A Novel Approach for Improving Breast Cancer Risk Prediction using Wavelet based Feature Extraction and SVM

[1]Madhuri Maru, [2]Saket Swarndeep
[1]Student of Master of Engineering, [2]Assistant Professor
[1]Information Technology,
[1]L.J. Institute of Engineering and Technology, Gujarat Technological University, Ahmedabad, Gujarat, India.

Abstract: Breast cancer represents one of the diseases that make a high number of deaths every year. Classification and data mining methods are an effective way to classify data. Here, a common misconception is that predictive analytics and machine learning are the same thing where in predictive analysis is a statistical learning and machine learning is pattern recognition and explores the notion that algorithms can learn from and make predictions on data.
In this paper, we are addressing the problem of predictive analysis by adding machine learning techniques for better prediction of breast cancer. In this, a performance comparison between different machine learning algorithms: Support Vector Machine (SVM) with GRB technique on the Machine Learning Repository Breast Cancer Coimbra (original) datasets is conducted. The main objective is to assess the correctness in classifying data with respect to efficiency and effectiveness of hybrid algorithm in terms of accuracy, precision, recall.

***Index Terms* – Breast Cancer, Machine learning Algorithms, Support Vector Machine with GRB technique.**

I. INTRODUCTION

Breast cancer is one of the most dangerous and common reproductive cancers that affect mostly women. Breast tumor is an abnormal growth of tissues in the breast, and it may be felt as a lump or nipple discharge or change of skin texture around the nipple region. Cancers are abnormal cells that divide uncontrollably and are able to invade other tissues [8].

- For every 2 women newly diagnosed with breast cancer, one woman dies of it in India [8].
- Mouth & Lungs and Breast cancer is 50% of total cancer cases.
- The incidence rates in India begin to rise in the early thirties and peak at ages 50-64 years.
- Breast cancer is the most common cancer in women in India and accounts for 28% of all cancers in women.

As a standard, the current diagnostic screening consists of a mammography to identify suspicious regions of the breast, followed by a biopsy of potentially cancerous areas. A breast biopsy is a diagnostic procedure that can determine if the suspicious area is malignant or benign [3]. Although criteria for diagnostic categories of radiologic and pathology are well established, manually detection and grading respectively is a tedious and subjective process and thus suffers from inter-observer and intra-observer variations [2].

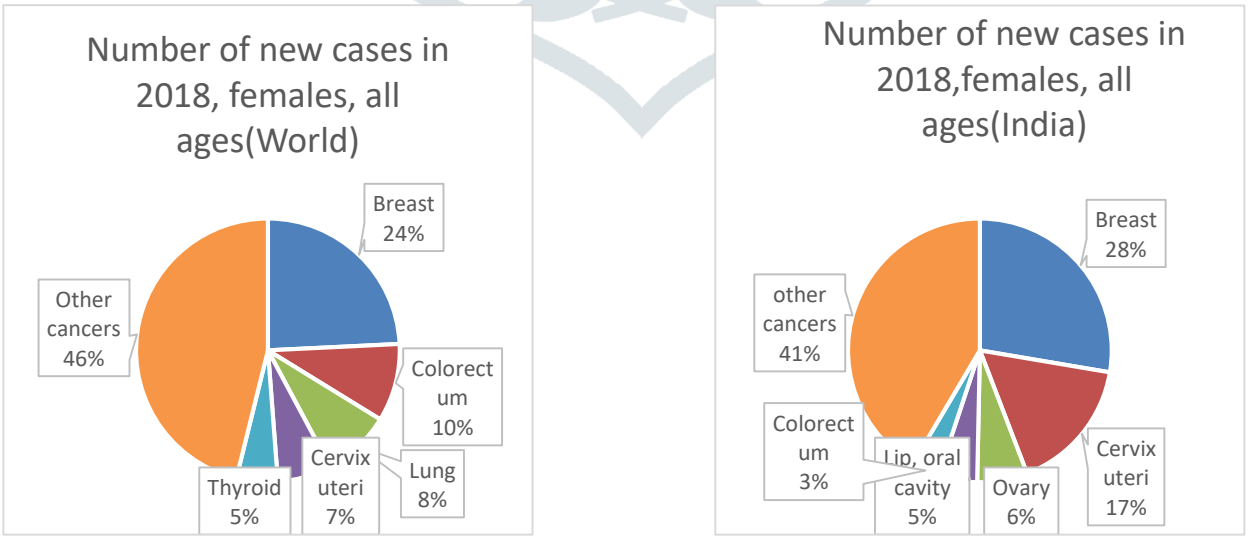


Figure 1: Source: Globocon 2018[8]

In mammography, breast images of a specific type are used to detect early cancer symptoms in women by the radiologist. It has been observed that due to the use of mammography for cancer detection, the death ratio has decreased [3].
Through biopsy, pathologist can determine two types of lesion: benign and malignant. The benign lesion is not cancerous; it is indeed the abnormalities in the epithelial cells, and most of these abnormalities are unable to become a source of breast cancer. The malignant or cancerous cells are those types of cells, which start divisions abnormally and grows irregularly [3].

II. BACKGROUND THEORY

Machine Learning

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves [6].

Supervised ML learning algorithm required the data to be labeled for training purposes. For example, in training a set of medical images to identify a specific breast tumour type, the label would be tumour pathologic results or genomic information [6].

Unsupervised ML clusters the data that have similar characteristics, and the unlabeled data are exposed to the algorithm with the goal of generating labels that will meaningfully organize the data. This is typically done by identifying useful clusters of data based on one or more dimensions. Compared with supervised techniques, unsupervised learning sometimes requires much larger training data sets [6].

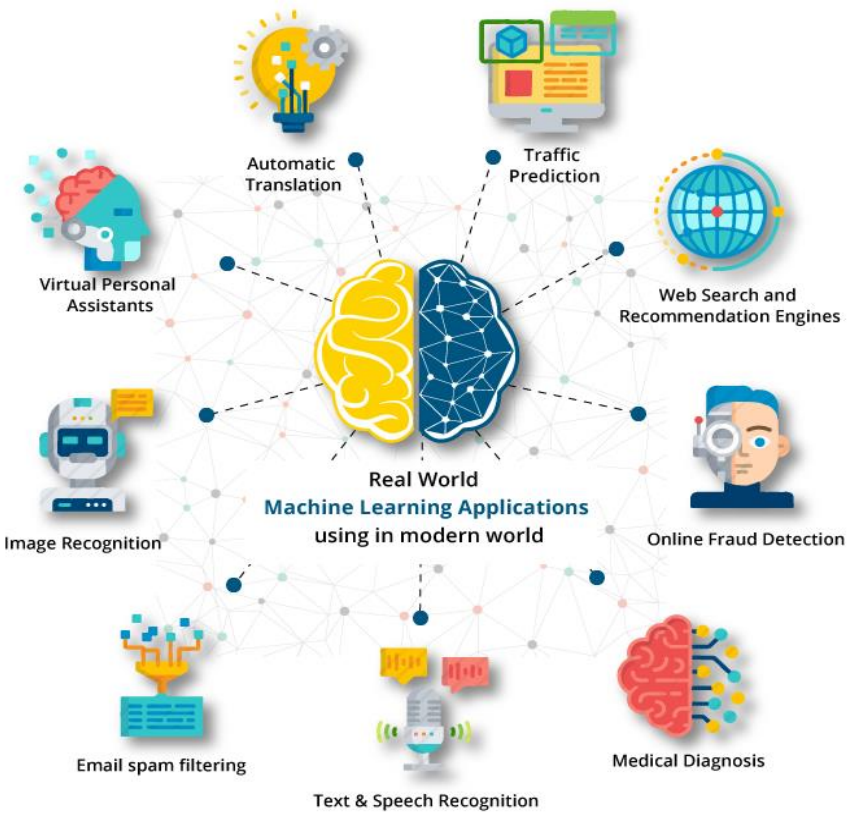


Figure 2: Machine Learning Applications [6]

Machine Learning application in Breast Cancer

The value of machine learning in healthcare is the capability to train & process large datasets that are beyond the scope of human and that processed data analyse and convert the data into clinical sight that helps physicians in planning & giving treatment ultimately leads towards much better outcomes.

With the help of machine learning we can train the data to look at different images (Mammograph or Biopsy report) and with the help of image processing techniques we can identify abnormality & able to find out principal areas that need analysis we can look into it. With the classification techniques, we can finally get our outcome. As machine learning learns from itself & improves after each iteration, we can say that Machine learning improves the Curing cycle, Efficiency & accuracy.

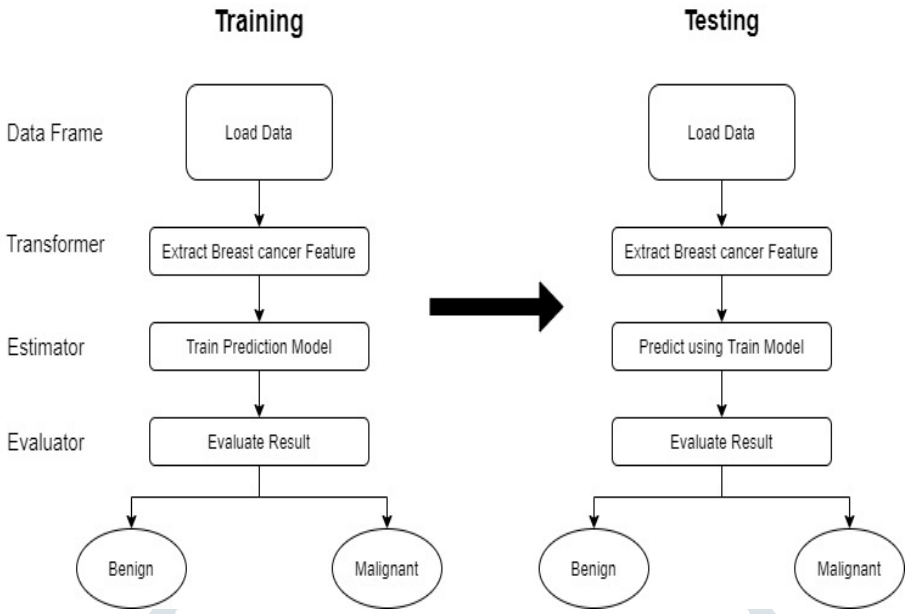


Figure 3: Machine Learning Steps for Predicting Breast Cancer

III. METHODOLOGY

Support Vector Machine

SVMs are a more recent approach of ML methods applied in the field of cancer prediction/prognosis. Initially SVMs map the input vector into a feature space of higher dimensionality and identify the hyperplane that separates the data points into two classes. The marginal distance between the decision hyperplane and the instances that are closest to boundary is maximized [7].

Below figure illustrates how an SVM might work in order to classify tumors among benign and malignant based on their size and patients' age. Obviously, the existence of a decision boundary allows for the detection of any misclassification produced by the method [7].

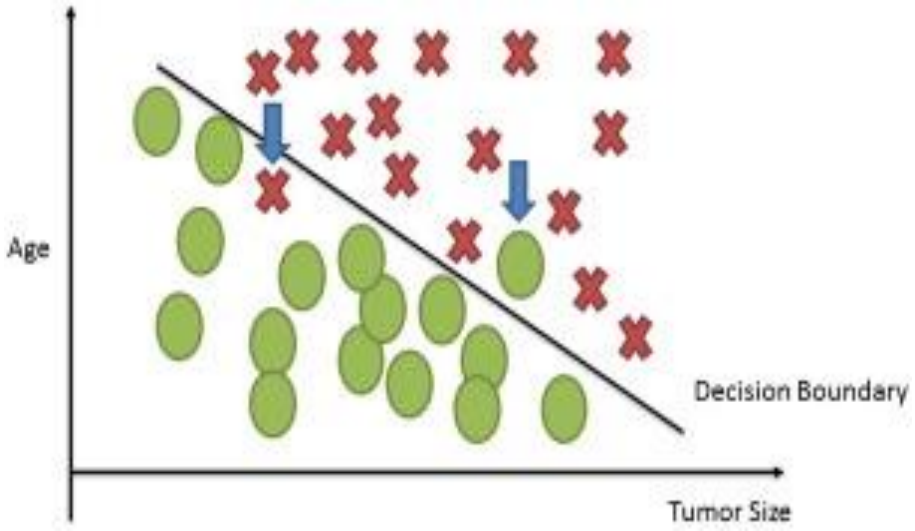


Figure 4: A simplified illustration of a SVM classification of the input data [7]

IV. DATASET

- Input Data collected from UPI Repository ataset:(https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra.
- Dicom images from extracted data. This data is collected from the UKM Medical Centre (UKMMC).

V. PERFORMANCE MEASURES AND IMPLEMENTATION STRATEGIES

The proposed approach mainly focuses on improving the accuracy parameter. The accuracy, Precision & Recall of the Breast Cancer prediction system is calculated based on the following formula.

Accuracy = (TP+TN) / (TP+TN+FP+FN)
Precision = (TP) / (TP+FN)
Recall = (TP) / (TP+FP)

- TP -True positive -reflects the number of instances which are sick and diagnosed accurately.
- FP – False positive -reflects the number of instances which are healthy and diagnosed wrongly as they are sick
- FN – False negative-reflects the number of instances which are sick but the instances are diagnosed wrongly.
- TN - True negative-reflects the number of instances which are healthy and the instances are diagnosed accurately.

Operating System: Windows 10
Programming Language: Python (Version: 3.6)
Tool: PyCharm (runtime version - 11.0.5 + 10-b520.38 amd64)

PyCharm is an integrated development environment (IDE) used for programming, specifically for the Python language . PyCharm provides code analysis, an integrated unit tester , a graphical debugger, built-in database tools, integration with version control systems and supports web development as well as Data Science with Anaconda [11].

Library used for Implementation

- OpenCV
- Pydicom
- Numpy
- PyWavelets (pywt)
- Matplotlib
- Pandas
- Seaborn
- Scikit-Learn (sklearn)

VI. PROPOSED WORK

The Image enhancement with spatial domain method is applied to enhance the image parts. Enhancement at any point of image depends on its grey level part. which can produce the contrast level of an image.

With image segmentation, the region based method is applied, which is controlled by connectivity between pixels. with the help of the prior knowledge of problem, which in turn generate a binary image.

In digital image processing, wavelet transform represent multi resolution images. The wavelet coefficients measure how closely correlated the wavelet is with each section of the signal in the form of matrices. the computation time can be accelerated by using wavelet transforms here 2D wavelet transform method is applied. The training dataset is breast cancer from UCI machine learning repository with 10 features.

We have stored the extracted data from image to extract_detail.csv. (note: the extracted data is in the form of glucose, resistin, BMI), the extracted information is based on the above parameters).We want to use this file in testing with analysis.py on the place of detailed.csv. which itself is a csv generated to test the records.

The training dataset is breast_cancer from UCI machine learning repository with 10 features.

The next process is to first clean the trained dataset and make sure no unwanted data exist. its a pre-processing step for dataset of breast cancer. It is removing the space values which can affect the result.

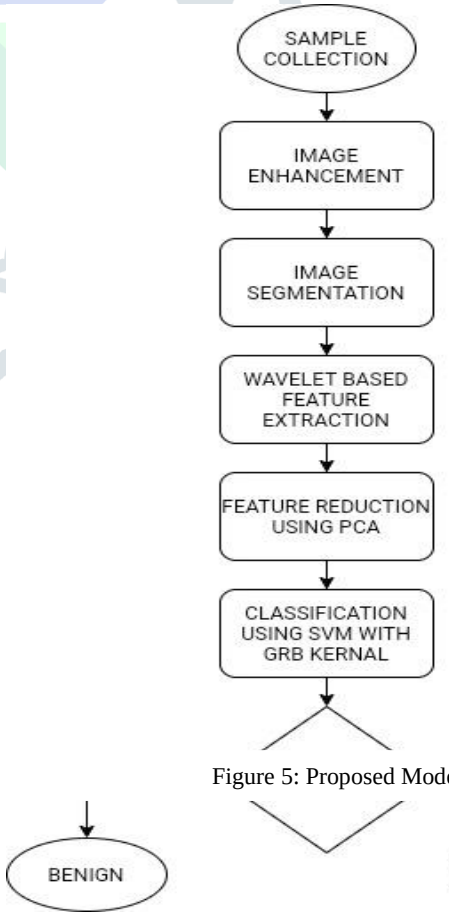


Figure 5: Proposed Model

Than we have extracted the features from the dataset where we have used PCA method of scikit learn. Linear dimensionality reduction using SVC (support vector classifier) of the data to project it to a lower dimensional space. The input data is centered but scaled for each feature with eda.py

The reduced dimensional data is further processed for classification with SVM with GRB technique, which means A classifier such as support vector machine (SVM) is used to make better classification with Gaussian Radial Basis (GRB) kernel , which is used for the classification method proposed and yields maximum accuracy of higher value compared to linear kernel. SVM with GRB itself a technique for better classification in segmented images.

VII. RESULTS AND DISCUSSION

Evaluation of Result

With SVM with GRB technique we got the results for different Parameters.

The proposed approach mainly focuses on improving the accuracy parameter. The accuracy, Precision & Recall of the Breast Cancer prediction system is calculated,

Accuracy: 99.72
Precision: 98.99
Recall : 97.20
Fractions of Errors = 0.053
The time taken by the trained classifier to assign labels = 2.87 seconds.

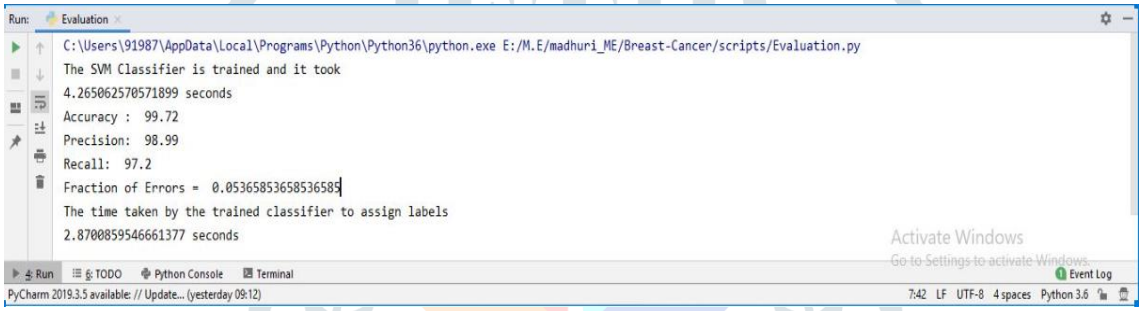
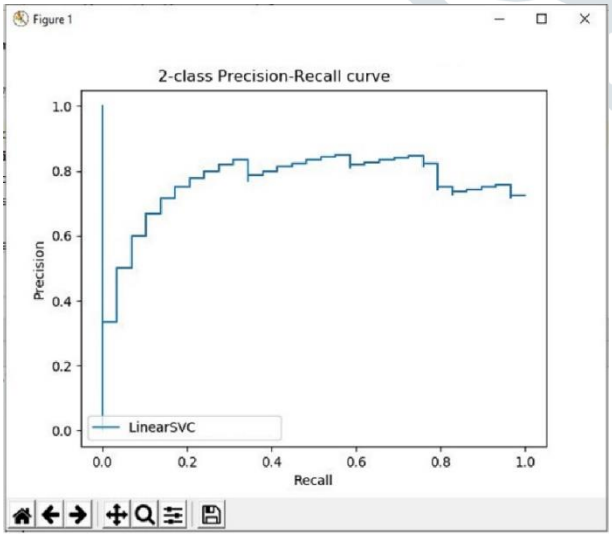


Figure 6: Evaluation of Result

Figure 7: Precision Recall Curve



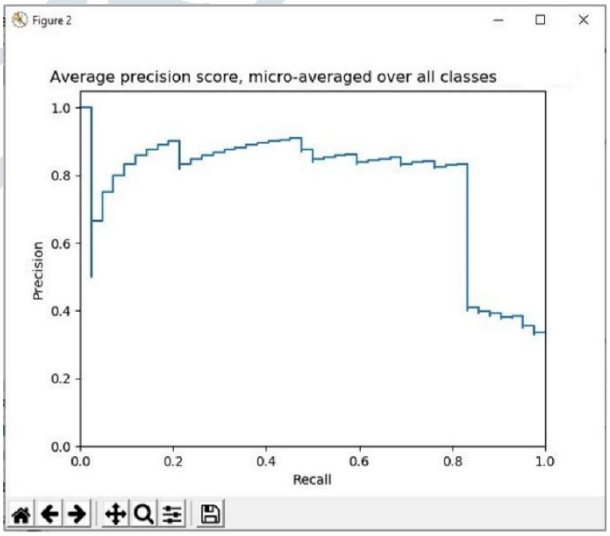


Figure 8: Average precision Score, micro averaged over all classes

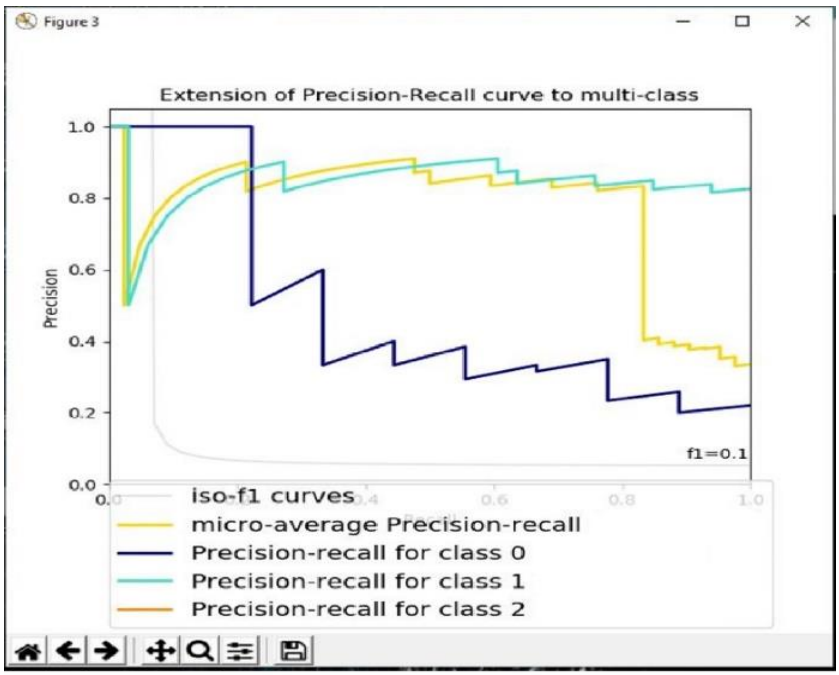


Figure 9: Features identified from images after processing with image enhancement

The finding from the study suggest that breast cancer risk prediction using SVM with GRB technique gives very higher and accurate results than only using SVM classifier [1].Also the performance in terms of precision and recall gives much reliable results in dataset.

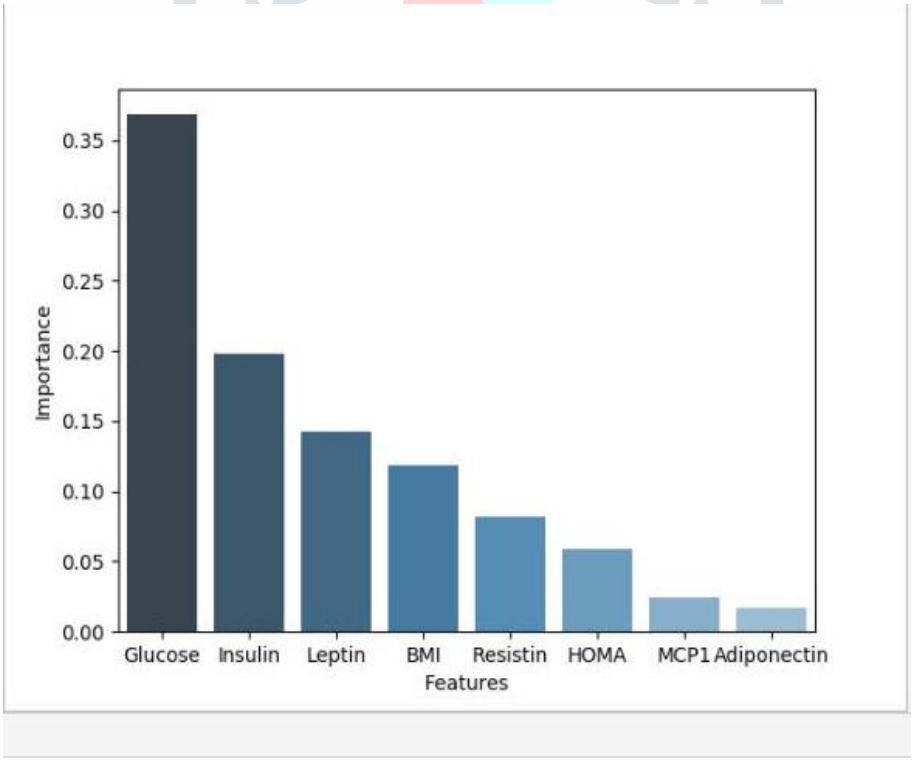


Figure 10: Plotation of Result for different parameters

Here from this result chary (Fig. 10) we can say that, Leptin Induces a Proliferative Response in Breast Cancer Cells but Not in Normal Breast Cells which is found high based on correlation of cells. adiponectin is found with less connected cysts as the high adiponectin levels were associated with an decreased risk for breast cancer. HOMA is identified with size of cysts. The association of resistin with breast cancer risk is evaluated by regression analysis. from image it depends on the softness of tissues which is identified with density of color in cysts. MCP1 is associated with skewness of cysts in image.

The key achievement in this paper is that, this paper classifies these three types of features from patients; using a Support Vector Machine (kernel based) classifier. The images are pre-processed, and its dimensionality is reduced before entering the classifier, and the difference in accuracy produced after pre-processing techniques is compared with Gaussian Radial Basis (GRB) for the classifier which is specially used when the data points are linearly separable. What GRB kernel SVM does is to create non-linear combinations of your features to uplift your samples onto a higher-dimensional feature space where you can use a linear

decision boundary to separate your classes. The experimental results showed that the proposed approach with pre-processed images by using GRB kernel achieves better performance than quadratic and linear kernels in terms of accuracy, precision, and recall.

Accuracy is an important metric to consider but it does not always give the full picture.so precision and recall are become vety important parameters to check.

Precision is the fraction of relevant instances among the retrieved instances (Value Indicates % of people don't have disease but in result shows as positive).

Recall is a measure that tells us how great our model is when all the actual values are positive (Value Indicates % of people have disease but in result shows as negative).

Here, we have shown Computation of precision-recall pairs for different probability thresholds. if the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall in average.

VIII. CONCLUSION

Survey & Study of research papers give me an insight of techniques and algorithms used in the prediction of breast cancer. Support Vector Machine (SVM) on the Breast Cancer datasets are conducted. Main parameters for the comparison were Accuracy, precision, Recall. The proposed system improves the accuracy of Breast Cancer prediction system with demonstrating the 99.72% accuracy, which is definitely more than the existing research methodologies in this domain.

REFERENCES

[1] SARA ALGHUNAIM AND HEYAM H. AL-BAITY , "On the Scalability of Machine-Learning Algorithms for Breast Cancer Prediction in Big Data Context", IEEE Access, 2019, vol. 7, pp. 91535-91546.

[2] Riku Turki et al., "Breast cancer outcome prediction with tumour tissue images and machine learning", Stringer 2019, pp. 41-52.

[3] SanaUllah Khan, Navdeep Islam et al., "A Novel Deep Learning based Framework for the Detection and Classification of Breast Cancer Using Transfer Learning", ELSEVIER 2019, pp. 1-8.

[4] Sara Reis, Patrycja Gazinska et al., "Automated Classification of Breast Cancer Stroma Maturity From Histological Images", IEEE TRANSACTION ON BIOMEDICAL ENGINNERING 2017, vol. 64, pp. 2344-2352.

[5] SungHwan Kim "Weighted K-means support vector machine for cancer prediction", Kim SpringerPlus 2016, vol. 5, pp. 1-11.

[6] Machine learning application, access on 10 August 2019, https://www.quora.com/What-are-some-real-world-examples-of-applications-of-machine-learning-in-the-field

[7] Konstantina Kouroua, Themis P.Exarchos et al., "Machine Learning applications in cancer prognosis and prediction", ELSEVIER 2019, pp. 8-17.

[8] Globocan 2018 report, access on 22 July 2019, https://gco.iarc.fr/today/data/factsheets/populations/900-world-fact-sheets.pdf

[9] Predicting Breast Cancer in Machine learning, access on 10 January 2020, https://mapr.com/blog/predicting-breast-cancer-using-apache-spark-machine-learning-logistic-regression/.

[10] Pycharm Feature, access on 15 January 2020, https://www.jetbrains.com/pycharm/features

[11] PyCharm Installation, access on 5 January 2020, https://www.jetbrains.com/help/pycharm/installation-guide.html.

[12] Pycharm DDownload, access on 4 January 2020, https://www.jetbrains.com/pycharm/download/#section=windows.

[13] Y.M. George, H. H. Zayed, M. I. Roushdy, B. M. Elbagoury,Remote computer-aided breast cancer detection and diagnosis system based on cytological images, IEEE Systems Journal 8 (3) (2014) 949–964.

[14] Bejnordi BE, Veta M, Johannes van Diest P et al (2017) Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 318:2199. https ://doi.org/10.1001/jama.2017.14585

[15] Image Processing, access on 14 August 2019, https://www.researchgate.net/figure/Image-processing-pipeline-developed-for-the-segmentation-of-the-different-compartements_fig6_287792024

[16] Nicoletta Biglia, Elisa Peano et al., "Body mass index(BMI) and breast cancer: impact on tumor histopatologic features, cancer subtypes and recurrence rate in pre and postmenopausal women", ReaserchGate, 2012 pp. 263-267.

[17] Paola Muti,2 Teresa Quattrin et al., "Fasting Glucose Is a Risk Factor For Breast Cancer: A Prospective Study1",Cancer Epidemiology, Biomarkers & Prevention, 2018 ,Vol. 11, pp. 1361–1368

[18] Eva S. Schernhammer, Jeff M. Holly et al., "Circulating Levels of Insulin-like Growth Factors, their Binding Proteins, and Breast Cancer Risk", Cancer Epidemiology, Biomarkers & Prevention, 2005, pp. 699-705.

[19] Ezinne Igwe1, Ahmad Zaid Fattah Azman1 et al., "ASSOCIATION BETWEEN HOMA IR AND CANCER IN A MEDICAL CENTRE IN SELANGOR, MALAYSIA", International Journal of Public Health and Clinical Sciences, 2015, vol. 2, pp. 2289-7577.

[20] Sebastiano Andò, Luca Gelsomino et al., "Obesity, Leptin and Breast Cancer: Epidemiological Evidence and Proposed Mechanisms", Cancers, 2019, pp. 2-27.

[21] Flora Sánchez-Jiménez , Antonio Pérez-Pérez et al., "Obesity and Breast Cancer: Role of Leptin", Frontiers in Oncology, 2019, vol. 9 , pp. 1-12.

[22] Li-Yuan Liu, Meng Wang et al., "The Role of Adiponectin in Breast Cancer: A Meta-Analysis", PLOS ONE, 2013, vol. 8, pp. 1-10.

[23] Jee-Hyun Kang, Byung-Yeon Yu, et al., "Relationship of Serum Adiponectin and Resistin Levels with Breast Cancer Risk", The Korean Academy of Medical Sciences, 2007, vol. 22, pp. 117-21.

[24] Liang-Shan Da1, Ying Zhang et al., "Association between MCP-1 -2518A/G Polymorphism and Cancer Risk: Evidence from 19 Case-Control Studies", PLOS ONE, 2013, vol. 8, pp. 1-7.

[25] M. HLAVNA1, L. KOHUT et al., "Relationship of resistin levels with endometrial cancer risk", Neoplasma, 2011, pp. 124-128.