An Efficient System for Early Diagnosis of Breast Cancer using Support Vector Machine

Ahatsham, Anupam Singh, Vivek Shahare, Nitin Arora

Abstract: ***There are many lives lost every year due to cancer and among them; among the women breast cancer causes the most deaths. For the better prediction of breast cancer risks, numerous studies have been undertaken incorporating data mining techniques. 1.1 million Cases of breast cancer were reported in 2004. It has been seen over the years that, that the numbers increase with the increasing industrialization and urbanization. It was earlier observed that mostly affected countries with breast cancer were high income countries such as America but now a days it is also very serious issue in middle and low income countries like Africa, Latin America and Asia. The main objective of this paper is to create a model which can more efficiently and accurately categorize a cancer as malignant or benevolent based on interpretation of the numerical values of attributes of ultrasound images of breast cancer. In this paper various data mining algorithm used like SVM(Support Vector Machine) for prediction and compared it with various other algorithms such as CART, Logistic Regression, KNN for the best training and test accuracy. SVM algorithm gives the most accurate results among the rest algorithm.***

Keywords ***: Predictive Analysis, SVM, Breast Cancer, KNN, CART, Logistic Regression.***

I. INTRODUCTION

A study carried out in India on Breast cancer showed that everyone in twenty eight women develop the disease of breast cancer during their life span only. The rate of its occurring is higher in the urban areas and the numbers being one in twenty two. The average age in India comprising the higher risk of developing this cancer is forty three to forty six years. The study also showed that women in the western parts like in America tend to develop breast cancer between fifty three to fifty seven years and are very much prone to it. According to the data given by Indian Council of Medical research (ICMR) (2006-2008) Women living in metro cities like Bangalore, Mumbai, Chennai, Delhi etc. have higher rates of having breast cancer accounting up to twenty eight to thirty five percent.

Revised Manuscript Received on October 30, 2019.

* Correspondence Author

Ahatsham*, Department of Computer Science, University of Petroleum & Energy Studies, Dehradun, India. Email: ahatsham@ddn.upes.ac.in

Anupam Singh, Department of Computer Science, University of Petroleum & Energy Studies, Dehradun, India., Email: anupam.singh@ddn.upes.ac.in

Vivek Shahare, Department of Computer Science, University of Petroleum & Energy Studies, Dehradun, India., Email: vshahare@ddn.upes.ac.in

Nitin Arora, Department of Computer Science, University of Petroleum & Energy Studies, Dehradun, India., Email: narora@ddn.upes.ac.in

Currently, the only pharmaceutical way to verify that the tumor is malignant or benign is biopsy. But it is a very painful and time taking process. Many times it happens that the results are not clear and the biopsy needs to be repeated in order to confirm the presence of tumor. Therefore, this procedure involves increased cost and delayed confirmed results.

As a result, Computer-Aided Detection (CAD) is developed to get rid of this tedious procedure of repeated and painful biopsies and helping radiologists to find out the accurate diagnosis of breast cancer, CAD system is use to automatically detect breast cancer, So to detect breast cancer automatically lots of scientists are contributing in developing CAD system. CAD systems can be used as an efficient tool to diagnose people in rural areas where high level treatment is not possible. This will also help the doctors in detection of breast cancer in early stages and this in turn will help reduce the death rates by timely medication. The ultrasound breast cancer images are described by various attributes like texture, color, smoothness, concavity, symmetry, fractal dimension etc. and we have used these values to develop a model to classify whether a tumor is malignant or benign. Here main objective is to detect a tumor as malignant or benign based on the attributes present in the dataset.

II. EXISTING SYSTEMS

1.1.1 Clinical Breast Examinations, Breast Self-Examinations and Mammography

To detect early breast cancer Breast self-examination (BSE) method is used, it is the first screening method, in this method women herself to check for the any possible lump in her breast.

On the other hand clinical breast examination (CBE) is used, in this health care provider will check for any possible lump in your breast. CBE has an affectability of 57.14% and a particularity of 97.11% [3]. Despite the fact that it doesn't allow one to decide harm with confirmation, it is valuable for recognizing suspicious breast sores.

For the most part alluded to as the highest quality level of breast imaging, the most popular type of breast imaging are mammography or screen-film mammography (SFM). In this process a low energy X-ray beam is used to examine the human breast for diagnosis. This process have a positive rate of 83% to 95% and a false-positive rate of 0.9% to 6.5%.

1.1.2 Full-Field Digital Mammography

Digital mammography, also known as Full-field computerized mammography (FFDM) is just an advanced mammogram.

As opposed to recording a picture in video form, FFDM records a picture in an electronic document. Mammography is basically divided into three discrete procedures: picture obtaining, picture show, and picture stockpiling, so each can be improved independently [7]. FFDM permits constant picture introduction, post image handling, and advanced stockpiling. Digital mammographic images can be transmitted over the network for helping the clinician and the radiologist for audit the picture in remote area this is also known as telemammpgraphy. [4]

1.1.3 Computer-Aided Detection (CAD)

Computer aided design (CAD) is used to read mammographic and find out the presence of breast cancer. It is a pattern recognition programming technique that distinguishes suspicious variations from the norm on pictures, checking them for the radiologist. Computer aided design likewise represents PC helped conclusion, which alludes to a framework that marks benign or malignant pictures, and the two abbreviations are frequently confounded. R2 image checker is the most prevalent CAD framework (R2 Technology, Inc., Santa Clara, CA), which consolidates discovery and finding of breast cancer in human body. [2]

1.1.4 Modalities Using Ultrasound

Sonomammography also known as Ultrasonography (US), is used for quick visualization of the breast tissue in human. Mostly it is done with the mammography for the study of lump in the body. Unlike various other techniques sonomammography can be done at any time without any special preparation. By using sonomammography breast cancer can be easily detected in the routine check-up.

III. LITERATURE REVIEW

A statistical study of the women diagnosed with breast cancer revealed a close association between the high mammographic breast density and the associated risk of breast cancer. Inference can be made that among women with a mammogram density less than 10% and women embodying this proportion in more than 75%, the latter was found to have the increased risk of breast cancer (odds ratio, 4.7; 95% confidence interval [CI], 3.0 to 7.4), when diagnosed under a test (odds ratio, 3.5; 95% CI, 2.0 to 6.2) or within a period of one year after a negative screening diagnosis (odds ratio, 17.8; 95% CI, 4.8 to 65.9). [5]

The conventional perceptions of women about the involved reduction technologies for breast cancer, the sense of associated risks, and the prevention guidelines, influence the advances in the risk assessment and curing technologies. These skeptical approaches are most commonly observed among the low-earning women or those which fall under minority. This difference in perception of a clinically adopted risk and a purposely overlooked risk needs to be explored furthermore. The qualitative techniques dig deep into the definitions of risk and prevention which are often neglected by the rigid scientific technologies. [7]

A hierarchical study showed that the chances of occurrence of breast cancer is higher in the women whose mother was confirmed with breast cancer before the age of 40 years (Relative Risk, 2.1 [95% confidence interval, 1.6 to 2.8]) than those without any history of breast cancer. This relative risk stooped low with 1.5 (95% confidence interval, 1.1 to 2.2) when diagnosed after the age of 70 years with the maternal age being more than the previously mentioned. A sibling with a medical background of breast cancer was also found to pose an increased risk of breast cancer. Women with a breast cancer diagnosed sibling was found to be under increased risk than the ones holding no such family log, with the relative risk of 2.3 (95% confidence interval, 1.6 to 3.4). Cases in which both the mother and the sibling were diagnosed with breast cancer, women were found to have a relative risk of 2.5 (95% confidence interval, 1.5 to 4.2) in contrast to those with no such tragic medical logs.

Twice the risk was observed in the cases where women's mother was diagnosed with breast cancer before an age of 40 years or a sister with breast cancer, and was just as high in the cases where the mothers were identified to have breast cancer at the age of 70 years or more. [6]

IV. METHODOLOGY

3.1 Identifying the problem and getting data

Identify the types of information contained in dataset. In this various modules are used to import external datasets for the purpose of getting to know/familiarize with the data to get a good grasp of the data and think about how to handle the data in different ways.

a) Identify the problem:

In recent years most widely recognized disease among the women is breast cancer, In United states almost 1 of every 3 women are suffering with this problem, and now it is spreading in all over the world it's the second driving reason for cancer passing among women's.

b) Expected outcome:

Given breast cancer results from breast fine needle aspiration (FNA) is a diagnostic procedure which is used to investigate lumps in the human breast it is a very quick and simple procedure to perform, In this we uses a very thin syringe which will removes a sample of cells, tissues or fluid from the abnormal area from the breast. Here we have a model that can be used to classify breast cancer by two training classification

- 1 denotes that Malignant(Cancerous) results are present
- 0 denotes that Benign (Not Cancerous) results are absent

3.2 Exploratory Data Analysis

Explore the variables to assess how they relate to the response variable in this notebook. The data is explored using data exploration and visualization techniques using python libraries (Pandas, Matplotlib, Seaborn). Familiarity with the data is important which will provide useful knowledge for data pre-processing.

After getting a good intuitive sense of the data, next step involves taking a closer look at attributes and data values.

Objectives of Data Exploration

Exploratory Data Analysis is a way to deal with examining data sets to condense their primary attributes, frequently with visual strategies and it ought to be done before any modeling. This is on the grounds that it is significant for a data researcher to understand data without assuming. The end results of data exploration can be amazingly helpful in getting a handle on the structure of the data, the values being distributed, and the presence of extreme values and interrelationships within the data set.

3.3 Pre-Processing the data

The aim in this notebook is to increase the predicting power of the analytical model being used. The concept of feature selection helps in reducing high-dimensional data, transformation and feature extraction. It is important to properly prepare the data before developing the predictive models.

Data preprocessing is a crucial step for any data analysis problem. It is advised to prepare the data in such a manner so that it can expose the structure in the best possible way by the use of machine learning algorithms that we intend to use.

Following activities are used for data pre-processing

- Replace all the categorical values with numerical values.
- Make a model to handle all the missing values present in the dataset.
- Perform normalization on the features so that the one feature does not dominates other feature present in the dataset.

This notebook deal with finding and filtering predictive features of the data such that it results to enhanced predictive power of the analytics model.

3.4 Predictive model using Support Vector Machine (SVM)

Construct predictive models to predict the diagnosis of a breast tumor. In this notebook a predictive model is constructed using SVM machine learning algorithm to predict the diagnosis of a breast tumor. The diagnosis of a breast tumor is a binary variable (benign or malignant).

Predictive model using SVM

One of the main implementations of SVM is predictive modeling. Support Vector machines are very popular as SVM can easily process nonlinear data. So by the help of SVM one can use a linear algorithm for fitting a linear model to the data.

3.5 Optimizing the Support Vector Classifier

Predictive models are constructed to predict the diagnosis of a breast tumor. In this notebook, parameters of the SVM Classification Model are tuned using scikit-learn.

Optimizing the SVM Classifier

Machine Learning models can use different number of parameters such that their behavior can be set according to the given problem. Here search problem can be seen as finding the best possible combination of attribute.

3.6 Comparison between different classifiers

Machine learning consists of many standard workflows that can be automated. Using Python scikit-learn, it’s easy to use define and automate these workflows clearly.

- They make your workflow much easier to read and understand.
- They enforce the implementation and order of steps in your project.
- These in turn make your work much more reproducible.

V. RESULT AND DISCUSSION

4.1 Histogram generated for data visualization

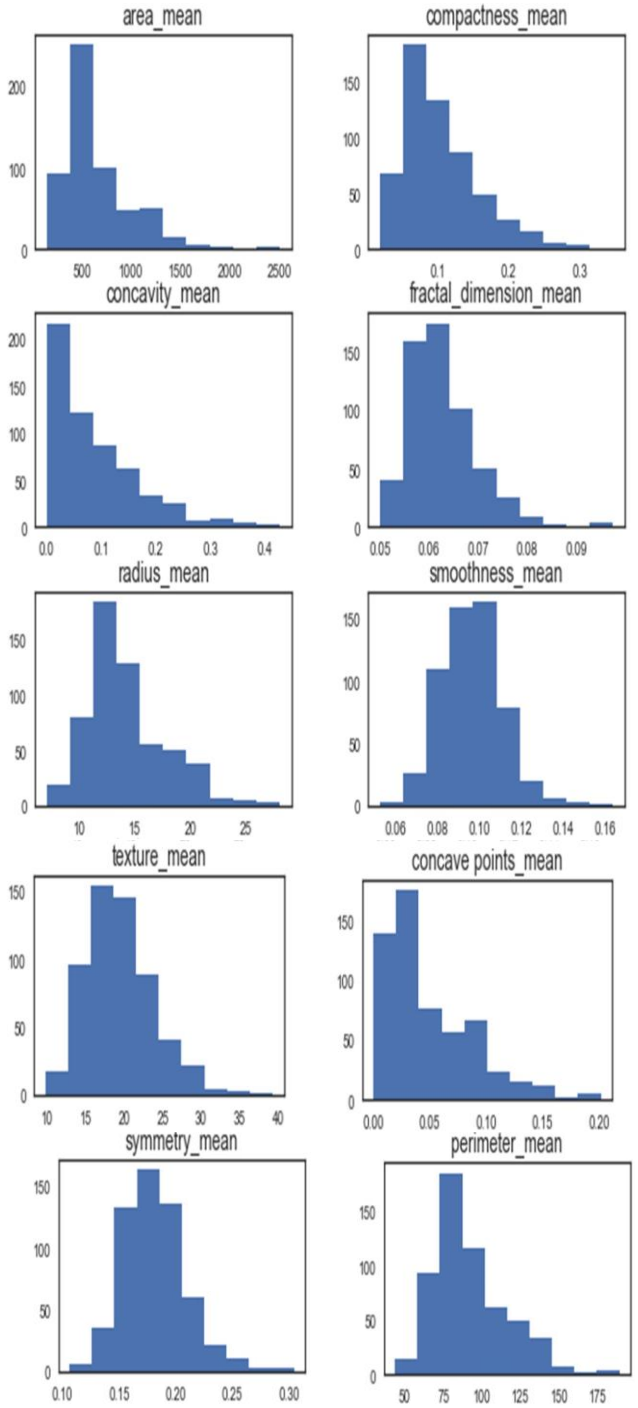


Fig. 1: Correlation of attribute

This shows that the attribute concavity, and concavity point follows the exponential distribution whereas texture, smooth and symmetry attributes may have a Gaussian or nearly Gaussian distribution.

4.2 Correlation Matrix between the attributes

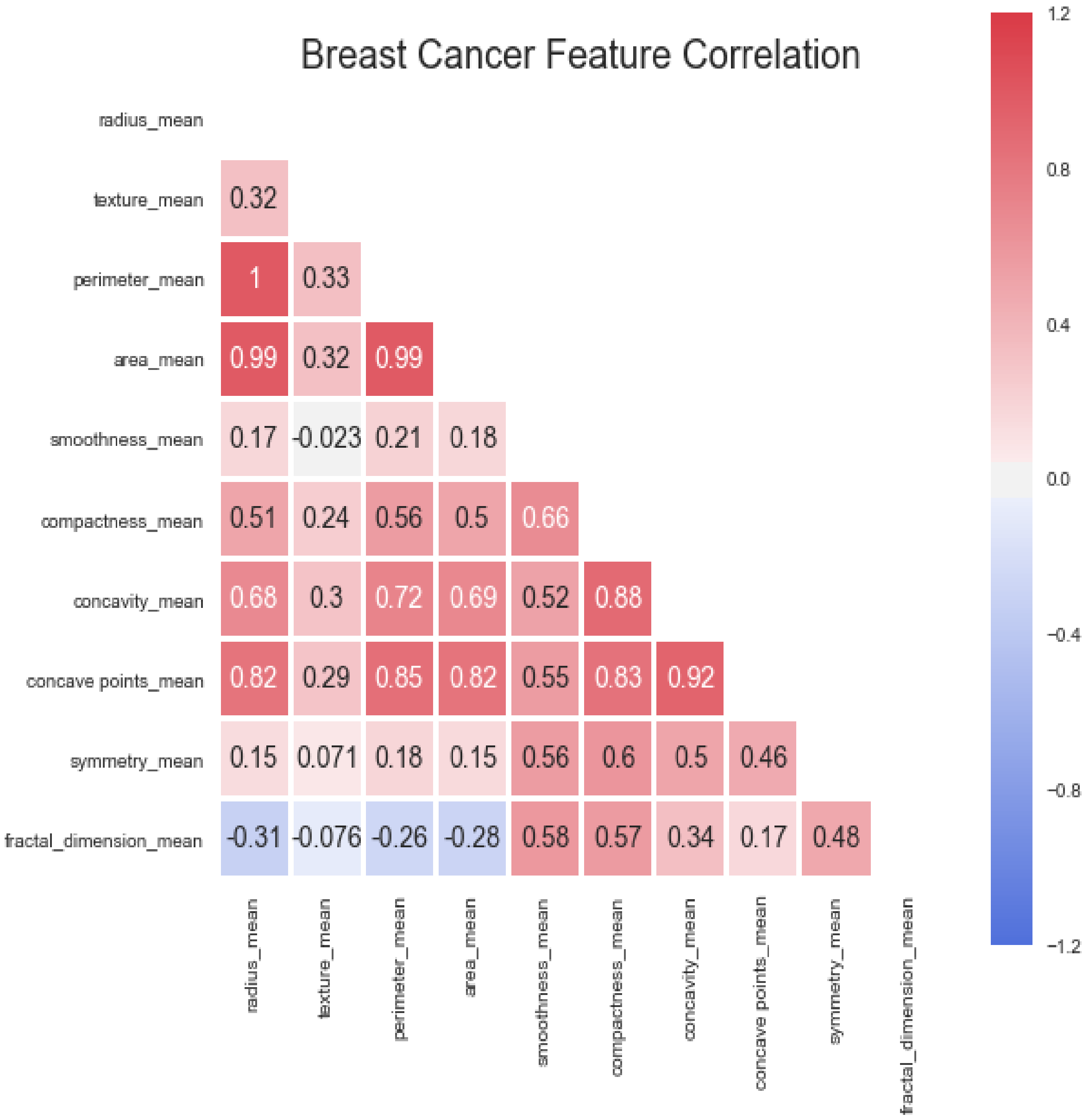


Fig. 2: Correlation Matrix

Correlation between various attributes is very useful when summarization of data is required. By seeing the correlation matrix it is observed that there is a strong relationship exist between the mean value parameters

- Here some strong positive correlation has been seen with mean area of the tissue nucleus and mean value of radius parameter.
- Correlation between the concavity and area, concavity and perimeter is showing the moderate positive correlation between them (0.5 to 0.75).
- Similarly by this correlation matrix some strong negative relation are also seen like fractal dimensions with other attributes like radios, texture etc.

4.3 Scatter Plot generated for data visualization

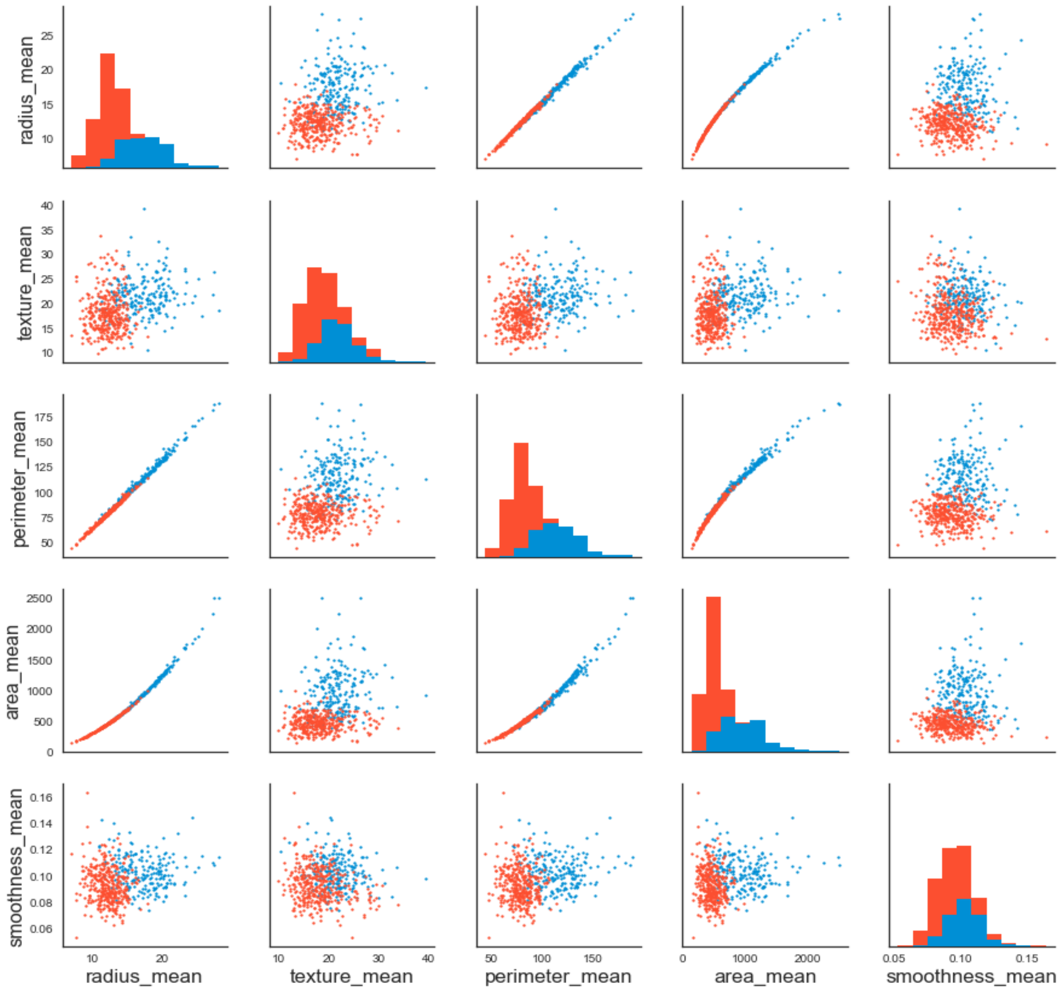


Fig. 3: Scatter Plot

- By seeing this we can understand that mean value of radius, perimeter, area points can be used for cancer detection in human body. These values are showing the correlation with malignant tumors.
- On the other hand mean values of the attributes like texture, smoothness, symmetry are not showing a particular preference of one diagnosis over the other.

4.4 Scree Plot

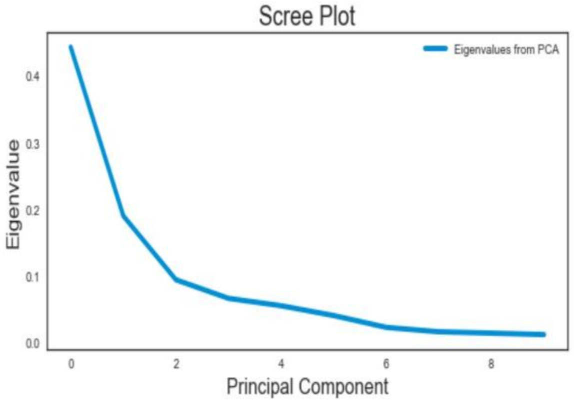
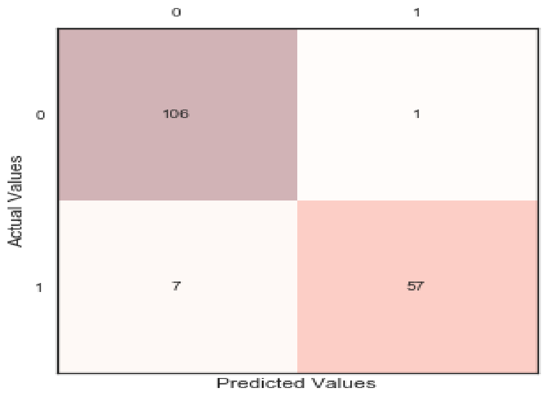


Fig. 4 : Scree Plot

By seeing the scree plot between Principal Component and Eigen value it is observed that after principal component value "2" there is obvious change in slope has been seen. Therefore on the basis of scree plot we can say that the first three elements should be retained.

4.5 Confusion Matrix



	0	1
0	106	1
1	7	57

Fig. 5: Confusion Matrix

Here are two classes for checking the cancer 1 and 0.
1 indicates the presence of cancer cells called Malignant
0 indicates the absence of cancer cell in the body

- In this 171 patients were being tested for cancer.

- In these 171 cases, the classifier predicted "1" (Malignant) 58 times, and "0" (Absence) 113 times.
- In reality, 64 patients in the sample have the disease, and 107 patients do not.

4.6 Decision Boundaries of different types of Classifiers

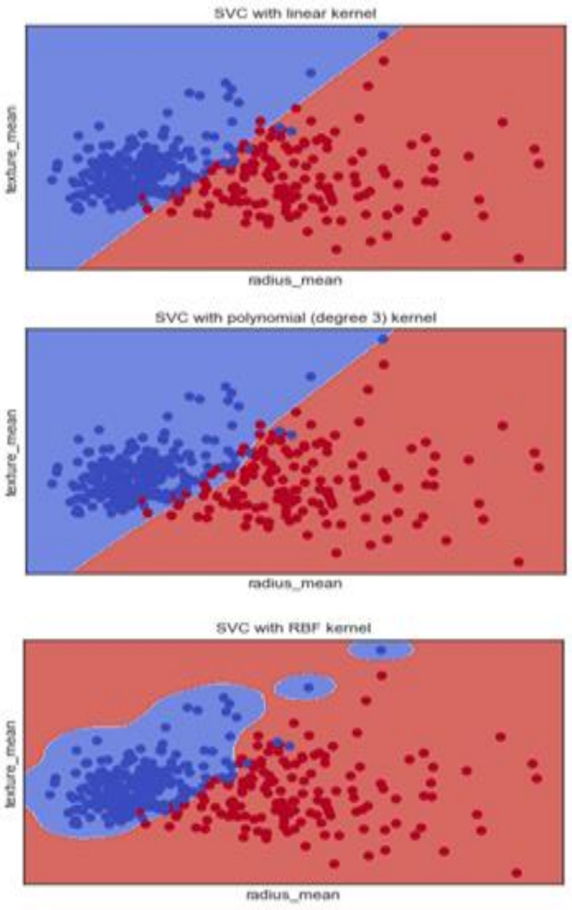
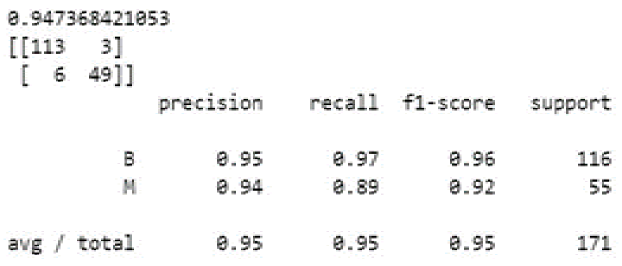


Fig.6: Decision Boundaries of different types of Classifiers

This shows that if the dataset is standardized (that means all the attributes should have mean value of zero and standard deviation of one) SVM will give the better result.

4.7 Final Model



```
0.947368421053
[[113   3]
 [  6  49]]
             precision    recall  f1-score   support

          B       0.95      0.97      0.96       116
          M       0.94      0.89      0.92        55

avg / total       0.95      0.95      0.95       171
```

In this paper SVM is applied to 171 patients to check for the breast cancer and SVM is giving the highest accuracy of 94%.

VI. CONCLUSIONS AND FUTURE SCOPE

Now a day's breast cancer detection is very significant in the field of Biomedical because of this several deaths occurred in the past years and it is spreading more and more in the metro cities as well as in the rural areas. This paper purposed a model to detect the breast cancer efficiently so that patient can be cured. Here SVM classifier gives the highest accuracy of 94%.

Only 171 patients are being examined here so running time of SVM can become challenge when dataset size increases, and some neural network techniques can also be applied for better and fast result.

REFERENCES

1. J. Tang, S. Agaian, I. Thompson, Computer aided detection or diagnosis (CAD) systems, IEEE Syst. J. 8 (3) (2004) 907–909.
2. H. Peng, F. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.
3. Y. Li, D.F. Hsu, S.M. Chung, Combining multiple feature selection methods for text categorization by using rank-score characteristics, in: Proceedings of 21st IEEE Conference on Tools with Artificial Intelligence, pp. 508–517.
4. W. Bouaguel, G.B. Mufti, M. Limam, A fusion approach based on wrapper and filter feature selection methods using majority vote and feature weighting, in: Proceedings of International Conference on Computer Applications Technology, 2013, pp. 1–6.
5. Boyd, Norman F.,et al. "Mammographic density and the risk and detection of breast cancer." New England Journal of Medicine 356.3 (2007): 227-236.
6. Salant, Talya, et al. ""Why take it if you don't have anything?" Breast cancer risk perceptions and prevention choices at a public hospital." Journal of general internal medicine 21.7 (2006): 779-785.
7. Colditz, Graham A., et al. "Family history, age, and risk of breast cancer: prospective data from the Nurses' Health Study." Jama 270.3 (1993): 338-343.
8. Hassan Khotanlou, OlivierColliot, JamalAtif, IsabelleBloch , "3D brain tumor segmentation in MRI using fuzzy classification, symmetry analysis and spatially constrained deformable models", Fuzzy Sets and Systems, Volume 160, Issue 10, pp.1457- 1473, 16 May 2009.
9. E.C.Fear, P.M.Meaney, and M.A.Stuchly,"Microwaves for breast cancer detection", IEEE potentials, vol.22, pp.12-18, February-March 2003.
10. Ahatsham, Nitin Arora and Kamal Preet Singh, "An Approach towards Real Time Smart Vehicular System Using Internet of Things", International Journal of Research in Engineering, IT and Social Sciences, ISSN 2250-0588, Volume 08 Issue 10, October 2018, Page 52-56
11. R. Sheikhpour, N. Ghassemi, P. Yaghmaei, J. M. Ardekani, and M. Shiryazd, "Immunohistochemical assessment of p53 protein and its correlation with clinicopathological characteristics in breast cancer patients," Indian Journal of Science and Technology, vol. 7, no. 4, pp. 472–479, 2014.
12. American Cancer Society (ACS), Breast cancer facts & figures, American Cancer Society, 2018.
13. Padmanabhan, S., Sundarajan, R.: "Enhanced Accuracy of Breast CancerDetection in Digital Mammograms using wavelet analysis." IEEE Trans. Imag. Proc. (2012)
14. Spandana, P., Rao, K.M.M., Jwalasrikala, J.: "Novel Image Processing Techniques for Early Detection of Breast Cancer. In: Matlab and Lab View Implementation." IEEE Point-of-Care Healthcare Technologies (PHT), Bangalore, India, pp. 16–18 (2013)

AUTHORS PROFILE

Ahatsham was born in Roorkee, India in 1993. He received his B.tech degree from Uttrakhand Technical University in 2014 and M.tech degree from NIT Nagpur in 2016.He is currently working as Assistant Professor in University of Petroleum ad Energy Studies, Dehradun, India. His research interests include data mining, machine learning and data streaming.

Anupam Singh is Assistant Professor in Department of Informatics, School of Computer Science, University of Petroleum & Energy Studies, Dehradun. He is pursuing Ph.D. from Dr. A P J Abdul Kalam Technical University Lucknow. Formerly UPTU Lucknow). He has done B. Tech. in 2004, M. Tech. in 2011 from Uttar Pradesh Technical University Lucknow. His area of interest are Formal Methods, Distributed System and Database System. He is Reviewer of many referred journals. He has evaluated many presentations in International Conference as Session Chair.

Vivek Shahare is Assistant Professor in School of Computer Science and Engineering, UPES, Dehradun. He has completed his M.Tech from Visvesvaraya National Institute of Technology (VNIT), Nagpur in 2016 and B.Tech from Government College of Engineering, Amravati in 2012. He has 5 years of experience in Teaching, Research and Industry. His areas of interest are Bioinformatics and Theoretical Computer Science.

Nitin Arora received B. Tech. (Comp. Sci. Engg.) from NIT ALLAHABAD in 2008 and M. Tech. (Comp. Sci. & Engg.) in 2012 with GOLD MEDAL from Uttarakhand Technical University, Dehradun, the member of IAENG (USA), ISOC (USA) has published over Fifteen plus research papers in National and International Journals/Conferences and IEEE proceeding publication in field of Data Structures and Algorithms, Mobile Ad-hoc networks & Digital Image Processing. He is working as an Assistant Professor (Senior Scale) at University of Petroleum and Energy Studies (UPES), Dehradun.