International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 4, Issue 5, pp 10-15, May-2020

Predicting Breast Cancer using Support Vector Machine Learning Algorithm

[1]Adhista Chapagain, [2]Ashutosh Ghimire, [3]Amira Joshi, [4]Anku Jaiswal

[1]Associate Technical Writer, LogPoint, Kathmandu, Nepal
[2]Software Engineer, Bent Ray Technologies (Pvt) Ltd., Kathmandu, Nepal
[3]Associate Quality Engineer, F1soft International Pvt. Ltd., Kathmandu, Nepal
[4]Assistant Professor, Department of Electronics and Computer Engineering, Pulchowk Campus, Institute of Engineering, Kathmandu, Nepal

***Abstract* - Breast cancer is one of the main reasons for death in the world, mostly for women over the last decade. By data mining techniques, the number of tests that are conventionally required, such as MRI, mammogram, ultrasound, and biopsy, can be reduced. Here, a method is proposed that focuses on detecting the presence of risk of breast cancer as 1(Malignant), i.e., present or 0(Benign), i.e., absent. The proposed method uses dataset available in machine learning repository maintained by the University of California, Irvine. The dataset consists of the unique ID numbers of the samples with corresponding diagnosis (malignant/benign), and real-value features (parameters) that are computed from digital images of the cell nuclei of the breast. Support vector machines (SVMs) learning algorithm is used to build a predictive model to identify whether a tumor is malignant or benign. It resulted in an accuracy score of 95.6%.**

Keywords: Tumor, Malignant, Benign, Support Vector Machines (SVMs).

I. INTRODUCTION

Breast cancer is one of the causes of death in many countries. At present, the techniques used for the correct detection and finding the presence of tumor like; Computed Tomography (CT), Magnetic Resonance Imaging (MRI), ultrasound examination, and mammography have not been very efficient having physical complication on the body. [1] For an accurate and reliable diagnosis, a paradigm shift from the traditional treatment-based methods to more sophisticated event-driven precision treatment is necessary.

The proposed method aims to predict whether breast cell tissue is malignant or benign in the human undergoing the test. The intake for the method is; radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetrical measures, and fractal dimension of the breast. The analysis is divided into four sections; identifying the problem and data sources, exploratory data analysis, pre-processing the data, and predicting the case. Our method can do the following:

1. Take the parameter from a candidate as input.
2. Identify the most predictive features and filter such that it enhances the predictive power of the analytical model.
3. Construct a predictive model to predict the breast tumor.

The end-goal was to achieve accuracy with a low error rate while analyzing data. After building the model, the effectiveness was validated from criteria's; accuracy, precision, sensitivity, specificity, and correctly and incorrectly classified instances. Thus, the maximum accuracy of 95.6% was achieved.

II. LITERATURE REVIEW

Chaurasia et al. [2] investigated the performance of BFTree (Best First Tree), IBK (K-nearest neighbor classifier), and SMO (Sequential Minimal Optimization) classification techniques on breast cancer data. The authors experimented with the Weka data mining tool and took three evaluation criteria such as time, correctly classified instances, and accuracy for assessing the superiority of each algorithm. The authors stated that the performance of SMO algorithms was better than the other two algorithms in terms of accuracy and low error rate.

Shajahaan et al. [3] explored the applicability of decision trees for breast cancer prediction. They analyzed the performance of conventional supervised learning algorithms such as; CART, ID3, C4.5, and Naïve Bayes. The experiment was conducted through the Weka tool. The authors concluded that the random tree served as the best classification algorithm for breast cancer with higher accuracy in prediction.

Shrivastava et al. [4] used classification techniques for classifying the benign or malignant instances of breast cancer datasets. The authors created a decision tree classifier model for the classification. They stated that most of the breast cancer analyses have done only through a neural network and decision tree approaches. Hence, these authors implemented a decision tree model using if-then rules for enhancing the performance of decision trees.

Venkatesan et al. [5] analyzed the breast cancer data using four classification algorithms, namely Classification and Regression Trees (CART), Alternating Decision Tree (AD Tree), and Best First Tree (BF Tree). The authors experimented with the Weka tool. The classifier was applied for two test beds cross-validation, which used ten folds with nine folds used for training each classifier, and one-fold is used for testing, and the percentage split uses 2/3 of the dataset for training and 1/3 of the dataset for testing. The authors claimed that the decision trees have a standard construct and easy to understand from which the rules can be extracted. The authors have also stated that the classifier has the highest accuracy, with 99%.

Majali et al. [6] presented a diagnostic system using classification and association approaches. The authors used Frequent Pattern (FP) in association to the rule mining for separating the patterns Prediction of Breast Cancer through Classification Algorithms: A Survey 361 that are frequently found with benign and malignant instances. The authors used a decision tree algorithm for predicting the possibility of cancer concerning age. The authors have implemented the Fp growth algorithm for generating frequent item set without candidate generation, which improves the performance of the algorithm. The authors have claimed that their algorithm can achieve 94% of prediction accuracy.

III. MODEL OF THE PROPOSED METHOD

The below flowchart illustrates the chronological order for the proposed method. The detailed explanation for every step is given below:

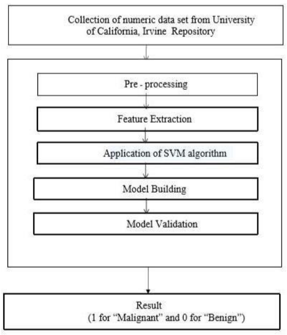
Collection of numeric data set from University of California, Irvine Repository
Pre - processing
Feature Extraction
Application of SVM algorithm
Model Building
Model Validation
Result
(1 for "Malignant" and 0 for "Benign")

Figure 1: Chronological model for the proposed method

3.1 Collection of data

The Breast Cancer dataset is available and maintained by the repository at the University of California, Irvine. The dataset contains a total 569 instances of malignant and benign tumor cells. The real-valued features were calculated for each nucleus of the cell. It includes; radius, texture (Standard deviation of related grey-scale values), perimeter, areas, smoothness (A local variation of the radius length), compactness ((Square of perimeter/area)-1), concavity (Severeness of the concave section of the contour), concave points (The number of concave parts of the contour), symmetry, and fractal dimension (Approximation of the coastline-1).

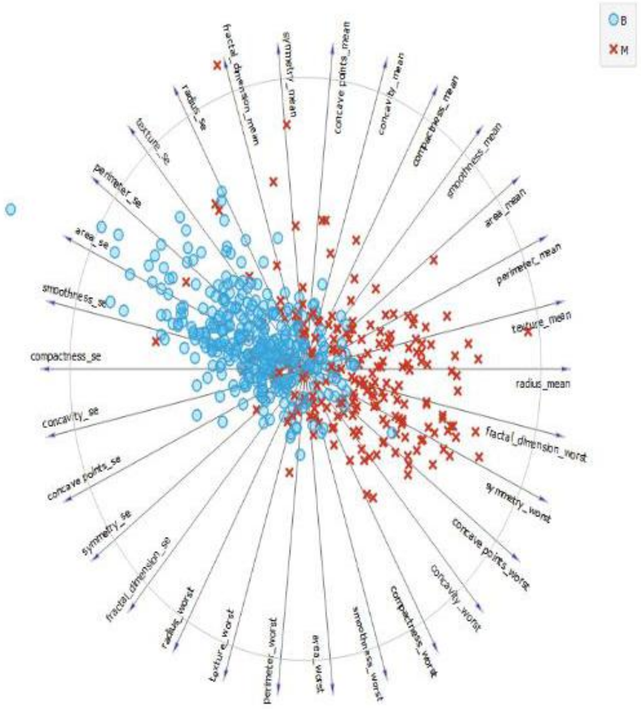

Figure 2: Data Circle

Fig-2 illustrates the 360-degree view of real-valued instances. The blue and red dots plot the distribution of the overall benign and malignant cases of the data set used respectively.

3.2 Pre-processing

Out of 33 columns, the first and second stored unique ID numbers and its diagnosis (M=malignant, B=benign), respectively. The columns 3-33 stored real-valued instances that were calculated from the digital images. These real-valued instances were used to predict whether a tumor was malignant or benign. The column 3-33 consists of the mean, standard deviation, and worst values (mean of three greatest values) of real-valued instances. For example, field 3 is texture_mean, field 13 is texture_se, and field 23 is texture_worst. During preprocessing, the data of two columns (NaN and Patient ID) were removed. It reduced the dimension of the column from 33 to 31.

3.3 Feature Extraction

Principal Component Analysis (PCA) is convenient when dealing with three or higher dimensional data, as it aids in finding the direction with the maximum variance of the data. PCA rotates the initial data space into the new coordinate system point in the directions with maximum variance of the data. The axes or new variables are principal components (PCs), ordered by variance. The first component, PC1 indicates the direction of the maximum variance of the data. It transforms a correlated variable to uncorrelated variables (PCs). The first principal component can be written as:

$$Z^{I} = \emptyset^{11}X^{1} + \emptyset^{21}X^{2} + \ldots. + \emptyset^{P^{I}}X^{P}$$

Where, Z^{I} is the first principal component and $\emptyset^{P^{I}}X^{P}$ is the loading vector that comprises loadings $\emptyset_1, \emptyset_2, \ldots..$ and $X^1, X^2, \ldots, X^P$ set of predictors. The second principal component is the linear combination of original predictor. It captures the remaining variance in the dataset and is not correlated with Z^{I} (The correlation between first and second component must be zero). The second principal component can be written as:

$$Z^{2} = \emptyset^{12}X^{1} + \emptyset^{22}X^{2} + \ldots. + \emptyset^{P^{2}}X^{P}$$

Low variance is assumed to represent undesired background noise. Hence, PCA extracts low dimensional dataset from high dimensional dataset. It helps in capturing as much information as possible. Thus, while performing the feature extraction, the components get filtered out using PCA and the actual extraction gets done by choosing all the components whose Eigen value is greater than 1.

The Support Vector Machines (SVMs) is a type of a supervised machine learning algorithm as the input and output pattern to the system was fed. SVM performs classification task by constructing a hyperplane (or a separator or decision-surface) in high-dimensional feature space separating different instances.

Following this, a prediction is made for any new instance based on its characteristics and mostly with the group it resonates. Point near to the hyperplane is called support vector and distance between the vectors from the hyperplane is called margin. The equation for hyperplane of "m" dimension is:

$$y = w_0 + \sum_{i=1}^{m} w_i x_i$$

Where, w_i are vectors $(w_0, w_1, \ldots\ldots, w_m)$ and x_i are variables.

The behavior of a support vector machine can be changed by using a different kernel function. Kernel is a method of calculating dot product between two vectors in high-dimensional feature space transformed by Ø. RBF and Polynomial kernel functions were used.

The formula for RBF kernel function is,

$$(Xi, Xj) = exp(-\gamma|Xi-Xj|\text{^}2)$$

The formula for Polynomial kernel function can be written as,

$$(Xi, Xj) = (\gamma Xi. Xj+C)\ \text{^}4$$

Where,

$$K(Xi,Xj)=\emptyset(Xi)*\emptyset(Xj)$$

γ: Gamma is used for RBF kernel only. As the value of ‘γ’ increases the model gets overfits and vice-versa the value of ‘γ’ decreases. Likewise, as the value of ‘C’ increases the model gets overfits and vice-versa if the value of ‘C’ decreases.

3.4 Model Building

SVM is used to build the predictive model.

3.5 Model Validation

Confusion matrix and ROC curve.

IV. RESULT AND ANALYSIS

4.1 Result from data visualizations

Fig-3 illustrates the malignant(M) and benign(B) cases in the dataset. The total number of benign(B) cases are 357 and malignant(M) cases are 212.

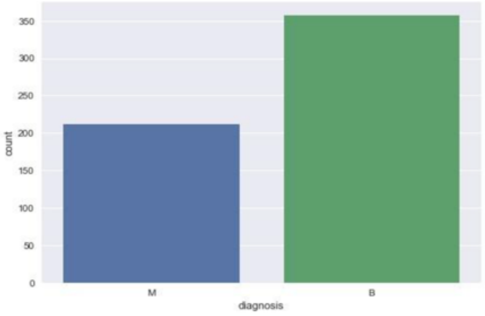

Figure 3: Total number of malignant and benign cases

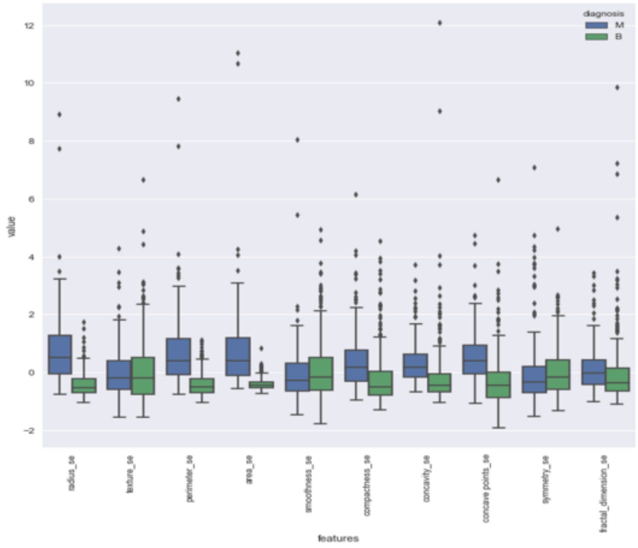

Figure 4: Box plot for features with suffix_se

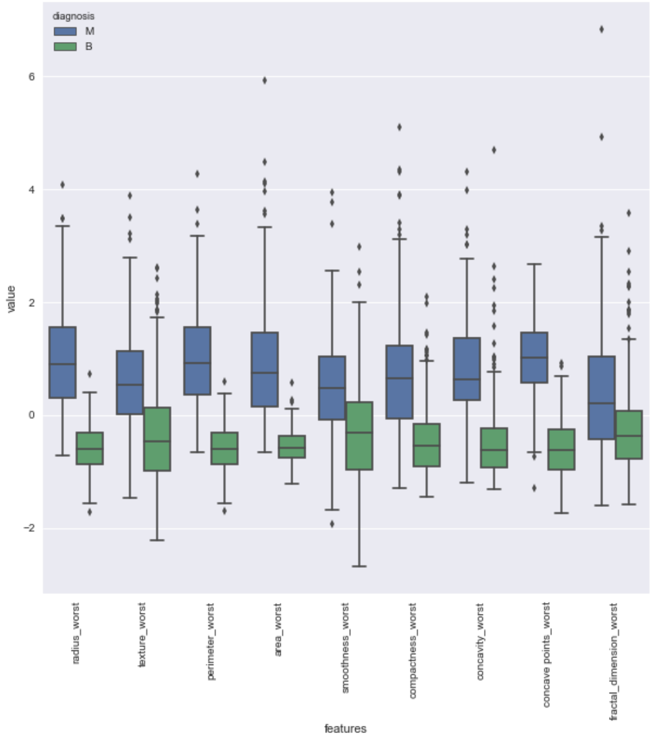

Figure 5: Box plot for features with suffix_worst

Fig-4 and Fig-5 illustrate the statistical data on a plot in which a rectangle is drawn to represent the second and third quartiles.

With a horizontal line inside the rectangle indicates the median value. The lower and upper quartiles are shown on the either side of the rectangle. Shorter the distances more data are bunched together, and greater the distance data spread out.

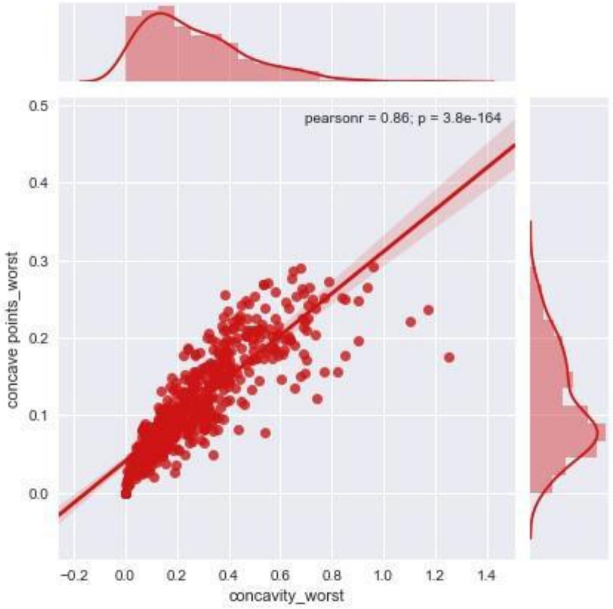

Figure 6: Correlation with Pearson value 0.8

Fig-6 represents the correlation plot where concave points_wave and concativity_worst are positively correlated.

4.2 Result from feature extraction

TABLE 1
Total Variance

Components	Eigen Value		
	Eigen Value	Percentage of Variance	Cumulative Percentage
1	13.28	44.27	44.27
2	5.89	18.97	63.24
3	2.31	9.39	72.63
4	1.88	6.60	79.23
5	1.64	5.49	84.73
6	1.20	4.02	88.75

Table-1 illustrates the six principal components extracted; whose eigenvalue is greater than 1. It represents the dimensionality reduction.

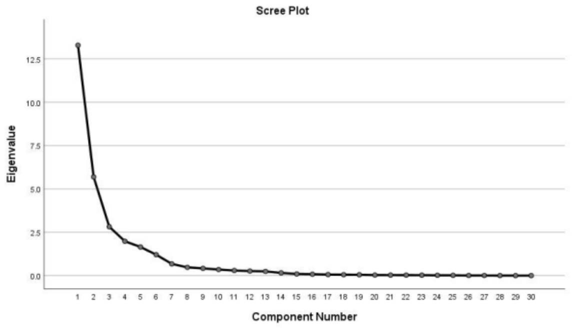

Figure 7: Scree Plot

Scree Plot is the plot for the eigenvalue. In the Scree Plot, there is a significant drop after component 1 and a minimal drop after component 6. Hence, six components out of all 30 components were chosen.

TABLE 2
Communalities Plot

Features	Initials	Extractions
Texture_mean	1.0	.978
Radius_mean	1.0	.972
Area_mean	1.0	.972
Perimeter_mean	1.0	.961
Concavepoints_mean	1.0	.958
Concavity_mean	1.0	.954

The communalities plot illustrates the six features extracted after PCA. It defines the percentage a component is accountable for the feature extraction.

TABLE 3
Confusion Matrix

	PREDICTED VALUES		
ACTUAL VALUES	**Positive**	**Negative**	**Total**
Positive	342	15	357
Negative	10	202	212
Total	352	217	569

Table-3 illustrates the total of 569 predictions made by the classifier. Out of those 569 predictions, the classifier predicted 352 true positive cases, 202 true negative cases, 10 false positive cases, and 15 false negative cases. From the given table accuracy, sensitivity and specificity are calculated as,

Accuracy: 95.67%
Sensitivity: 95.79%
Specificity: 95.28%

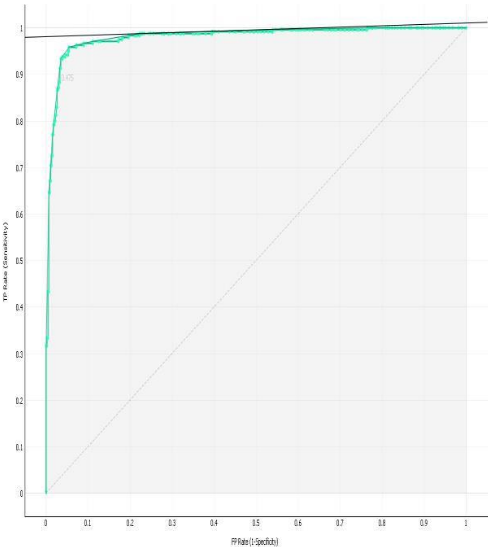

Figure 8: ROC Curve for SVM Model

Fig-8 illustrates the plot between the true positive rate (tpr) versus false positive rate(fpr). For all the points above the diagonal tpr>fpr. Hence, on holding the value of fpr constant, the more vertically above the diagonal, a graph is positioned, the better a classification model performs.

4.3 Result from parameter tuning

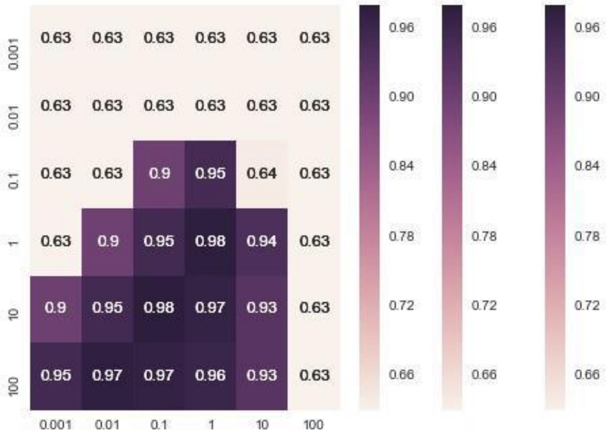

Figure 9: Grid parameter search tuning to identify parameter dependency (X-axis defines Gamma and Y-axis defines C)

Grid search parameter tuning has been used to find the best parameters for the model. List of parameters defined for the model were C [0.001, 0.01, 0.1, 1,10,100] and Gamma [0.001,0.01,0.1,1,10,10,100]. The best parameter for the model is C [10] and Gamma [0.1]. Hence, using these values the decision boundaries were plotted.

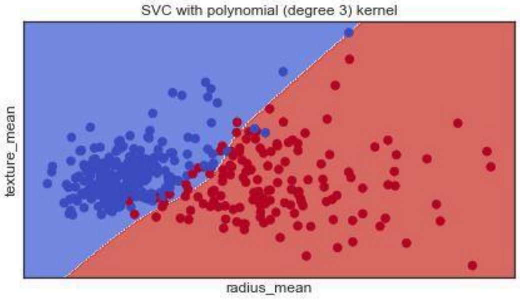

Figure 10: Decision boundary produced by Polynomial (degree 3) kernel

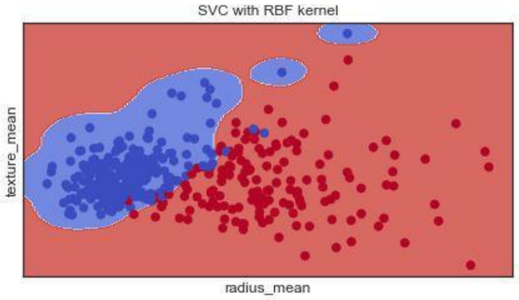

Figure 11: Decision boundary produced by RBF kernel

The decision boundaries produced by the Polynomial and RBF kernel. Here, RBF kernel gives the best classification of malignant data over benign.

V. CONCLUSION

In conclusion, the implementation of SVM can produce almost enough accuracy to be termed as a medically acceptable level of diagnostic accuracy for the dataset used. However, the dataset is not highly normalized resulted in an over-fitting problem due to the number of prominent outliers as seen while tuning the parameters. This limitation partly affected the accuracy of the model.

REFERENCES

[1] Kornack, D. R., & Rakic, P. (2001, December 7). Cell Proliferation Without Neurogenesis in Adult Primate Neocortex. Retrieved June 2018, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.407.3043&rep=rep1&type=pdf

[2] Chaurasia, V., & Pal, S. (2014, January 1). A Novel Approach for Breast Cancer Detection using Data MiningTechniques. RetrievedJune6, 2018, from https://www.researchgate.net/publication/259979477_A_Novel_Approach_for_Breast_Cancer_Detection_using_Data_Mining_Techniques

[3] Selvaraj, S., Shajahaan, S., & Chitra, M. (2013, January). Retrieved June 2018, from https://www.researchgate.net/publication/322635659_Application_of_Data_Mining_techniques_to_model_breast_cancer_data

[4] Shrivastava, S. S., Sant, A., & Aharwal, R. P. (2013). Retrieved July 2018, from https://www.semanticscholar.org/paper/An-Overview-on-Data-Mining-Approach-on-Breast-data-Shrivastava Sant/8d4ae4b9325cd9dfd42a9c70db947ff884c28026

[5] Venkatesan, E., & Velmurugan, T. (2015, November). Performance Analysis of Decision Tree Algorithms for Breast Cancer Classification. Retrieved from https://www.researchgate.net/publication/293191667_Performance_Analysis_of_Decision_Tree_Algorithms_for_Breast_Cancer_Classification

[6] Jaimini, M., Niranjan, R., Phatak, V., & Tadakhe, O. (2015, March). Data Mining Techniques For Diagnosis And Prognosis Of Cancer. Retrieved July 2018, from https://www.researchgate.net/publication/281700941_Data_Mining_Techniques_For_Diagnosis_And_Prognosis_Of_Cancer

Citation of this Article:

Adhista Chapagain, Ashutosh Ghimire, Amira Joshi, Anku Jaiswal, “Predicting Breast Cancer using Support Vector Machine Learning Algorithm” Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 4, Issue 5, pp 10-15, May 2020.

****