# BREAST CANCER USING MACHINE LEARNING

**DURGA DEVI[1], Venkat Surya[2]**

[1]Assistant Professor MCA dept, (Ph.D.) , **Dantuluri Narayana Raju College** , Bhimavaram, Andharapradesh
[2]PG Student of MCA, **Dantuluri Narayana Raju College**, Bhimavaram, Andharapradesh

• **Abstract Cancer is the common problem for all people in the world with all types. Particularly, Breast Cancer is the most frequent disease as a cancer type for women. Therefore, any development for diagnosis and prediction of cancer disease is capital important for a healthy life. Machine learning techniques can make a huge contribute on the process of early diagnosis and prediction of cancer. In this paper, two of the most popular machine learning techniques have been used for classification of Wisconsin Breast Cancer (Original) dataset and the classification performance of these techniques have been compared with each other using the values of accuracy, precision, recall and ROC Area. The best performance has been obtained by Support Vector Machine technique with the highest accuracy.**

**Index Terms—** SVM, k-NN, RF, NB, ,DT, AI-SIEM, FCNN, CNN,LSTM

## I Introduction

- Cancer is the second reason of human death all over the world and accounts for roughly 9.6 million deaths in 2018. Globally, for 1 human death in 6 can be said that is caused by cancer. Almost 70 percent of the deaths from cancer disease happen in countries that have low and middle income [1]. The most common cancer type among women are breast, lung and colorectal, which totally symbolize half of the all cancer cases. Also, breast cancer is responsible for the thirty percent of all new cancer diagnoses in women [2]. Machine learning (ML) methods ensure analyzing the data and extracting key characteristics of relationships and information from dataset. Also, it creates a computational model for best description of the data. Especially, according to in researches about cancer disease, it can be said that ML techniques can be handled on early detection and prognosis of cancer [3]. Asri et al. have compared some machine learning algorithms for the risk prediction and diagnosis of breast cancer. Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Naive Bayes (NB) and Decision Tree (C4.5) have been applied Wisconsin Breast Cancer (Original) dataset. SVM classification method has been given the highest accuracy value (97.13 %) with least error rate when the experimental results were compared.

Breast Cancer was used as a dataset and Weka software was used as a Machine Learning tool. The key performance parameters of machine learning classifiers have been compared according to accuracy, recall, precision and ROC area. They have suggested that BN has the best performance according to recall and precision values and RF technique has optimum performance in term of ROC area [5]. Ahmad et al. have exercised machine learning algorithms for predicting the rate of two years recurrence of breast cancer disease. The dataset has been obtained from Iranian Center of Breast Cancer (ICBC) program, collected the time period of 1997-2008 years. The dataset is consisted of population characteristics and 22 input variables also the cases have been collected from 1189 women of diagnosed breast cancer. Artificial Neural Network (ANN), Support Vector Machine (SVM) and Decision Tree (DT) have been applied

and SVM has been showed the best performance with highest accuracy and least error rate

## 2 Literature survey

In the literature, many works applied Artificial Intelligence (AI) technics for breast cancer diagnosis to improve classification accuracy and time response. In this section, we give some works related to the solution of medical breast cancer diagnosis using machine and deep learning approaches. Arpit B. and Aruna T. [5] proposed a genetically optimized neural network (GONN) for breast cancer classification (malignant and benign). They optimized the neural network architecture by introducing new crossover and mutation oper ators. To evaluate their work, they used WBCD and compared the classification accuracy, sensitivity, specificity, confusion matrix, ROC curves and AUC un der ROC curves of GONN with classical model and classical Back propagation model. This method presents a good accuracy classification. However, it can be improved by using a larger dataset than WBCD, feature extraction to make GONN more efficient for real time diagnosis of Breast Cancer. Ashraf O. I. and Siti M. S. [6] proposed a computer-based method to classify automatically breast cancer disease. The method applied multilayer perceptron (MLP) neural network based on enhanced non-dominated sorting genetic algo rithm (NSGA-II) to optimize both the accuracy and network structure. Com pared to other methods, this work improves classification accuracy. However, MLP can get stuck in local minima. Na L. et al. [7] proposed an intelligent classification model for breast cancer diagnosis based on a hybrid feature selection approach: gain directed simulated annealing genetic algorithm wrapper (IGSAGAW), to remove redundant and irrelevant feature from the feature space and cost sensitive support vector ma chine (CSSVM) learning algorithm. This process can improve the classification accuracy and reduce the computational cost. The proposed method is applied on Wisconsin Original Breast Cancer (WBC) and WBCD to verify its effectiveness. The proposed work shows a good performance and decreases the calculation complexity. Nawel Z. et al. [8] presented a conception and implementation of Computer As sisted Detection (CAD) for mammogram images classification. The system is based on a GA-based features selection algorithm to reduce the dimensional ity of the feature vector and semi supervised support vector machine (S3VM) for classification. Experiments were validated on Digital Database for Screening Mammography (DDSM) dataset. The proposed approach improved accuracy. Abdulkader H. et al. [9] developed an automated system for classification of breast tissue. The system uses two machine learning techniques: Feed forward neural network using the back propagation learning algorithm (BPNN) and ra dial basis function network (RBFN). Breast cancer tissues were classified into 6 different tissues, Carcinoma, Fibro-adenoma, Mastopathy, Glandular, Con nective, and Adipose tissue. Data were acquired using an electrical impedance spectroscopy (EIS) method. The Radial basis function network outperformed the back propagation network for classifying six different breast tissues in terms of accuracy, minimum error, maximum epochs and training time. The proposed system improved accuracy and decreases training time. However, learning with ML Techniques for Breast Cancer Diagnosis: Literature Review 5 neural networks can be weak in generalizing and can get stuck in local optima. Haifeng W. et al. [10] designed an SVM-based ensemble learning model for breast cancer diagnosis. The proposed ensemble model includes two types of SVM struc tures, i.e., a C-SVM and a -SVM, and six types of kernel functions. To import the expertise of different base classifiers on diagnostic tasks, a Weighted Area Under the Receiver Operating Characteristic Curve Ensemble (WAUCE) mechanism is proposed for model hybridization. The model was evaluated based on two datasets: the Wisconsin Breast Cancer (WBC) dataset and the Wisconsin Diag nostic Breast Cancer (WDBC) dataset, and one large dataset, the Surveillance, Epidemiology, and End Results (SEER) dataset. The proposed model increases

diagnosis accuracy compared to other works based on single SVM. However, it is a computationally expensive method and the training time is high. Kemal P. et al. [11] proposed a hybrid approach based on mad normalization, KMC based feature weighting and AdaBoostM1 classifier. The detection of the presence of breast cancer is done in three steps: In the first step, the dataset was first normalized by the MAD normalization method. In the second step, k-means clustering (KMC) based feature weighting has been used for weighting the nor malized data. Finally, the AdaBoostM1 classifier has been used to classify the weighted data set. The Breast Cancer Coimbra dataset (BCC) taken from UCI machine learning database was used. This method shows good results in terms of accuracy. However, it is a computationally expensive method. Teresa A. j. et al. [12] proposed a classification method of hematoxylin and eosin stained breast biopsy images using Convolutional Neural Networks (CNNs). They provide four classes of medical relevance: normal tissue, benign lesion, in situ carcinoma and invasive carcinoma. The proposed CNN architecture is designed to integrate information from multiple histological scales. The model is applied on image dataset composed of high resolution uncompressed, and anno tated HE stain images from the Bioimaging 2015 breast histology classification challenge

**3 Implementation Study**

We have applied SVM and ANN techniques for prediction of the classification of breast cancer to find which machine learning methods performance is better. Support Vector Machines (SVMs) have been first explained by Vladimir Vapnik and the good performances of SVMs have been noticed in many pattern recognition problems. SVMs can indicate better classification performance when it is compared with many other classification techniques . SVM is one of the most popular machine learning classification technique that is used for the prognosis and diagnosis of cancer. According to SVM, the classes are separated with hyperplane that is consisted of support vectors that are critical samples from all classes. The hyperplane is a separator that is identified as decision boundary among the two sample clusters. SVM can be used for classifying tumors as benign or malignant based on patient's age and tumors size . Artificial Neural Network (ANN) can be expressed in terms of biological neuron system. Especially, it is similar to human brain process system. It is consisted of a lot of nodes that connect each node [12]. ANN have the ability of modelling typical and powerful non-linear functions. It is consisted of a network of lots of artificial neurons.

**3.1proposed methodology**

Machine learning involves predicting and classifying data and to do so we employ various machine learning algorithms according to the dataset. SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes. In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification. As a simple example, for a classification task with only two features (like the image above), you can think of a hyperplane as a line that linearly separates and classifies a set of data. Intuitively, the further from the hyperplane our data points lie, the more confident we are that they have been correctly classified. We therefore want our data points to be as far away from the hyperplane as possible, while still being on the correct side of

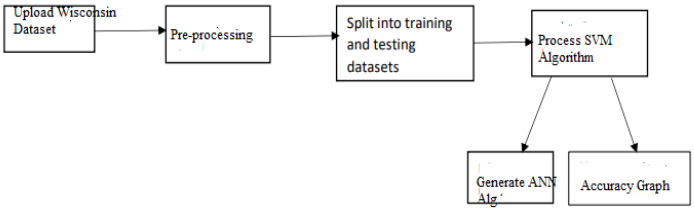


Fig 1: - flow of proposed system

In this the input log data taken is pre-processed to find any inconsistent and null data. Then it is split into training and testing to fit it into the classifier. The SVM Algorihm is trained against training data and is tested using the test data and the accuracy is predicted. Therefore it classifies the user as genuine.

**3.2 Methodology**

The model implementation consists of modules:

1. Upload Dataset
2. Pre-processing
3. Process On Training and Testing Model
4. SVM Model

**3.2.1 Upload Dataset**

Upload Dataset is the process of importing raw data sets into your analytical platform. It can be acquired from traditional databases (SQL and query browsers), remote data (web services), text files (scripting languages), NoSQL storage (web services, programming interfaces), etc. Upload Dataset involves the identification of data sets, retrieval of data, query of data from the dataset. The dataset used in the project is collected from Wiconsin Dataset . We used additional tools to get other information, such as, server country with Whois. The final dataset consists of around 1780 values which can serve as a training set for Machine Learning models

**3.2.2 Pre-Process Data :**

Pre-processing of data involves 2 criteria:

**Cleaning Data:** Data cleaning involves removal of inconsistent values, duplicate records, missing values, invalid data and outliers.

**3.2.3 Data Munging / Data Wrangling:** Data Wrangling techniques involve scaling, transformation, feature selection, dimensionality reduction and data manipulation. Scaling is performed over the dataset to avoid having certain features with large values from dominating the results. The transformation technique reduced the noise and variability present in the dataset. Multiple features are handpicked for the removal of redundant/irrelevant features present in the dataset. Dimensionality reduction helped in eliminating irrelevant features and made analysis simpler.

**3.2. 4 Support Vector Machine :-**

“Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well (look at the below snapshot).

**4 Results and Evolution Metrics**

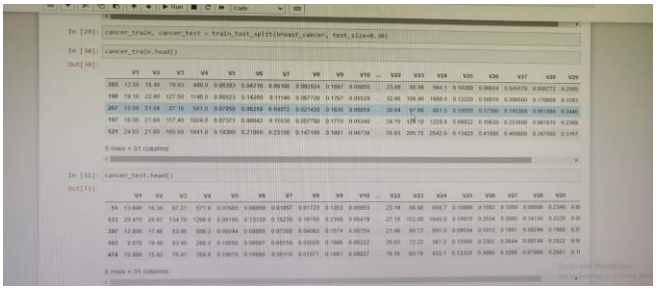

**Fig 2:_** Main screen

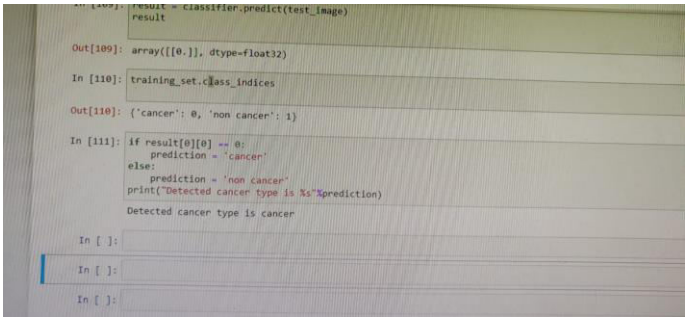

**Fig 3:** predicted outcome

## 5 Conclusion

Breast Cancer is the most frequent disease as a cancer type for women. Therefore, any development for diagnosis and prediction of cancer disease is capital important for a healthy life. In this paper, we have discussed two popular machine learning techniques for Wisconsin Breast Cancer classification.

## 6 References

[1]Cancer,https://www.who.int/en/news-room/fact sheets/detail/cancer. Last Access: 25.01.2019.

[2] Siegel, R. L., Miller, K. D., & Jemal, A. (2018). Cancer statistics, Ca-a Cancer Journal for Clinicians, 68 (1), pp. 7-30.

[3] Maity, N. G., & Das, S. (2017). Machine learning for improved diagnosis and prognosis in healthcare. In 2017 IEEE Aerospace Conference, pp. 1-9.

[4] Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83, pp. 1064-1069.

[5] Bazazeh, D., & Shubair, R. (2016). Comparative study of machine learning algorithms for breast cancer detection and diagnosis. In 2016 5th International Conference on Electronic Devices, Systems and Applications, pp. 1-4.