Breast Cancer Prediction Using Support Vector Machine

Ankit Verma[1], Ankit Kumar[2], Mr. Sanjeev Kumar[3]

[1]B.TECH. Scholar (CSE), ABES Institute of Technology, Ghaziabad
[2]B.TECH. Scholar (CSE), ABES Institute of Technology, Ghaziabad
[3]Professor, Dept. of Computer Science & Engineering, ABES Institute of Technology, U.P., INDIA

---***---

Abstract - *Cancer is the second leading cause of death in the world.9.6 million patients died due to cancer in 2018. Breast cancer is one of the most delicate and endogenous disease in any medical disease. This is one of the important reasons for the death of women around the world. In the world 1 out of 11 women die from breast cancer. A well known statement in cancer society is "Early detection means better chances of survival". So early detection is necessary as to prevent breast cancer with success and reduce morality. Breast Cancer Diagnosis and prediction have been one of the most important challenges faced by mankind in the last few decades. Accurately detecting cancer can save millions of lives. Effective tools for diagnosing cancerous breasts help healthcare professionals timely and accurately diagnose and treat patients. In this work, experiments were carried out using Wisconsin Diagnosis Breast Cancer (WDBC) database to classify the breast cancer as either benign or malignant. Supervised learning algorithm -Support Vector Machine (SVM). The classification performance of SVM classifier is evaluated. Experimental results show that SVM model has achieved a remarkable performance with 96.09% classification accuracy on testing subset.*

Key Words: **Breast Cancer, Prediction Support Vector Machine, Wisconsin Diagnosis Breast Cancer (WDBC), Benign, Malignant.**

1. INTRODUCTION

Breast cancer is the most common type of cancer found in women in the world. After years of technical and scientific researches, breast cancer is still the most common and the second leading reason for life taker of women. X- ray mammography was the orthodox method for diagnosis of breast cancer. Because of many variations in interpreting mammography, sharp needle aspiration cytology (FNAC) is used. The average accurate identification rate of FNAC is only 90%. So, it is very necessary to generate alternative identification methods to identify breast cancer. Data mining methods is a good choice to decrease the number of false diagnosis.

Early recognition of cancer can increase of survival of life upto 98%, Figure 1 indicates different types of cancers where by breast cancer is leading with 24%.

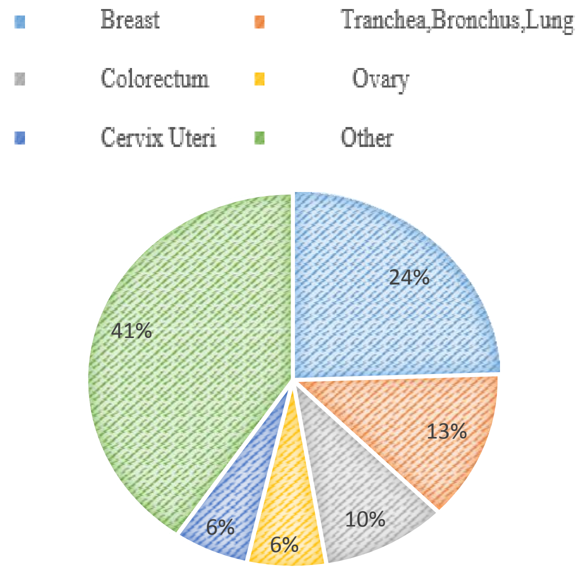

Figure 1: Types of Cancer

Artificial Intelligence (AI) can be used to for better and accurate detection and diagnosis of breast cancer, as well as to prevent overtreatment. Nevertheless, combining Artificial Intelligence (AI) and Machine Learning (ML) methods enables the prediction and increase the accuracy rate and decision making. For example, deciding on the biopsy output for detecting breast cancer if the patient needs surgery or not.

The survey of World Health Organization (WHO) reports that the breast cancer is the most common cancer amongst women. Around 5% of Indian women are have risk of breast cancer while Europe and in the U S, it is around 12.5%. Usually, breast cancer can be easily identified if specific symptoms arises. However, some women who are suffering from breast cancer have no symptoms. Hence, breast cancer recognition at early stage is very important.

Early detection of breast cancer helps in early diagnosis and treatment, because the prediction is very important for long-term survival. Early detection, diagnosis, and treatment of breast cancer can save a life of a patient.

1.1 RELATED WORKS

Many researchers have been conducted on the implantation of Machine Learning (ML) on Breast Cancer recognition, detection and diagnosis using different methods or combination of several algorithms for getting higher rate of accuracy.

[1] S. Gc, R. Kasaudhan [2] worked on withdrawing features including variance, range, and compactness. They used SVM classification to evaluate the performance. Their works output showed the highest variance of 95%, range 94%, compactness 86%. According to their output SVM can be considered as a good method for Breast Cancer Detection.

[2] Durai et al.[3] used Data Mining technique for recognising diseases including breast cancer. He used LRC and compared it with other techniques including BFI, ID3, J48 and SVM. The output shows that LRC is the most accurate one with 99.25% accuracy.

[3] Hafizah et al.[4] compared SVM and ANN using different data sets of breast cancer including WBCD, BUPA JNC, Data, Ovarian. The studies have demonstrated that both methods are having high performance but still, SVM was better than ANN.

[4]Tsirogiannis [5] applied bagging techniques using decision trees, neural networks and SVM on medical databases. As per the analysis, bagging techniques show better accuracy.

[5]Avramov and Si [6] worked on feature extraction and the impact of the selecting on performance. They applied different ways of correlation selection (PCA, T-Test Significance and Random feature selection) and five models of classification (LR, DT, KNN, LSVM, and CSVM). Best result was achieved by stacking the logistic, SVM and CSVM improve accuracy to 98.56%.

[6]Azar and El-Said [7] worked on six different methods of SVM. he compared ST-SVM with LPSVM, LSVM, SSVM, PSVM, and NSVM to find out which method performs the better in accuracy, sensitivity, specificity, and ROC. LPSVM proved to be the best with accuracy mark of 97.1429%, sensitivity 98.2456%, specificity 95.082%, and ROC 99.38%. Therefore, LPSVM has the highest performance and accuracy.

[7]Angeline [8] compared the performance of Naïve Bayes, Decision tree (C4.5), K-Nearest Neighbor and Support Vector Machine to find the preeminent classifier in Wisconsin Breast Cancer (WBC) to predict the primary site of cancer. As per the analysis, SVM performs better than other.

[8] by Mehmet Fatih Akay[9] had proposed medical decision making system based on SVM combined with feature selection has been applied on the task of diagnosing breast cancer. Considering the results, the SVM-based models have developed very promising results in classifying the breast cancer.

[9] by Leena Vig[10] had presented an analysis using Random Forest classifiers, Artificial Neural Networks, Naïve Bayes and Support Vector Machines. Results show that ANN's, Random Forests and SVMs are able to yield models with high accuracy, sensitivity and specificity whereas Naïve Bayes performs poorly.

2. OUR CONTRIBUTION

I ran a series of experiments in the python programming language using the Anaconda 5.3.0 SPYDER IDE and the Jupyter Notebook. To diagnose breast cancer. To diagnose Breast

cancer we follow some steps:

2.1 Data Collection & Preparation:

The breast cancer dataset named as Wisconsin Breast Cancer (WBC) data set is retrieved from UCI machine learning repository dataset [11]. This dataset comprises of 569 instances, where the cases are labelled as either benign or malignant and 357 (62.74%) cases are benign and 212 (37.25%) are malignant. The dataset is partitioned into two classes, B and M, where B denotes the benign class and M denotes the malignant class. Breast cancer is the most prominent disease in the field of medical diagnostics and is increasing each year. The dataset has 32 features that are Radius mean, Texture mean, Area mean, Smoothness, Compactness, Concavity except sample code number and class. The benign instances are represented as positive class as they don't affect the body badly and the malignant instances are represented as negative class as they are the cancerous cells that affect the body badly in our study. There are 16 missing values of features in the data set. The missing features are substituted by the mean for that feature. Finally, the data set is randomized to ensure correct propagation of data.

2.2 The Performance Measure Indices:

The performance of machine learning techniques is measured by several performance indicators. A confusion matrix for actual and predicted class is formed comprising of TP (True Positive), FP (False Positive), TN (True Negative), and FN (False Negative) to evaluate the parameter. The significance of the terms is given below.

TP (True Positive) = Correctly Identified

TN (True Negative) = Incorrectly Identified

FP (False Positive) = Correctly Rejected

FN (False Negative) = Incorrectly Rejected

2.3 Machine Learning Algorithm:

In our study, supervised learning algorithms are used. In fact, we use supervised machine learning algorithms, support vector machines.

Support Vector Machine:

Support Vector Machine is a supervised learning model for classification, and its classification performance is very good. In the SVM algorithm model, each data item is drawn as coordinates in n-dimensional space. Where n is the total number of features used for classification. The value of each feature is expressed in data point coordinates. The SVM contains decision hyperplanes to divide different classes of data points using maximum margin. Data points near hyperplanes are called support vectors. The classification process generates non-linear decision boundaries and classifies data points not represented in vector space.

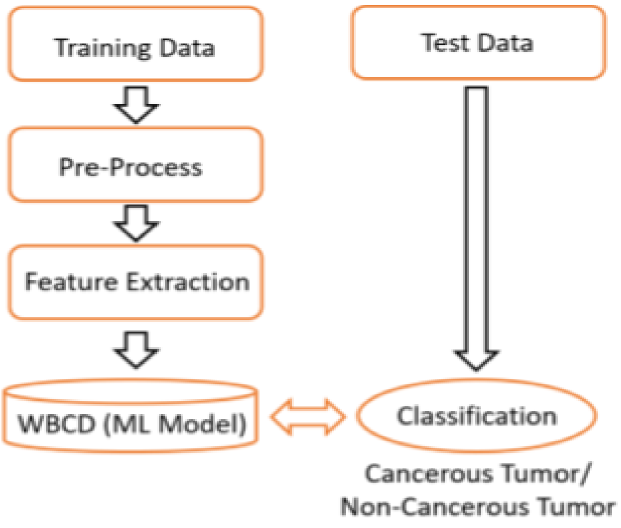

Figure 2: Flow diagram of work

Training Data: Training data is that in which train the model on the basis of classes of sample.

Pre-Process: Pre-process calculate mean and standard deviation using describe method in machine learning and remove the id and class.

Feature Extraction: Feature extraction is that in which irrelevant features are removed and train the model on relevant features

Dataset: WBCD it is Wisconsin Breast Cancer Dataset used to train our model. It updates every year so that it can be more useful in training model.

After classification the model that use SVM (Support Vector Machine) algorithm model of Machine Learning is able to classify faster that a cancerous raw data is actually a benign or malignant.

3. CONCLUSIONS

In this paper, we focused on a very dangerous disease that causes death for many women over the world which is the breast cancer. Breast cancer prediction is very significant in the area of Medicare and Biomedical. In this paper we focused on building a classifier which aims at predicting the most severe cancer known as breast cancer. In this we proposed a contributed method to diagnosis this disease and give information about the patient status. This article describes the breast cancer model as a classification task and describes the implementation of the Support Vector Machine (SVM) method to classify breast cancer as benign or malignant. The results of SVM consist of accuracy and precision. To summarize the developed method, the initial step, based on data gathering of patients in the form of text/csv file. Now extract the non-relevant feature like id and other.

Finally, the SVM classifier is used for classification, which train models to categorize cancer patients according to their diagnosis. Experimental results show that the effectiveness of model. SVM achieve 96.09% classification accuracy on test subsets.

REFERENCES

[1] E. D. Ubeyli,"Implementing automated diagnostic systems for breast cancer detection", Elsevier, Expert systems with applications, vol.33.

[2] S. Gc, R. Kasaudhan, T. K. Heo, and H.D. Choi, "Variability Measurement for Breast Cancer Classification of Mammographic Masses," in *Proceedings of the 2015 Conference on research in adaptive and convergent systems (RACS)*, Prague, Czech Republic, 2015, pp. 177–182.+

[3] S. G. Durai, S. H. Ganesh, and A. J. Christy, "Novel Linear Regressive Classifier for the Diagnosis of Breast Cancer," In Computing and Communication Technologies (WCCCT), 2017 World Congress on 2018.

[4]S. Hafizah, S. Ahmad, R. Sallehuddin, and N. Azizah, "Cancer Detection Using Artificial Neural Network and Support Vector Machine: A Comparative Study," J. Teknol, vol. 65, pp. 73–81, 2018.

[5]Tsirogiannis, G. L., et al. "Classification of medical data with a robust multi-level combination scheme." Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on. Vol. 3. IEEE, (2018).

[6] T. K. Avramov and D. Si, "Comparison of Feature Reduction Methods and Machine Learning Models for Breast

Cancer Diagnosis,” Proc. Int. Conf. Comput. Data Anal. - ICCDA ’17, pp. 69–74, 2018.

[7] A. T. Azar, and S. A. El-Said, “Performance analysis of support vector machines classifiers in breast cancer mammography recognition,” Neural Comput. Appl., vol. 24, no. 5, pp. 1163–1177, 2018.

[8].Christobel, Angeline, and Y. Sivaprakasam. "An empirical comparison of data mining classification methods." International Journal of Computer Information Systems 3.2 (2011): 24-28.

[9] Mehmet Fatih Akay, “Support Vector Machines Combined With Feature Selection For Breast Cancer Diagnosis”, Expert Systems with Applications 36, 3240–3247, 2018.

[10] LeenaVig , “Comparative Analysis of Different Classifiers for the Wisconsin Breast Cancer Dataset”, Open Access Library Journal, Volume 1 | e660,2018.

[11] Md. Milon Islam, Hasib Iqbal, Md. Rezwanul Haque, and Md. Kamrul Hasan,” Prediction of Breast Cancer Using Support Vector Machine and K-Nearest Neighbors”(September 2018)

[12].http://www.who.int/cancer/detection/breastcancer/en/ (Accessed November 2018)

[13]. http://www.stopcancerfund.org/pz-diet-habits-behaviors/lung-canceris-a-womens-health-issue/ (Accessed November, 2018)

[14].http://www.breastcancerindia.net/statistics/stat_global.html (Accessed November, 2018)

[15]. M. Lichman. 2015. UCI Machine Learning Repository. Retrieve from http://archive.ics.uci.edu/ml (Accessed November, 2018)

[16]http://www.openml.org/a/estimation-procedures/1(Accessed December, 2018)

[17].https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(D iagnostic) (Accessed 15 Dec 2018)

[18]http://scikitlearn.org/stable/modules/tree.html (accessed December, 2018)

[19] Soman K P, Loganathan R and Ajay V, “Machine Learning with SVM and Other Kernel Methods”, PHI, India.

[20]CD Katsis, I Gkogkou, CA Papadopoulos, Y Goletsis, PV Boufounou, and G Stylios. Using artificial immune recognition systems in order to detect early breast cancer. International Journal of Intelligent Systems & Applications, 5(2).

[21] Asuncion and Newman,"UCI machine learning repository" (accessed December 2018).

[22] Maglogiannis, Zafiropoulos, and Anagnostopoulos,"An Intelligent System for Automated Breast Cancer Diagnosis and Prognosis Using SVM Based Classifiers", Applied intelligence, vol. 30, no.1, pp. 24-36.

[23] Wisconsin Diagnostic Breast Cancer (WDBC) Dataset and Wisconsin Prognostic Breast Cancer (WPBC) Dataset. http://ftp.ics.uci.edu/pub/machine-learning-databases/breast-cancer-wisconsin