision tree classifier with feature selection and boosting ensemble method has been considered to evaluate the performance of classifier. They applied CART algorithm, CART with Feature Selection Method and CART with Feature selection and Boosting on different Breast cancer data sets and compared the Accuracy

7)]Sarvestani A. Soltani, et al. "Predicting Breast Cancer Survivability using data mining techniques." Software Technology and Engineering (ICSTE), 2010 2nd International Conference on. Vol. 2. IEEE, 2010

Sarvestani, A. Soltani, et al. , applied self-organizing map(SOM), radial basis function network (RBF), general regression neural network (GRNN) and probabilistic neural network (PNN) are tested on the Wisconsin breast cancer data (WBCD) and on the Shiraz Namazi Hospital breast cancer data (NHBCD). And concluded RBF and PNN were proved as the best classifiers in the training set. However the PNN gives the best classification accuracy when the test set is considered.

8)]Shah Chintan, and Anjali G. Jivani. "Comparison of data mining classification algorithms for breast cancer prediction." Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on. IEEE, 2013.

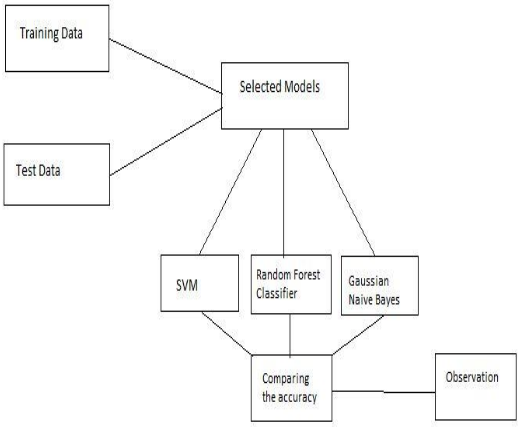
Shah, Chintan, and Anjali G. Jivani conducted the comparison between three algorithms namely Compares Decision tree, Bayesian Network and K-Nearest Neighbor with help of WEKA (The Waikato Environment for Knowledge Analysis), which is an open source software. The author concluded that Naïve Bayes is a superior algorithm compared to the two others because it takes lowest time i.e. 0.02 seconds and at the same time is providing highest accuracy.

9)B. Padmapriya, T. Velmurugan, "A Survey on BreastCancer Analysis Using Data Mining Techniques",Computational Intelligence and Computing Research(ICCIC), 2014 IEEE International Conference on.
IEEE, 2014.

Data mining (DM) comprises the core algorithms that enable to gain fundamental insights and knowledge from massive data. In fact, data mining is a part of a larger knowledge discovery process. One of the new researches in data mining application involves analyzing Breast cancer, which are the deadliest disease and most common of all cancers in the leading cause of cancer deaths in women worldwide. Among the various DM techniques, classification plays a vital role in DM research. Breast cancer diagnosis and prognosis are two medical applications pose a great challenge to the researchers in medical field. This survey work analyses the various review and technical articles on breast cancer diagnosis. The main goal of this research is to explore the overview of the current research being carried out using the data mining techniques to enhance the breast cancer diagnosis. Particularly, this survey discusses about use of the classification algorithms ID3 and C4.5 in breast cancer analysis.

IV. PROPOSED SYSTEM

On Comparison with the voting algorithm, the Random Forest Classifier can handle large data set with higher dimensionlity.It handles the missing values and maintain the accuracy for missing data



.It does both classification and regression tasks. Datasets are divided into training data and test data and are tested for the prediction of breast cancer.Here, by tuning the data sets the accuracy of the prediction increases.The datasets are splitted into three categories and are predicted for the best, normal and the worst case of features.Corresponding Interactive Maps are generated with the help of the diagonisis reports.

REFERENCES

[1] Salama Gouda I., M. B. Abdelhalim, and MagdyAbd-elghanyZeid. "Experimental comparison of classifiers for breast cancer diagnosis." Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on. IEEE, 2012.

[2] GhoshSoumadip, SujoyMondal, and Bhaskar Ghosh. "A comparative study of breast cancer detection based on SVM and MLP BPN classifier." Automation, Control, Energy and Systems (ACES), 2014 First International Conference on. IEEE, 2014

[3] Elouedi Hind, et al. "A hybrid approach based on decision trees and clustering for breast cancer classification." Soft Computing and Pattern Recognition (SoCPaR), 2014 6th International Conference of. IEEE, 2014

[4] MittalDishant, Dev Gaurav, and SanjibanSekhar Roy. "An effective hybridized classifier for breast cancer diagnosis." 2015 IEEE International Conference on Advanced Intelligent Mechatronics (AIM). IEEE, 2015

[5] MittalDishant, Dev Gaurav, and SanjibanSekhar Roy. "An effective hybridized classifier for breast cancer diagnosis." 2015 IEEE International Conference on Advanced Intelligent Mechatronics (AIM). IEEE, 2015

[6] Lavanya D., and K. Usha Rani. "Ensemble decision making system for breast cancer data." International Journal of Computer Applications 51.17 (2012)

[7]]Sarvestani A. Soltani, et al. "Predicting Breast Cancer Survivability using data mining techniques." Software Technology and Engineering (ICSTE), 2010 2nd International Conference on. Vol. 2. IEEE, 2010

[8]]ShahChintan, and Anjali G. Jivani. "Comparison of data mining classification algorithms for breast cancer prediction." Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on. IEEE, 2013.

[9] B. Padmapriya, T. Velmurugan, “A Survey on BreastCancer Analysis Using Data Mining Techniques”,Computational Intelligence and Computing Research(ICCIC), 2014 IEEE International Conference on.IEEE, 2014.