CLASSIFICATION OF BREAST CANCER INTO BENIGN AND MALIGNANT USING SUPPORT VECTOR MACHINES

K.S.NS. Gopala Krishna[1], B.L.S. Suraj[2], M. Trupthi[3]

[1,2]Student, [3]Assistant Professor, Department of Information Technology,
Chaitanya Bharathi Institute of Technology, Hyderabad (India)

ABSTRACT

Cancer has been characterized as a heterogeneous disease consisting of various subtypes. Breast Cancer is the second leading cause of cancer deaths after lung cancer, the principle cause of death from cancer among women globally [7]. Early detection is the most effective way to reduce breast cancer deaths. Premature diagnosis requires an accurate and reliable procedure to distinguish between benign tumours from malignant tumours of breast. The prognosis and early detection of a cancer type have become an essential requirement in cancer research, as it can help in the consequent clinical management of patients. The importance of classifying cancer patients into benign or malignant groups has led many research teams, from the bioinformatics and biomedical field, to study the application of machine learning (ML) methods and algorithms. Therefore, these techniques have been utilized as an aim to model the progression and treatment of cancerous conditions. In addition, the ability of ML tools to detect key features from complex datasets reveals their importance. In this paper Support Vector Machines (SVMs) technique has been applied for the development of breast cancer predictive models which is effective and accurate decision making. This paper acts as a data science analysis tool which assists the user to distinguish between the benign and malignant tumours of breast. With the help of this tool breast cancer diagnosis can be achieved. Graphs and tabular columns are used to represent the analysed and classified data.

Keywords: Benign, Feature Extraction, Malignant, Supervised Machine Learning, Principal Component Analysis.

I. INTRODUCTION

A. Preliminaries

Breast cancer occurs when a malignant (cancerous) tumour originates in the breast. As breast cancer tumours mature, they may metastasize to other parts of the body. The primary route of metastasis is the lymphatic system which is the body’s elementary system for transporting and producing white blood cells and other cancer-fighting immune system cells throughout the body. Metastasized cancer cells that aren’t destroyed by the white blood cells move through the lymphatic vessels and settle in remote locations of the body, forming tumours and perpetuating the disease process. For diagnosing the different stages of the breast cancer, Ultrasound, MRI, Chest X-ray, Mammography are extensively used [1]. Number of breast cancer disease is calculated to be around 1.2 million women every year according to the statistics of World Health Organization. For the diagnosis and treatment of cancer, precise prediction of tumours is critically significant. The Latest machine learning techniques are increasingly being used by scientists to obtain relevant tumour information from the databases. Among the current techniques, supervised machine learning methods are the most popular in cancer diagnosis. The Breast Cancer datasets for this paper has been obtained from Wisconsin [8]. The dataset contains the samples of malignant and benign tumour cells.

B. Basic concepts used in cancer cell detection

In this paper we have used Principal Component Analysis for Feature Extraction and Support Vector Machines for classification of data into benign and malignant stages. The concepts are discussed as follows:

1) Principal Component Analysis: Principal component analysis is a method of extracting important variables in form of components from a large set of components available in a data set. It extracts low dimensional set of features from a high dimensional data set with a motive to capture as much information as possible. It is always performed on a symmetric correlation or covariance matrix. This means that the matrix should be numeric and have standardized data. Let us consider an example having a dataset with the dimension of NxP. The image below shows the transformation of a 3 dimensional data to 2-dimensional data using PCA. Not to forget, each resultant dimension is a linear combination of p features. In figure 1 PC1 and PC2 are principal components 1 and principal components 2 respectively. Let us say we have a set of predictors as x1, x2, x3……xp. The Principal Components can be written as: $\boldsymbol{Z^{1} = \varphi^{11}x + \varphi^{21}x + \varphi^{31}x + \cdots + \varphi^{p1}x(1)}$ where Z^{1}, is first principal component, φ^{p1} is the loading vector comprising of loadings of the first principal component. The loadings are constrained to a sum of square equals to 1 [6]. This is because large magnitude of loadings may lead to large variance. It also defines the direction of the principal component Z along which data varies the most. It results in a line in p dimensional space which is closest to the n observations. Closeness is measured using average squared Euclidean distance x1, x2....., xp are normalized predictors. Normalized predictors have mean equals to zero and standard deviation equal to one. First principal component is a linear combination of original predictor variables which captures the maximum variance in the data set [6]. It determines the direction of highest variability in the data. Larger the variability captured in first component, larger the information captured by component. No other component can have variability higher than first principal component.

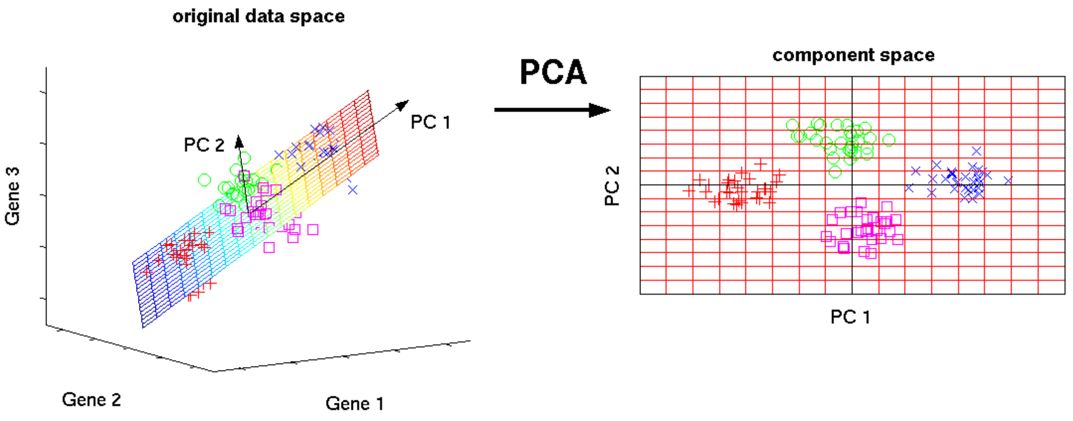


Fig. 1. An example of showing how a PCA is useful in transforming a 3-D graph to a 2-D graph

2) Support Vector Machine: Support vector machines (SVMs) learning algorithm will be used to build the predictive model. SVMs are one of the most popular classification algorithms, and have an elegant way of transforming nonlinear data so that one can use a linear algorithm to fit a linear model to the data [12]. SVMs allow for complex decision boundaries, even if the data has only a few features. They work well on low dimensional and high-dimensional data.

II. PROPOSED METHODOLOGY

The first step is to import external data sets and identify the types of information contained in our data set. Then explore the variables to assess how they relate to the response variable and find the most predictive features of the data and filter it so it will enhance the predictive power of the analytics model. In the final step we construct predictive models to predict the diagnosis of a breast tumour. The Breast Cancer datasets is available in Wisconsin (Diagnostic). The first two columns in the dataset store the unique ID numbers of the samples and the corresponding diagnosis (M=malignant, B=benign), respectively. The columns 3-32 contain 30 real-value features that have been computed from digitized images of the cell nuclei, which can be used to build a model to predict whether a tumour is benign or malignant. The entire procedure can be represented in form of a flowchart as shown below.

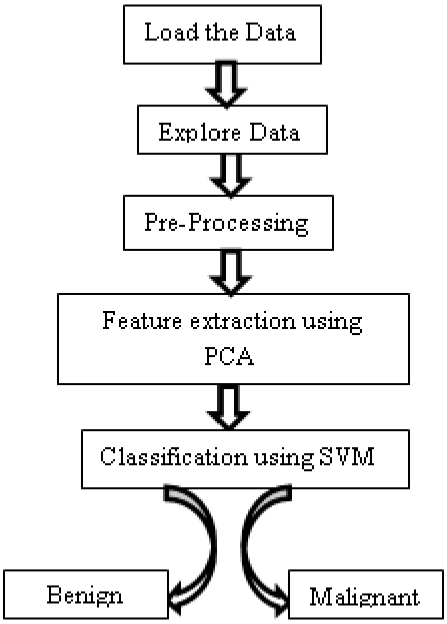


Fig. 2. Flow chart representing the entire procedure

III. IMPLEMENTATION

A. Load the Data

We collected data from Wisconsin and using Python modules like pandas and numpy we imported the data for the purpose of getting to know the data and to get a good grasp of the data and think about how to handle the data in different ways. The dataset consists of 569 rows and 32 columns [8]. Here M refers as malignant tumour and B refers as benign tumour. Unnecessary columns (like ID, unnamed) are dropped.

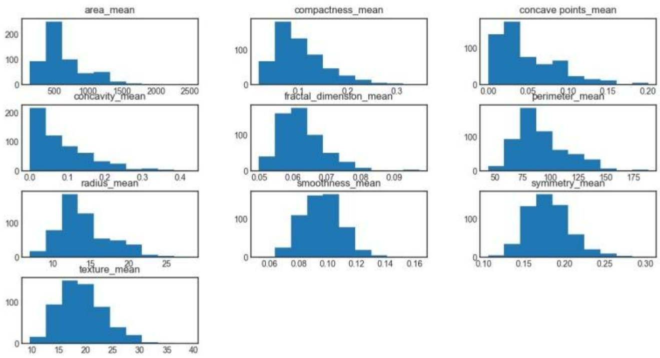


Fig. 3 Histograms obtained by EDA

B. Explore Data

Exploratory data analysis is a very important step that takes place after feature engineering and acquiring data and it should be done before any modelling. This is very important for a data scientist to be able to understand the nature of the data without making assumptions. The results of exploratory data analysis can be extremely useful in grasping the structure of the data, the distribution of the values, and the presence of extreme values and interrelationships within the data set. In this paper, we have plotted histograms. Histograms group data into bins and provide us a count of the number of observations in each bin. From the histogram in the fig.3, we can quickly get a feeling whether an attribute is Gaussian, skewed or even has an exponential distribution. It can also help us see possible outliers. From the fig. 3 we can see that perhaps the attributes like concavity and concavity_point may have an exponential distribution. We can also see that the texture and smooth and symmetry attributes may have a Gaussian or nearly Gaussian distribution. This is very interesting because many machine learning techniques assume a Gaussian uni-variate data distribution on the input variables.

C. Pre-Processing

Data pre-processing is an important step for any data analysis problem. It is often appropriate to prepare your data in such way to best expose the structure of the problem to the machine learning algorithms that you intend to use. Pre-processing involves a number of activities like Data cleaning to fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies; Data integration for using multiple databases, data cubes, or files; Data transformation which is used for normalization and aggregation; Data reduction for reducing the volume but producing the same or similar analytical results; Data discretization which includes part of data reduction where replacing numerical attributes with nominal ones takes place. The simplest method to evaluate the performance of a machine learning algorithm is to use different training and testing datasets. In this paper the available data is split into a training set and a testing set. (70% training, 30% test). The first set is used to Train the algorithm and predictions are made on the second part. Then we evaluate the predictions against the expected results.

Feature Standardization: Standardization is a useful technique to transform attributes with a Gaussian distribution and differing means and standard deviations to a standard Gaussian distribution with a mean of 0 and a standard deviation of 1. As seen in exploratory data analysis mentioned above in this paper, the raw data has differing distributions which may have an impact on the most ML algorithms. Most machine learning and optimization algorithms behave much better if features are on the same scale. Then evaluate the same algorithms with a standardized copy of the dataset. Sklearn is used to scale and transform the data such that each attribute has a mean value of zero and a standard deviation of one. After feature standardization we use Principal component analysis for dimensionality reduction i.e. to transform attributes with a Gaussian distribution and differing means and standard deviations to a standard Gaussian distribution with a mean of 0 and a standard deviation of 1 [6]. After applying PCA we plotted a graph as shown below.

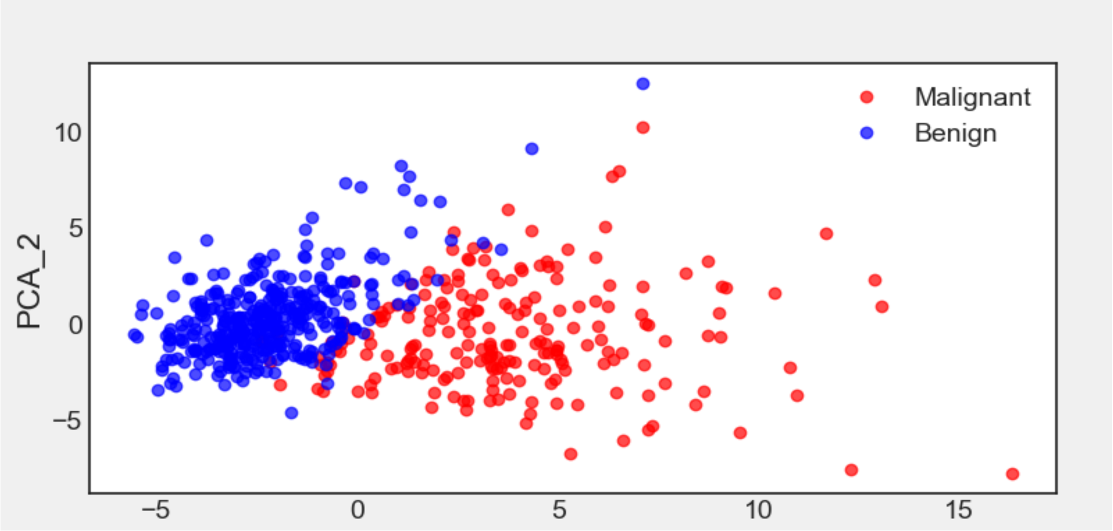
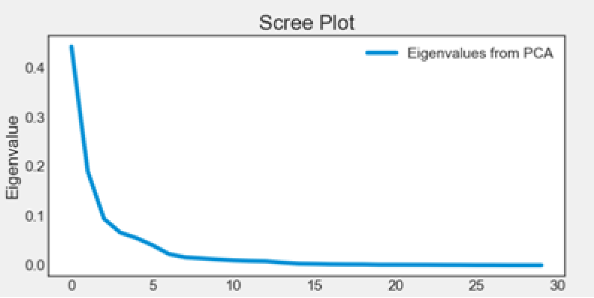


Fig. 4 Dimensionality reduction using PCA and plotting a 2-D graph.

Now, what we got after applying the linear PCA transformation is a lower dimensional subspace (from 3D to 2D in this case), where the samples are most spread along the new feature axes. Choosing the number of principal components can be a challenging task. To make this easier, we plot a scree plot as shown in the figure given below. From fig 5 and 6 we could see that the graph is an exponential graph until the value 6. Later on it has become stationary [6]. So, we can conclude that we require atleast 6 Principal components. For our convenience, we took 10.



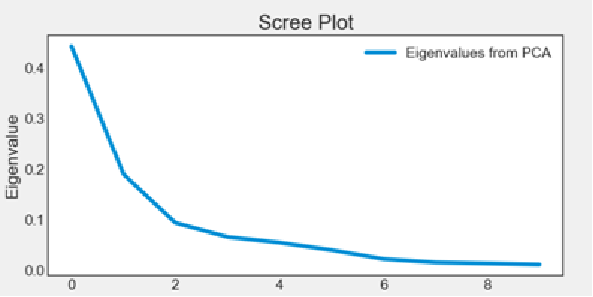


Fig. 5 Scree Plot for 30 components **Fig.6 Scree Plot for 6 components**

D. Classification of Data using SVM

As said earlier, we are using SVM classifier for our classification. They work well on low-dimensional and high-dimensional data (i.e., few and many features), but don't scale very well with the number of samples. SVMs require careful pre-processing of the data and tuning of the parameters.

Important Parameters: The important parameters in kernel SVMs are:

- The regularization parameter C-Here, C is a regularization parameter that controls the trade-off between the achieving a low training error and a low testing error that is the ability to generalize the classifier to unseen data.
- The choice of the kernel, (linear, radial basis function (RBF) or polynomial)
- Kernel-specific parameters.

Gamma and C both control the complexity of the model, with large values in either resulting in a more complex model. Therefore, good settings for the two parameters are usually strongly correlated, and C and gamma should be adjusted together. As discussed earlier, we are splitting data into 70% training and 30% testing, it is crucial to avoid over fitting(i.e. when a model is excessively complex, like having too many parameters relative to the number of observations, it leads to poor predictive performance.). To overcome this we use Cross Validation. Instead of having a single train/test split, we specify so-called folds so that the data is divided into similarly-sized folds. Training occurs by taking all folds except one referred to as the holdout sample. On the completion of the training, you test the performance of your fitted model using the holdout sample. The holdout sample is then thrown back with the rest of the other folds, and a different fold is pulled out as the new holdout sample. Training is repeated again with the remaining folds and we measure performance using the holdout sample. This process is repeated until each fold has had a chance to be a test or holdout sample. The expected performance of the classifier, called cross-validation error, is then simply an average of error rates computed on each holdout sample. After the classification is done, we calculate the accuracy of the classifier as:

- Accuracy = predicted values/actual values.
- Precision is calculated for individual tumour i.e. when it is 1, how does it predict correctly?
- Precision=Predicted 1 or 0/Actual 1 or 0.
- F1 score (also called F-score or F-measure) is the measure of test's accuracy. It considers the precision p and the recall r of the test to compute the score. The F1 score is average of p and r. The best value for an F1-score is 1 and worst score is 0.

A table as show below shows the precision, recall, f1-score accuracy.

The accuracy that we achieved using SVM classifier is 95%.

TABLE-1

	Precision	Recall	F1-Score	Support
0	0.94	0.99	0.96	107
1	0.98	0.89	0.93	64
Avg/Total	0.95	0.95	0.95	171

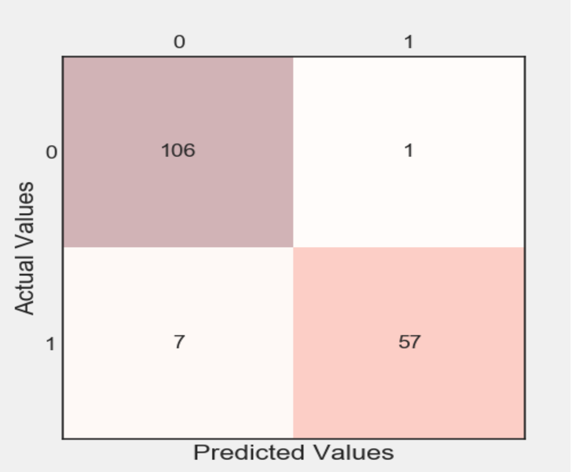


Fig. 7 Prediction using SVM classifier

E. Optimizing the Classifier

Machine learning models are parameterized so that their behaviour can be tuned for a given problem. Models can have many parameters and finding the best combination of parameters can be treated as a search problem. We can tune two key parameters of the SVM algorithm:

- The value of C (how much to relax the margin)
- The type of kernel.

The default for SVM (the SVC class) is to use the Radial Basis Function (RBF) kernel with a C value set to 1.0. Like with KNN, we will perform a grid search using 10-fold cross validation with a standardized copy of the training dataset. We will try a number of simpler kernel types and C values with less bias and more bias (less than and more than 1.0 respectively).

Python scikit-learn provides two simple methods for algorithm parameter tuning:

- Grid Search Parameter Tuning.
- Random Search Parameter Tuning.

Optimizing the classifier in this paper helped us to achieve a 100% accuracy in calculating precision rate for Benign tumour and increased the accuracy of precision rate of Malignant tumour as shown in the figure below.

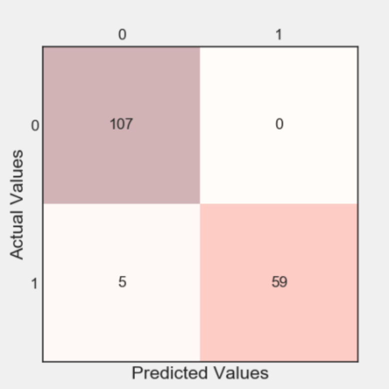


Fig.8 Output using Optimized Classifier.

V. CONCLUSION

Compared to other cancers, breast cancer is the major cause of deaths in women. So, it's early detection is required to reduce life loss. In this paper we have applied techniques namely, data cleaning, exploratory data analysis, pre-processing the data, feature extraction using Principal Component Analysis, Classification of data into either benign or malignant and finally optimizing the classifier. Our study reveals that using SVM classifier gives an accuracy of 95% and optimizing it increases the precision of benign accuracy from 99% to 100% and malignant accuracy from 89% to 92%. This work can be further enhanced by identification of particular stage of breast cancer or if it is known to be malignant cancer, the survival prediction can be made in the near future.

REFERENCES

[1] Subrata Kumar Mandal, IJECS Volume 6 Issue 2 Feb., 2017 Page No.20388-20391.

[2] Mohd.F, Thomas, M., 2007.Comparison of different classification techniques using WEKA for Breast cancer.IFMBE proceedings 15:520- 523.

[3] Kim W,Kim KS, Lee JE, Noh DY, Kim SW, Jung YS, Park MY, Park RW, "Development of novel breast cancer recurrence prediction model using support vector machine"Journal of breast cancer 15.2(2012): 230-238.

[4] K.Balachandran &R.Anitha,"Ensemble based optimal classification model for pre-diagnosis of lung cancer", 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), IEEE,(2013)

[5] Jiawei Han, Jian Pei, Micheline Kamber "Data Mining Concepts and Techniques", Third Edition, Elsevier Inc, 2012, ISBN:978-0-12-381479- 1.

[6] https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/

[7] http://www.nationalbreastcancer.org/breast-cancer-facts

[8] https://archive.ics.uci.edu/ml/machine-learning-databases/

[9] Priyanka Jain & Santosh Kr. Vishwakarma (2016). Collaborative Analysis of Cancer Patient Data using Rapid Miner. International Journal of Computer Applications, 145, 8-13.

[10] M. Kumar, S. S. Tomar and B.Gaur, "Mining based Optimization for Breast Cancer Analysis: A Review",International Journal of Computer Applications, vol. 19, no. 13,(2015).

[11] LeenaVig, "Comparative Analysis of Different Classifiers for the Wisconsin Breast Cancer Dataset", Open Access Library Journal,Volume 1 | e660,2014.

[12] http://www.jmlr.org/papers/volume2/tong01a/tong01a.pdf