Asian Journal of Convergence in Technology
ISSN NO: 2350-1146 I.F-5.11

Volume V Issue III

Application of Machine Learning in Classification and Prediction of Breast Cancer

1.Pavan N. Kunchur,2.Vidyadheesh Pandurangi,3.Khasgatesh Hiremath,4.Mallikarjun Kolar

1,2.Asst,Professor,Department of Computer Science and engineering,KLS,Gogte institute of Technology,Belagavi
3,4 Department of Computer Science and engineering,KLS,Gogte institute of Technology,Belagavi
pnkunchur@git.edu, vjpandurangi@git.edu

***Abstract*—Cancer misdiagnosis is extremely common. We attempt to build different machine learning models that can predict occurrences of cancer traits in a patients. Being said that cancer is often misdiagnosed, when it comes to cancer, spotting the disease earlier can quite literally mean the difference between life and death. Predictive models obtained by using machine algorithms may be a key in such cases. This can be used by any medical institutes for faster, economical and accurate cancer diagnosis. Machine learning incorporates varieties of statistical, probabilistic and optimization techniques that allow computers to "learn" from past examples and to detect hard-to-diagnosed patterns from massive, noisy or complex datasets This project allows us make fast, real-time and accurate diagnosis and prediction of breast cancer. The software uses support vector machine algorithm to do the prediction and diagnosis of breast cancer. The simplicity and almost accurate results for support vector machine algorithm is very suitable for implementation.**

Keywords—component; formatting; style; styling; insert (key words)

I. INTRODUCTION

Cancer is life threatening diseases. Proper treatment of cancer saves lives of cancer patients. Identification of benign and malignant is a crucial task since and it facilitates the further treatment of cancer. In identification of benign and malignant conditions, image of the targeted area helps the doctor in further diagnosis. With the advent of modern photography techniques, image of targeted part of the body are more reliable [3]. Breast cancer tumors can be categorized into two cases -

- *Benign(Noncancerous):* These cases are supposed to be non cancerous. They do not pose a threat. On rare occasions it could turn into cancer status. An immune system called "sac" segregates benign tumors from other cells which can be easily removed from the body.
- *Malignant(Cancerous):* Malignant cancer begins from abnormal cell growth and quickly spread or invade nearby tissue. Usually the nuclei of the malignant tissue are bigger than the ones present in normal tissue, which can be life threatening in future.

The genetic mutation in the DNA of breast cancer cells causes breast cancer. Some mutations may be inherited, some may develop randomly over time, or may be due to environmental exposures or lifestyle factors.

The second major cause of women's death is breast cancer (after lung cancer). 246,660 of women's new cases of invasive breast cancer are expected to be diagnosed in the US during 2016 and 40,450 of women's death is estimated. Breast cancer represents about 12% of all new cancer cases and 25% of all cancers in women [5].

Some of the major concerns of healthcare and medical industries are the misdiagnosis, costly treatment, transportation constraints, slow and time-consuming process, unavailability of a required number of specialized doctors. Machine learning is a type of artificial intelligence (AI) that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. Machine learning algorithms are categorized as being unsupervised or supervised. The algorithms that need to provide both input and the correct output, and then the algorithm improves the model by reducing the error during training are called supervised learning algorithms. We can input new test cases for prediction once the model is built. The algorithms that do not require to be trained with actual outcome data are called unsupervised learning algorithms. It learns the model on its own based on the parameters.

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression problems [1]. The main application of Support Vector Algorithm (SVM) is for classification problem. In this algorithm, we plot each data item as a point in n-dimensional space involving input features, with the value of each feature being the value of a particular coordinate plane. We can then perform classification by finding the hyperplane that differentiates the two classes very well.

II. LITERATURE REVIEW

Breast cancer has been major concern in the present world. Various diagnoses has been approached by several machine learning techniques, this paper presents study on classification of breast cancer using support vector machines.

Methodologies used:
1.tensor flow library was used to code the above mentioned algorithm with the help of other libraries like matplotlib and numpy.

2.Data set used were Wisconsin diagnostic breast cancer[WDBC] data set and features were computed from digitized image of fine needle aspirate[FNA]

3.dataset consists of features which were computed from digitized images of fine needle aspirate(FNA) tests on a breast mass. with 569 data point inputs consisting of 212 malignant and 357 benign.

4.comparing with other machine learning algorithms this above proposed theory of machine learning can also be done in other ML algorithms like multi-layer perceptron, nearest labor search and softmax regression.

Abien Fred M. Agarap presents an application of different machine learning algorithms, including the proposed GRU-SVM model for the diagnosis of breast cancer. All presented ML algorithms exhibited high performance on the binary classification of breast cancer, i.e. determining whether benign tumor or malignant tumor. Consequently, the statistical measures on the classification problem were also satisfactory. Hiba Asri [2] analyses medical data, various data mining and machine learning methods are available. An important challenge in data mining and machine learning areas is to build accurate and computationally efficient classifiers for Medical applications. In this study, he employed four main algorithms: SVM, NB, k-NN and C4.5 on the Wisconsin Breast Cancer (original) datasets. Later in 1992 Vapnik, Boser& Guyon suggested a way for building a non-linear classifier. They suggested using kernel trick in SVM latest paper. Vapnik& Cortes published this paper in the year 1995 From then, SVM classifier treated as one of the dominant classification algorithms. However, Svm is a supervised learning technique. When we have a dataset with features & class labels both then we can use Support Vector Machine. But if in our dataset do not have class labels or outputs of our feature set then it is considered as an unsupervised learning algorithm. Ahmad LG [3] has explored risk factors for predicting breast cancer by using data mining techniques. Each method has its own limitations and strengths specific to the type of application. His results show that SVM outperforms both Decision Tree and MLP in all the parameters of sensitivity, specificity and accuracy. SVM is the best predictor of breast cancer recurrence

III. EXISTING SYSTEM

Current system that exists is manual and repetitive. The current cancer diagnostic methods are highly expensive and time consuming. Moreover, these diagnostic methods are not accurate enough and often misdiagnosed. Recent survey reports say that in India there are currently, around 1,80,00,000 people are suffering from cancer. The ratio of number of oncologist to number of cancer patients is 1:2000. Expertise of the oncologist is rarely known. In 2012, the Indian government stated 22% of its population is below its official poverty limit. Thus, many people cannot afford the initial diagnosis tests itself. Also, Transportation and culturing of cancer takes 7-10 days on an average, thus, faster diagnosis very critical for higher cancer stage patients.

IV. IMPLEMENTATION FLOWCHART

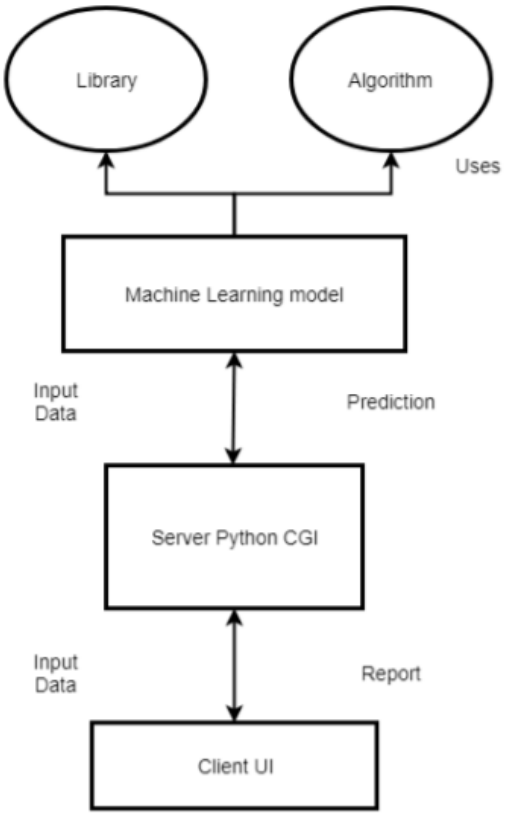

Fig. 1. System Architecture

V. APPLICATION OF CLASSIFICATION AND PREDICTION ALGORITHM

- Bioinformatics – It includes protein classification and cancer classification. We use SVM for identifying the classification of genes, patients on the basis of genes and other biological problems.
- Face detection – SVM classify parts of the image as a face and non-face and create a square boundary around the face.
- Text and hypertext categorization – SVMs allow Text and hypertext categorization for both inductive andtransudative models. They use training data to classify documents into different categories. It categorizes on the basis of the score generated and then compares with the threshold value.
- Classification of images – Use of SVMs provides better search accuracy for image classification. It provides better accuracy in comparison to the traditional querybased searching techniques.
- Handwriting recognition – We use SVMs to recognize hand written characters used widely.

VI. EXPLORATORY DATA ANALYSIS

We use Wisconsin (Diagnostic) Breast Cancer Data Set for training the model[6].

The features chosen for diagnosis are:

- ID number
- Diagnosis (M = malignant, B = benign)

Ten real-valued features are computed for each cell nucleus:

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness (perimeter^2 / area - 1.0)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry fractal dimension ("coastline approximation" - 1)

Combination of some features may lead to breast cancer traits in the patient. What and How features contribute the cancerous traits can be more understood by making use of exploratory data analysis techniques which is discussed next.

A. Swarm plot

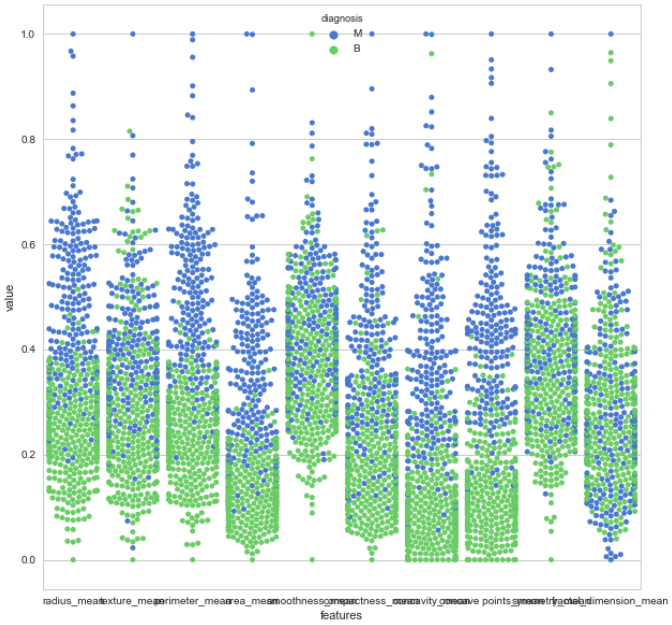

Fig 2. Swarm plot representation of features

This graph shows the distribution of various features involved according to classes. It show that it is possible to differentiate between malignant and benign tumors.

B. Heatmap

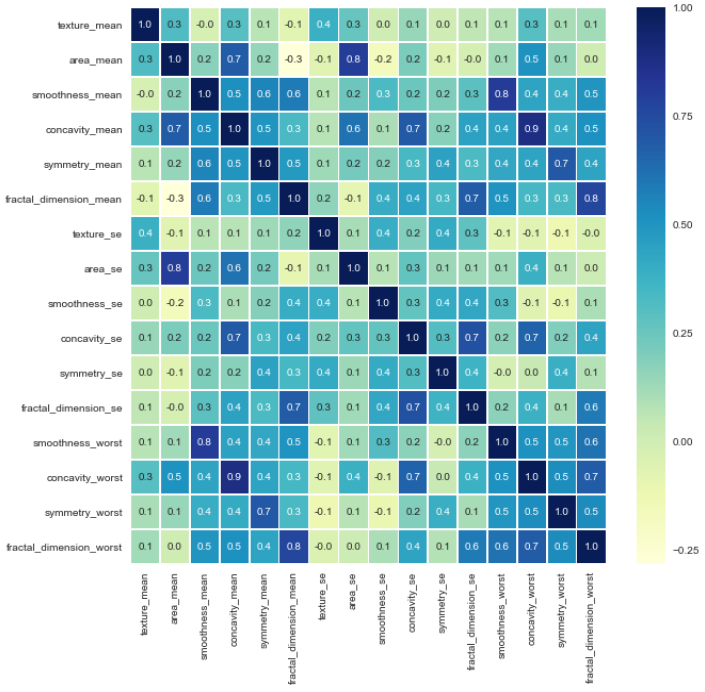

Fig. 1. Heatmap plot representation of features

This heat map shows the correlation of different features in the dataset. Darker shades represent high correlation.

C. Violin plot

The following plot shows the distribution of features according to their median values.

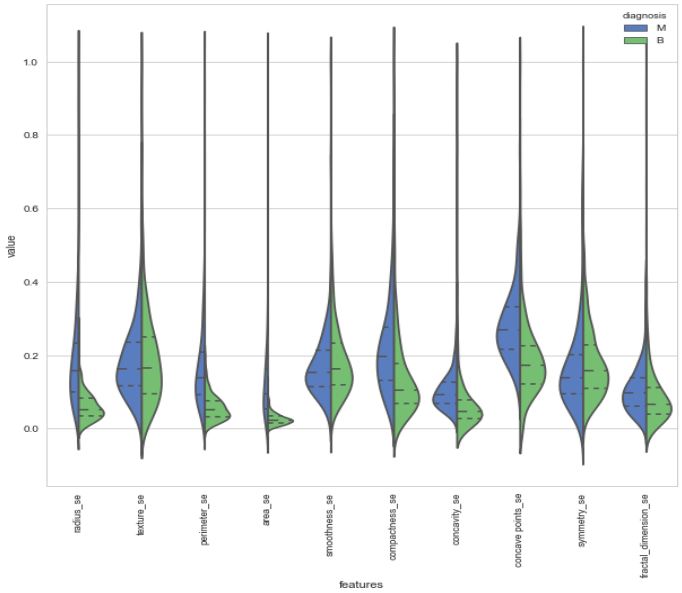

Fig. 2. Violin plot representation of features

VII. SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for regression as well as classification purposes [2].We will make use of SVM for classification in this post. They work by finding a hyperplane that divides the dataset into two different classes as shown in the below figure.

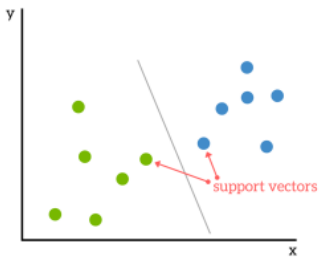

Fig. 3. Working of Support Vector Machines

C. Support Vectors

Support Vector Machine falls in the supervised machine learning algorithm. It is used for regression and also to classify. They find a hyperplane which in turn bifurcates the features.

D. Hyperplane

An easy to understand example is while classifying two features. The hyperplane in this case separates the data linearly by a line and classifies the data. The farther are the points from the hyperplane, the more confidence that the classification has been done correctly. We therefore expect our data points to be far from hyperplane, while being on right side. Whenever new testing data is added, the side on which the points lie will decide the class to which it belongs to.

E. Designing right hyperplane

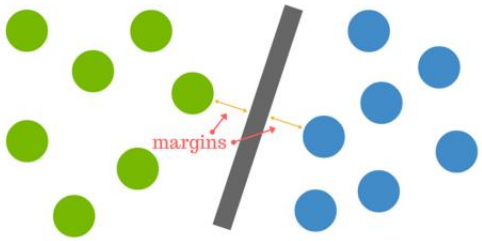

Fig. 4. Large margin classifier

The distance in between the data point from any of the set nearest to the hyperplane and the hyperplane itself is called as margin. The primary objective here is to select greatest possible margin between any random point on the training set and hyperplane. This in turn increases the chances of classifying the new data correctly[4].

F. Advantages of SVMs

- Accuracy.
- Works well on smaller cleaner datasets.
- It can be more efficient because it uses a subset of training points.

G. Disadvantages of SVMs

- It suited to larger datasets as the training time with SVMs can be high.
- Less effective on noisier datasets with overlapping classes.

VIII. IMPLEMENTATION

We use Python for the implementation of the code. We have used Scipy, Numpy, Scikit-learn libraries are used implementation of designed algorithm [7]. We split the dataset into test and train dataset and then train the model using train datatest

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...

Fig. 5. Features in dataset

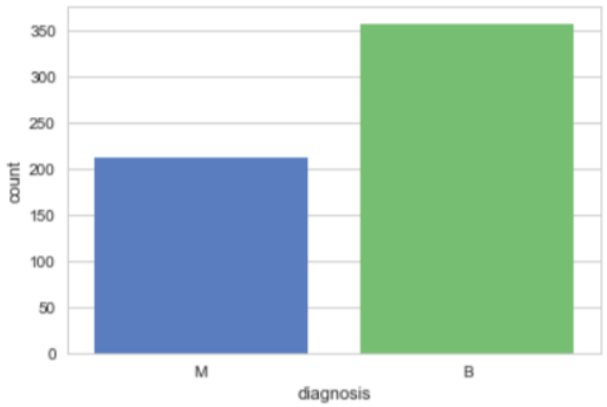

Fig. 6. Count plot

```
X_train,X_test,y_train,y_test=train_test_split(x,
y, test_size=0.33, random_state=42)
svc = SVC(kernel ='linear',C=.1, gamma=10,
probability =True)
svc.fit(x,y)
y_pred=svc.fit(X_train,y_train).predict(X_test)
```

IX. Conclusion and Future Work

This paper meets its objective of being able to classify the test case to possibility of being cancerous or non-cancerous using support vector machines. The result obtained by this project were highly accurate with accuracy almost 95% over the test cases that were examined.

The parameters contributing to breast cancer were scrutinized using different graphical methodologies. The project has been implemented in Python programming language along with CGI to handle client-server requests. Modern and scientific libraries were used for numerical computation. The project has been version controlled on GitHub and hence making it easier for further evolution of the project. More complex analytical and predictive models like Deep Neural Networks, Restricted Boltzmann Machines can be used to have a comparative prediction for test cases. In conclusion, SVM hasproven its efficiency in Breast Cancer prediction and diagnosis and achieves the best performance in terms ofprecision and low error rate.

X. References

[1] Boser, B. E., I. Guyon, and V. Vapnik (1992). A training algorithm for optimal margin classifiers . In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pages. 144 -152. ACM Press 1992.

[2] Chih-Wei Hsu, Chih-Chung Chang, and Chih Jen Lin. “A Practical Guide to Support Vector Classification” . Deptt of Computer Sci. National Taiwan Uni, Taipei, 106, Taiwan http://www.csie.ntu.edu.tw/~cjlin 2007

[3] Mikhail V. Blagosklonny (2005) Molecular theory of cancer, Cancer Biology & Therapy, 4:6, 621-627, DOI: 10.4161/cbt.4.6.1818

[4] Bernhard Scholkopf, and Alexander J. Smola. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, 2001.

[5] Cancer Key Facts, [online] Available: http://www.who.int.

[6] Breast Cancer Wisconsin Dataset. Available at: UCI Machine Learning Repository.

[7] Lars Buitinck (ILPS), Gilles Louppe, Mathieu Blondel, Fabian Pedregosa (INRIA Saclay - Ile de France), Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort (INRIA Saclay - Ile de France, LTCI), Jaques Grobler (INRIA Saclay - Ile de France), Robert Layton, Jake Vanderplas, Arnaud Joly, Brian Holt, GaëlVaroquaux (INRIA Saclay - Ile de France). “API design for machine learning software: experiences from the scikit-learn project”. European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases (2013).

[8] Lorne Mason and Peter L. Bartlett and Jonathan Baxter. Improved Generalization Through Explicit Optimization of Margins. Machine Learning, 38. 2000.

[9] P. S and Bradley K. P and Bennett A. Demiriz. Constrained K-Means Clustering. Microsoft Research Dept. of Mathematical Sciences One Microsoft Way Dept. of Decision Sciences and Eng. Sys. 2000.

[10] EndreBoros and Peter Hammer and Toshihide Ibaraki and Alexander Kogan and Eddy Mayoraz and Ilya B. Muchnik. An Implementation of Logical Analysis of Data. IEEE Trans. Knowl. Data Eng, 12. 2000.

[11] Yuh-Jeng Lee. Smooth Support Vector Machines. Preliminary Thesis Proposal Computer Sciences Department University of Wisconsin. 2000.

[12]] Justin Bradley and Kristin P. Bennett and Bennett A. Demiriz. Constrained KMeans Clustering. Microsoft Research Dept. of Mathematical Sciences One Microsoft Way Dept. of Decision Sciences and Eng. Sys. 2000.