ISSN No:-2456-2165

Classification of Breast Cancer Detection by Using Machine Learning Technique

Tushar Khandelwal
Sumati Gupta

Abstract:- Breast cancer causes more death in women and it also curable if it is early diagnosed. Hence, early detection of cancer in women will be helpful in taking necessary actions. In order to detect the disease supervised machine learning techniques is discussed in this paper. With the help of Sequential ForwardSelection (SFS) best feature will be selected for support vector machines (SVM) model. Wisconsin breast cancerdataset (WBCD) is used for diagnosis of breast cancer. The SVM result shows 96% precision because of random permutation on the data set.

Keywords:- Sequential Forward Selection SFS; Support Vector Machine; Breast Cancer; Classification; Machine Learning; Wisconsin Breast Cancer Dataset.

I. INTRODUCTION

Cancer or cancer cell are the cells that have lost the ability to follow the normal control that the bodyexerts on all cells. Cancer can occur anywhere in our body because there are cells everywhere in our body. In women one of the most common cancer is breast cancer and in men prostate cancer and in both menand women lung cancer and colorectal cancers. Generally, cancer has number of types which are Carcinomas, Sarcomas, Leukemias and Lymphomas. Carcinomas it a type of cancer which starts from skin or tissues that covers outer layers of internal organs and breast cancer, prostate cancer, lung cancer are example of Carcinomas. Breast cancer begins when there is irregular development or unusual change in healthy cells forming a sheet of cells known as tumour. Tumours can either be non-cancerous (benign) or cancerous (malignant). Healthy body tissues are destroyed by Cancerous tumour when they break in.

Women 40 to 50 years of age die with breast cancer and this rate of death is ranked second in the causes ofdeath in the women. There are almost 145000 cases inIndia according to world health organization. Huge innovation in medical science has caused decreased inthe cases of breast cancer as there are effective treatments methods now. Early detection and diagnosing accurately is key factor for decrease in breast cancer.

Advances in medicine in past few decades have improved health care immensely. Allowing doctors tomore efficiently diagnose and treat diseases. The biggest difference between doctors is not their levelof intelligence it's how they approach patient problems and the types of health system that supports them. This combination is what causes such wide variations in clinical outcomes and it's the reason why machine learning is the best solution out there to improve doctor's capabilities. There is so much potential here studies show that over half of all women in the U.S.A who get regular mammograms will receive at least one false positive which is a test that wrongly indicates the possibility of cancer. Radiologists regularly disagree on their respected interpretations of medical images. Artificial Intelligence can do what no radiologist can it canlearn from hundred and thousands of medical images and its estimated to be up to 10 percent more accurate than average radiologist that accuracy gap willincrease as computing power gets cheaper and can be applied to any of the countless subfields of medicine not just radiology. Doctors also have to interpret patient medical which can be very complex task NLP a branch of artificial intelligence that helps computers understand and interpret human language can review thousands of medical records and output the optimal steps for evaluating and managing patients with illness. Doctors have natural biases artificial intelligence is more likely produce objective diagnosis for patents without preconceived socio- economic notion which can produce disparities in care machine learning will become an essential tool for doctors. It helps in minimizing and optimizing the error in short time and it can be examined in more detailed way. In this study, SFS and SVM feature are used to diagnose the breast cancer. WBCD form university of California at Irvine (UCI) machine learning repository was use for training and testing experiment. The observation was that when we shuffle the data by using random permutation on it and then applied SFS. By applying SFS on dataset it gave 96.4% accuracy by using best ten feature of the dataset those are texture mean, perimeter mean, smoothness mean, texture, area, fractal dimension, texture worst, smoothness worst, concave points worst, diagnosis. With the help these features a new dataset is created on which SVM is applied.

II. LITERATURE REVIEW

In (2002) Vinterbo, Ohno-Machado, Wong, Lappas and Albrecht 98.8% accuracy was record when logarithmic simulated annealing learning and the perceptron algorithm are combined together [1]. In (1999) Sipper and Pena-Reyes, reached 97.36%accuracy in fuzzy-GA method [2]. In (2000) Setiono 98.10% accuracy was reported in feed forward neural network rule extraction algorithm [3]. By using 10- fold cross-validation with C4.5 decision tree method 94.74% accuracy was reported by (Quinlan) in 1996. RIAC method was used by Cercone, Shan, &Hamiton, in 1996 and they obtained 94.99% accuracy[6]. In 1996 by Dobnikar & Ster used linear discreet analysis method to obtained 96.8% accuracy [5]. Neuron-fuzzy techniques are used by Kruse and Nauck in (1999) to obtained accuracy 95.06% [6]. In Goodmen, Bogess, and Watkeens (2002), three different

methods were used first is optimized learning vector quantization (LVQ) and 96.7% accuracy was record, by using artificial immune recognition system (AIRS) 97.2% accuracy wasreported and big LVQ were applied and the obtained accuracy was 96.8% [7]. In (2003) 95.57% accuracy was obtained by Szeifert and Abonyi by using the application of supervised fuzzy clustering technique [10]. In (2007) 98.53% accuracy was obtained by Gunes and Polat by using least square SVM. Mehmet Fatih Akay states SVM with feature selection results categorize patients whether they are suffering from cancer or not and the f-score f-score accuracy score achieved was 99.5% by using 5 feature. Feature like range, compactness and variance were extracted by S.GC et al and then SVM classification was used to evaluate the performance. This is how they figuredout that SVM is the best method as it showed 95 % variance, 86% compactness, 94% range [11]. chunqiu wang et al used ANN to classify image and with help of MTI Microwave Tomography Imaging extracted features. KNN and GMM techniques were used and compared. KNN recorded 87% accuracy where as GMM recorded 67. In term of accuracyKNN did better but in terms of specificity GMM is a better option [12]. For cheap, effective and efficient research Chowdhary and Acharjya used mammogram image. In order to improve performance extracting and selecting the features matters, image quality was increased by (FHH) Fuzzy Histogram Hyperbolization, to extracts feature Grey level dependence model was applied and for segmentation fuzzy C-mean. An accuracy of 94% was detected for malignant breast lesions in their research [13]. Aminikhanghahi et al. conducted a research to explore images with help of wireless cyber mammography. Then features are extracted andselected so that machine learning techniques can be performed it. The two machine learning techniques they used were SVM and GMM. Results showed that SVM was more accurate than GNN[14].

III. PROPOSED METHODOLOGY

Proposed method architecture is illustrated in fig 1. Injupyter implementing classification learnerapplication on proposed algorithm.

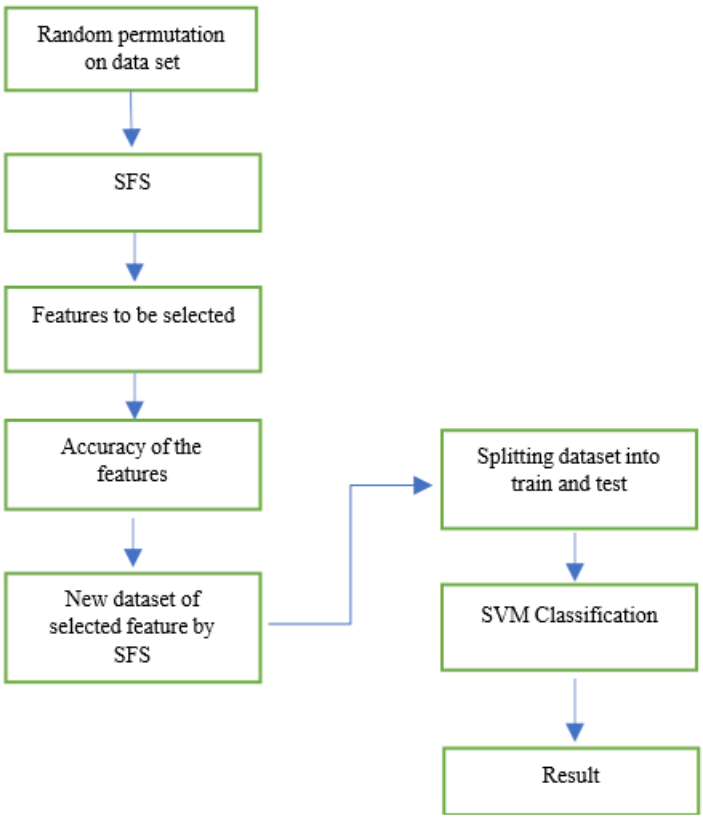

Fig. 1: Flow chart for cancer detection

Step 1: The dataset is taken as an input for random permutation. So that random features of a dataset are selected.
Step 2: The motivation behind feature selection algorithms is to automatically select a subset of features that is most relevant to the problem. The goalof feature selection is two-fold: We want to improve the computational efficiency and reduce the generalization error of the model by removing irrelevant features or noise

Step 3: The feature selection algorithm will give mostrelevant feature of the dataset on which accuracy will be tested.
Step 4: On the basis of accuracy a new dataset will be created. Those features who has the maximum accuracy will be selected for new dataset.
Step 5: Those feature who has the maximum accuracyare selected for new dataset and now on these features classification will be conducted

ISSN No:-2456-2165

Step 6: The extracted features data are used for training different models with Classification Learner Application. With the help of Cross validation machine learning algorithm predicts new datasets. By splitting the dataset into testing set and train set this isachieved. Once the datasets are trained, based on the accuracy of different techniques, select the best modelfor testing. For testing export model will be used and new input features for the diagnoses of tumour.

IV. RESULTS

An experiment is conducted on the Wisconsin breast cancer dataset WBCD in order calculate the efficiency of our method. The feature selection algorithm will give most relevant feature of the dataset on which accuracy will be tested.

```
feat_cols = list(sfs1.k_feature_idx_)
print(feat_cols)
```

```
[0, 2, 3, 5, 12, 14, 20, 22, 25, 28]
```

```
from sklearn.metrics import accuracy_score

print(accuracy_score(y_test, prediction))
```

```
0.9649122807017544
```

Fig2: SFS feature to be selected and their accuracy

Since SFS results with highest accuracy during training, SVM model will be exported for testing or prediction of new datasets. Selected features will be taken as an input and SVM will be implemented onthe extracted features shown in fig.2.

The F-score measures the importance of each feature. Grid search optimizes the SVM parameters. The F score can be taken as a weighted average of the recall and precision, where an F1 score reaches its worst score at 0 and best value at 1. The relative contribution of precision and recall to the F1 score are equal. Table 1 to 3 shows classification of accuracies.

Confusion matrix shows true negative rate and true positive rate of each class taken. The precision of the classification models is based on features that has been selected.

	Precision	Recall	F1-score	Support
1	0.98	0.86	0.92	64
0	0.92	0.99	0.95	107
Micro avg	.94	0.94	0.94	171
Macro avg	0.95	0.93	0.94	171
Weighted avg	0.94	0.94	0.94	171

Table 1- The accuracy achieved when 70-30% data was divided into training and test respectively and accuracy is 94%.

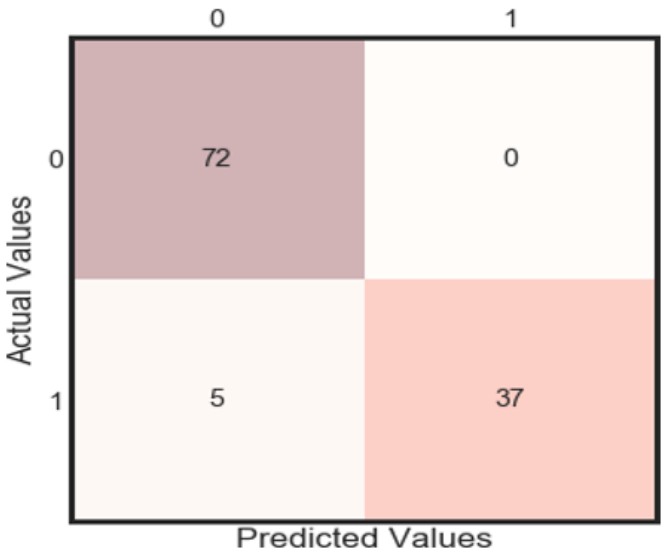

Fig.3: Confusion Matrix of 20% testing - 80% training.

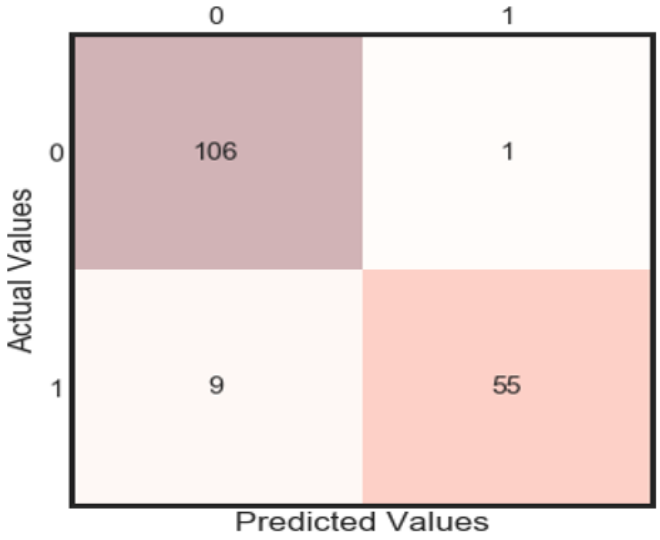

Fig.4: Confusion Matrix of 30 testing -70% training.

The possibilities are “1” meaning malignant and “0” meaning Benign. Here a test is done on 171 patients for the presence of breast cancer. According to dataset107 patients are not suffering from breast cancer and 64 patients are suffering from breast cancer. Prediction made by classifier are 56 times “yes” and 117 times “no”.

Accuracy: (true positive + true negative)/total = “(55+106/171) = 0.94”

	Precision	Recall	F1-score	Support
1	1.00	0.88	0.94	42
0	0.94	1.00	0.97	72
Micro avg	0.96	0.96	0.96	114
Macro avg	0.97	0.94	0.95	114
Weighted avg	0.96	0.96	0.96	114

Table 2- The accuracy achieved when 80-20% data was divided into training and test respectively and accuracy is 96%.

ISSN No:-2456-2165

Here a test is done on 114 patients for the presence of breast cancer. According to dataset 72 patients are not suffering from breast cancer and 42 patients are suffering from breast cancer. Prediction made by classifier are 37 times "yes" and 77 times "no".

Accuracy: (true positive + true negative)/total = "(37+72)/114=0.96"

	Precision	Recall	F1-score	Support
1	0.99	0.89	0.94	106
0	0.94	0.99	0.96	179
Micro avg	0.95	0.95	0.95	285
Macro avg	0.96	0.94	0.95	285
Weighted avg	0.96	0.95	0.95	285

Table 3- The accuracy achieved when 50-50% data was divided into training and test respectively and accuracy is 95%.

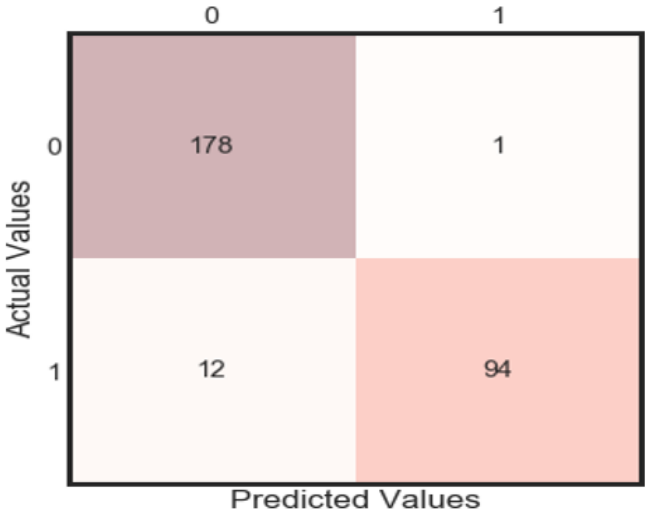

Fig.5: Confusion Matrix of 50% testing -50% training.

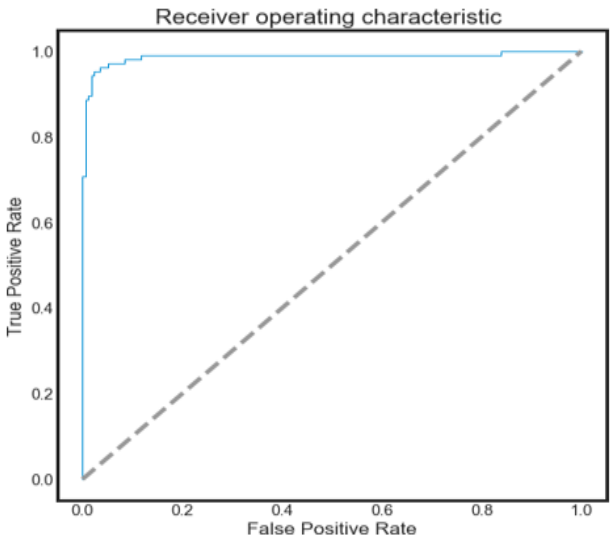

Fig.6: ROC curve for 20% testing -80% training.

Here a test is done on 285 patients for the presence of breast cancer. According to dataset 179 patients are not suffering from breast cancer and 106 patients are suffering from breast cancer. Prediction made by classifier are 95 times "yes" and 190 times "no".

Accuracy: (true positive + true negative)/total = "(94+178)/285 = .95"
Matrix y-axis defines the true class and matrix x-axis depicts predicted class.

The Receiver Operating Characteristic (ROC) of linear SVM. Subsequently, to know the accuracy a model should be able to differentiate between patients being benign and malignant. The performance visualization is done through ROC graph. Whereas summarizing a single value of overall performance is done through area under curve(AUC)

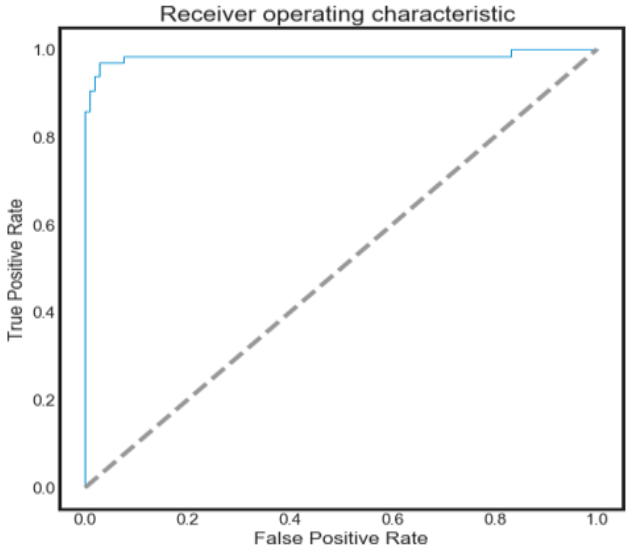

Fig.7: ROC curve for 30% testing -70% training.

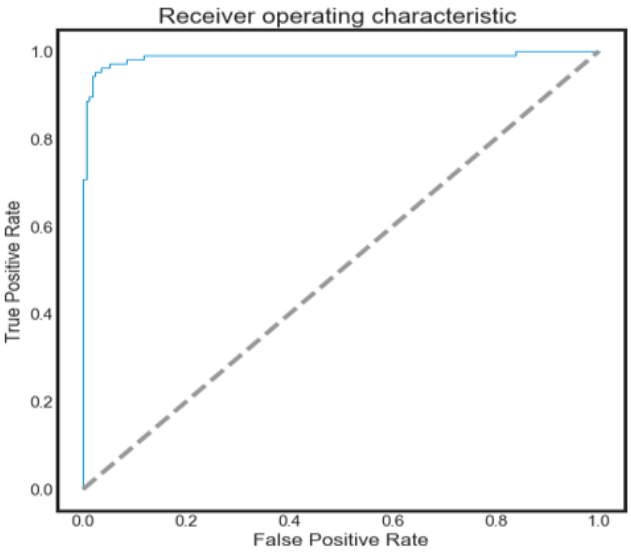

Fig.8: ROC curve for 50% testing -50% training.

ISSN No:-2456-2165

V. CONCLUSION

In this we have used random permutation so that random feature is select and based on those featurewe are able to classify patients is benign and malignant. In this way machine will be trained on different features and it will be able to classify random possibilities of benign and malignant. An experiment was conducted on WBCD and SFS technique which randomly selects features and gave 10 best feature from the data set and yield 96% accuracy. These 10 feature are selected for further implementation SVM based model. SVM is learning technique which helped in categorizing how manypatients are benign and malignant. Additional measures present in the model are ROC curve, predictive values (positive and negative), confusion matrices. SVM model will increase the performance and accurate prognosis. Selected 10 features were used in SVM model and 96% accuracy was observed in the classification method. We believe that continuous use machine learning in field of health and medical will improve the quality of studies.

REFERENCES

[1]. Allbrecht, aandreas A., et al. "Two applications of the LSA machine." Global Seminar on Neural Dispensation, 02. ICONP'02... V0l. I. IEEE, 2002.
[2]. P.Reyes, C.Andres, and M.Ssipper. "A fuzzy-genetic approach to breast cancer diagnosis." Artificial intelligence in medicines 17.2 (1999): 131-155.
[3]. Seetiono, Rody, B.Baesenes, and C.Muees. "Recursive neural network rule extraction for data with mixed attributes." IEEE Neural Networks 19.2 (2008): 299-307.
[4]. archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wiscon sin+(Origina)
[5]. B Bhardwaj, Arpeet, and Arunaa Tiwari. "Breast cancer diagnosis using genetically optimized neural network model." Expert Systems with Application 42.10 (2015): 4611-4620., 18(3), 205–217.
[6]. Hamillton, H.John, N. Cerconee, and N Sahan.: a instruction introduction procedure built on estimated sorting. Computer Branch, School of Regina, 1996.
[7]. Marcaano-Cedeeño, Aleexis, Joeel Quintanilila-Domíinguez, and D Andina. "WBCD breast cancer database classification applying artificial metaplasticity neural network." Professional Organizations using Application 38.8 (2011): 9573-9579.
[8]. Nauckk, Detleff, and Rodolf Krusee. "A fuzzy neural network learning fuzzy control rules and membership functions by fuzzy error backpropagation." global seminar on neural system. 1993.
[9]. Abdeel-Zaheer, Ahmeed M., and Aymaan M. Eldeeb. "Breast cancer classification using deep belief networks." Experts Systems with Application 46 (2016): 139-144
[10]. Abonyee, Janes, Robertt Babbuska, and Ference Szeifeert. "Modified Gath-Geva fuzzy clustering for identification of Takagi- Sugeno fuzzy models." IEEE Transactions on Systems, Man, and Cybernetics, 32.5 (2002): 612-621.
[11]. Polaat, Kemaal, and Saleeh Guneeş. "Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fast Fourier transform." Practical Maths and Computation 187.2 (2007): 1017-1026.
[12]. Chan, Feng, et al. "Sensitive determination of endogenous hexanal and heptanal in urine by hollow-fibre liquid-phase microextraction prior to capillary electrophoresis with aerometric detection." Atlanta 119 (2014): 83-89.
[13]. Chaudhary, C Lal, and D. P. Acharjya. "A hybrid scheme for breast cancer detection using intuitionistic fuzzy rough set technique." International Journal of Healthcare Info Systems and Informatics 11.2 (2016): 38-61.
[14]. Aminikhanghaee, Samaneh, et al. "Effective tumor feature extraction for smart phone based microwave tomography breast cancer screening." Proceedings of the 29th Yearly ACM Symposium on Applied Computing, 2014.