An Improved Model for Breast Cancer Classification Using Svm with Grid Search Method

Anita Paneri, Mayank Patel

Abstract: ***In today's era, unhealthy lifestyle is the main cause of increase in diseases among human beings. Breast cancer (BC) is one such disease responsible for a sudden increase in death rate among women. Breast cancer mass is mainly classified into two Benign and Malignant. Benign refer to non-cancerous which means it cannot spread through other parts of the body while malignant is cancerous that is it can spread through other parts of the body. If it is detected on early stage, it can be treated on time. In this paper, we will use machine learning algorithms for breast cancer classification into B (benign) and M (malignant). Here, we will use Wisconsin Breast cancer Data Set and will apply Support Vector Machine (SVM) using python and then will develop an improved model using SVM-GSM (Grid Search Method) model for breast cancer classification and will analyze their results accordingly.***

Index Terms**: *Benign, Malignant, Support Vector Machine, Grid Search Method.***

I. INTRODUCTION

Breast cancer is one of the most critical cancers which are responsible for death among women. BC mainly occurs in the mammary glands or milk ducts of the breast, where masses are formed due to many reasons such as lactation issues, menopause etc. Due to which lump formation takes place in women. Although, all the lesions formed are not cancerous. To identify cancerous and non-cancerous lumps in women, the physician prescribes various techniques which are as follows:

1. Self examination to check if any kind of mass can be felt.
2. Using X-Ray for screening.
3. Using ultrasound
4. Using Biopsy.

These are some of the pathological techniques advised by physician to identify BC, which is sometimes not accurate or fast. Therefore to overcome this problem, we will use Machine learning algorithms to achieve accuracy. Machine learning algorithms are an intelligent and accurate method to predict the Benign and Malignant from the given WBCD data of Fine Needle Aspirate images. Here, we will use SVM along with Grid Search Method to give an improve model to detect 'B' and 'M' from FNA images in digitized format.

II. LITERATURE SURVEY

In the recent years, various methods have been proposed and added to the literature regarding breast cancer diagnosis and prognosis. In year 2016, [1]They have proposed a method based on Wrapper Feature Selection method applied on Mammograms. They are then classified into 'B' or 'M' from Mammographic images. In 2017,[2]The paper presents different machine learning algorithms applications which includes the GRU-SVM model which is proposed for the diagnosis of breast cancer. All ML algorithms exhibit high performance measure on the binary classification of breast cancer, i.e. determining whether benign or malignant. In year 2018,[3]" They proposed a paper in which they investigated various ML approaches such as ANNs, K-NN, Decision Tree, SVM and applied on WBCD data set for breast cancer classification, which is the database used for comparing the results through different algorithms. Finally, a healthcare system model of our recent work is developed."In year 2019,[4] They present an original approach to integration of data from two cancer studies concerning all details of the records. A set of ML-based models have been applied for survival time prognosis in breast cancer and the corresponding results are analyzed in the paper The SVR-Linear, Lasso, Kernel Ridge, KNR, and DTR showed most accurate survival prognosis results. In the research paper, breast cancer classification is carried out using Wisconsin Diagnostic Breast Cancer (WDBC) database which consists of the details of Fine Needle Aspirate images data having several features. Support Vector Machine along with grid search method used to predict 'B' or 'M' from breast mass

III. MACHINE LEARNING

Machine learning is a trending field of research study that mainly concerned with the statistical design of algorithms that allow computers to learn. The term "Machine Learning" is derived from the artificial intelligence branch but nowadays, it is mainly the focusing area for many branches of science and engineering. Learning mainly refers to learning from data set or feature set. The main aim of machine learning is to create intelligent machines that can think and work like human beings. There are various machine learning methods are used for analysis of statistical data. They are supervised learning, unsupervised learning and reinforcement learning.

A. Support vector Machine:

Support vector machine is a type of supervised machine learning algorithm which is employed for solving classification and regression problem in real time. SVM is generally used when the number of cases and the number of attributes are very large in number. SVM was originally built for binary classification, but in case, it can be extended for multiple class problems. The SVM classifier is used to find the best of the hyper plane which divides the dataset into two classes. The principle of SVM initiates from solving linear separable problems and then extends to solve non-linear problems. In this research work, we will use it for Breast cancer classification. An example of the SVM classifier used for BC classification is shown in figure 1

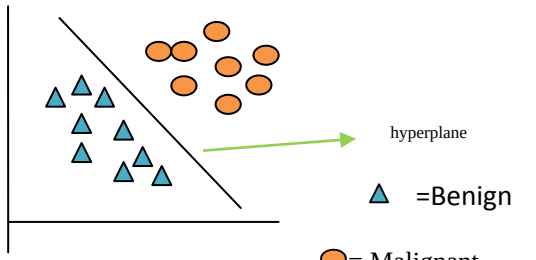

Figure1. How an SVM might w between Benign and Malignant

The parameters used in SVM are:

1. Radial bias function(RBF) kernel
2. c value

B. Grid Search Method:

It is an optimization method which is used to search best subset of hyper-parameters from the list of parameters. The method followed in this pattern is similar to the grid, where all the values are placed in the form of a matrix. The selected hyper-parameters are the best of performing parameters selected from training set. In this method we simply build a model for every combination of various hyper parameter and we evaluate each model. The final hyper-parameters are the best subset of parameters used in final model of our testing set. We have applied grid search using cross validation on the training dataset of the breast cancer. In grid search we have used sklearn model selection method.

The Best hyper parameters selected using grid search methods mentioned below:

1.C = 100,
2.Gamma = "0.01"
3.kernel = 'rbf'

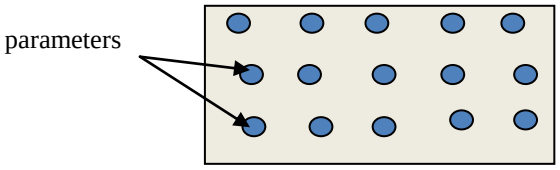

Figure2. Grid Layout

IV. PROPOSED METHODOLOGY

A. *Data Collection*: The data used in this research work is taken from the WBCD from the UCI repository. It has data collection of FNA images containing 569 instances, where 357=Benign and 212= Malignant.

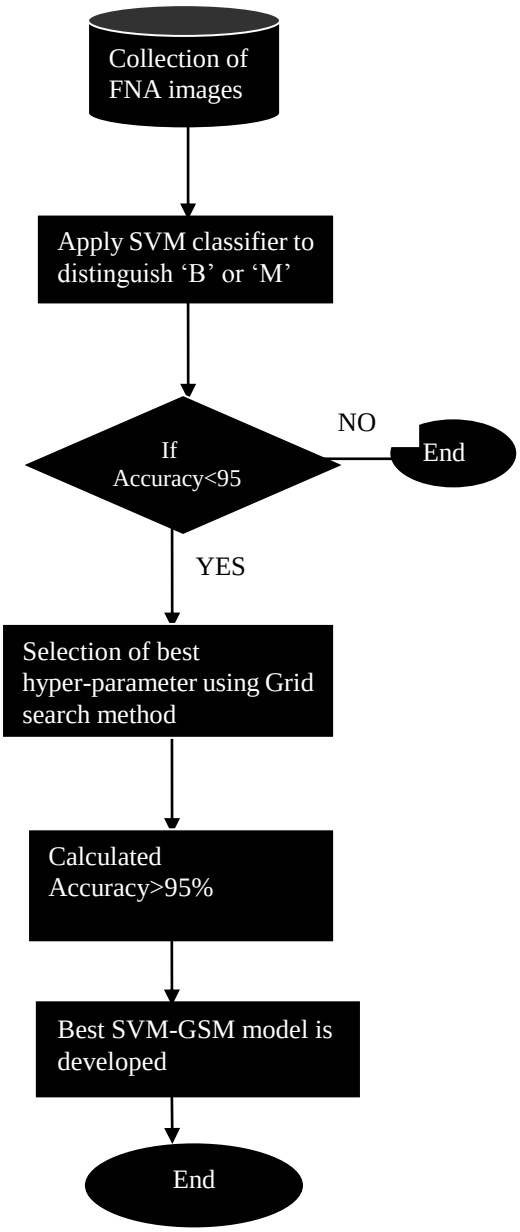

Figure3. Basic work flow diagram

B. Performance Measure Indices: In field of machine learning there are various classification algorithms employed to predict results. Classification is an integral part of supervised learning, which mainly aims on data classification depending on different features. The Training data set is used to train the model which in turn predicts the unknown labels of population data. To find the best classification algorithm for distinguish between benign and malignant. In this work, we will use certain evaluation methods to measure performance of algorithms [5]. We have split the training data and test data into the ratio of 4:1 The performance classifier depends on various factors like Accuracy, Precision, Recall, F1-score and confusion matrix. The confusion matrix is described as below:

Table1. Confusion Matrix

Confusion matrix	Actual Results	
Predicted Results	TP	FP
	FN	TN

1.Accuracy=TP+TN
TP+TN+FP+FN
2. Precision= TP ⁄ TP+FP
3. Recall = TP ⁄ TP+ FN
4. F1Score = 2(precision*recall) ⁄ (precision +Recall)
Where, TP= Number of True Positive predictions
FP= Number of False positive predictions
TN=Number of True Negative predictions
FN= Number of False negative predictions

V. EXPERIMENTAL RESULTS

We have proposed a model using SVM classifier along with grid search method to achieve highly accurate results.

Table2.The confusion matrix for SVM model is as follows:

	Predicted as Cancer	Predicted as healthy
Cancer	71	0
Healthy	43	0

Here, the total number of false predictions = 43**.** So we have to improve our model to achieve higher accuracy

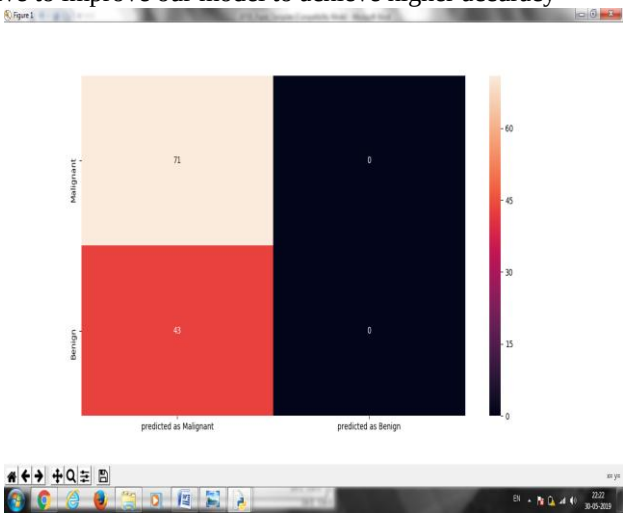

Figure2. Heat map of predicted results using SVM

The above heat map shows 43 false predictions which mean 43 people are not having cancer but are declared as cancer.

Table3: Table below shows the precision, recall, F1-score and support using SVM

	Precision	Recall	F1-score	Support
M= 0.0	0.00	0.00	0.00	43
B=1.0	0.62	1.00	0.77	71
average	0.39	0.62	0.48	114

To improve the model we will first normalize the data then we will use grid search method to choose best of the hyper parameters to build an appropriate model to classify Breast cancer into benign and malignant.

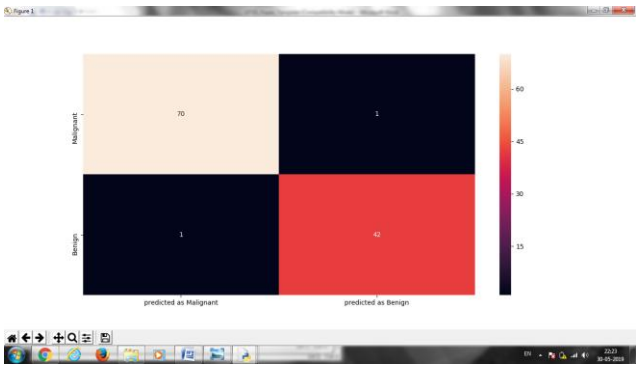

Figure2. The Heat map of predicted data SVM with Normalized data followed by Grid Search method

The above heat map shows that there are '1' false predictions which indicates 99% accuracy. This indicates that it is the best model for classification with high accuracy.

Table 4: Table below shows the precision, recall, F1-score and support using SVM (normalized data) with Grid search method

	Precision	Recall	F1-score	Support
M= 0.0	0.98	0.98	0.98	43
B=1.0	0.99	0.99	0.99	71
average	0.98	0.98	0.98	114

Table 5: Comparison of various performance measures with SVM classifier and using SVM classifier with grid search method.

	Precision	Recall	F1-score	Support
SVM	0.39	0.62	0.48	114
SVM-GSM	0.98	0.98	0.98	114

VI. .. CONCLUSION AND FUTURE SCOPE

This paper presents an optimal analysis of machine learning algorithms used for classification purpose. It includes the Support Vector Machine algorithm and their drawbacks which encouraged to build a proposed SVM-GSM model, for the classification of breast cancer. The SVM alone can't give the appropriate results. So, as to improve the performance of our model we have used SVM with Normalized data along with Grid Search Method .This SVM-GSM model gives us 99% accurate results with '1' false predictions and all the other performance measures: Precision, Recall, F1-score and Support are also higher than those with SVM .This proposed model will be very helpful in the medical field for diagnosis or classification of other diseases such as lung cancer, brain cancer etc.

APPENDIX

A.Sklearn: Scikit-learn or sklearn is an open software available for machine learning library for

the Python programming language. It includes different classification, regression and clustering algorithms which includes support vector machines,, random forests ,k-means and DBSCAN, and is programmed such as it can be interoperate with the Python numerical and scientific libraries NumPy and SciPy.[6]

REFERENCES

1. Abien Fred M. Agarap," On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset" ICMLSC 2018, February 2–4, 2018, Phu Quoc Island, Viet Nam..
2. Meenakshi M. Pawar, et al., "Genetic Fuzzy System (GFS) Based Wavelet co-Occurrence Feature Selection In mammogram Classification For Breast Cancer Diagnosis". Elsevier Gmbh 2016.
3. Wenbin Yue, et al., " Machine learning with applications in breast cancer diagnosis and prgnoisis"2018
4. Iliyan Mihaylov "Application of Machine Learning Models for Survival Prognosis in Breast Cancer Studies" The 18th International Conference on Artificial Intelligence: Methodology, Systems, Applications, 2019.
5. Fukunaga, K. & Hostetler, L. (1975) "k nearest-neighbor, Bayes risk estimation", IEEE Trans. Information Theory, 21(3): 285-293.
6. https://en.wikipedia.org/wiki/Scikit-learn.

AUTHOR PROFILE

Anita Paneri, Mtech student, Department of CSE, Geetanjali Institute of Technical Studies, Udaipur.