Improving Cancer Prediction and Diagnosis Using PCA-based Machine Learning Algorithms and Classifiers.

Abinaya K[1]

[1]*School of Bio Sciences and Technology (SBST), Vellore Institute of Technology (VIT), Vellore-632014, Tamil Nadu, India*

---***---

Abstract - In this project, we implemented machine learning methods to build a Model with higher testing accuracy in predicting breast cancer. For predicting breast cancer, we used a dataset from the Wisconsin Repository. The main idea of this paper is the application of PCA, min-max scalar, and different hyperparameter tuning to ML algorithms for developing Models with high testing accuracy. We used Gaussian naïve Bayes (NB), Logistic Regression (LR), Support Vector Machine (Linear classifier), Support Vector Machine (RBF classifier), Decision Tree (DT), Random Forest (RF), and K Nearest Neighbor (KNN) algorithms to develop Models. Later various performance metrics are performed to analyze each Model. Our Models SVM (linear) and SVM (RBF) obtained a very high testing accuracy of about 99.30% is outperforming other ML Models.

***Key Words*:** Algorithm; Breast Cancer; Machine Learning; SVM; Tuning; PCA; Logistic Regression; KNN; Random Forest; Decision Tree.

1. INTRODUCTION

Breast cancer occurs when cells in the breast grow and divide uncontrollably, creating a lump of tissue called a tumor. We can detect breast cancer by differentiating malignant and benign tumor cells. One in four women is diagnosed with breast cancer globally. The International Agency for Research on Cancer produces the GLOBOCAN 2020 estimation for cancer mortality and incidence. According to the report, women breast cancer rate has increased lately compared with other cancers, with a new 2.3 million (11.7%) cases [1].

Breast cancer can develop by a genetic mutation that disrupts the cell division process, resulting in unregulated cell proliferation and tumor growth. Lumps and nodules then start to form in the inner lining of milk ducts [2]. These cancers are of two broad classifications, sarcomas and carcinomas; Carcinomas begin from the breast epithelial cells. It includes ducts for ductal cancers and glands for lobular cancers. The uncommon type of breast cancer among the two is Sarcomas, which comprises less than 1% of breast cancer[3].

In the advanced stage, cancer cells get into the bloodstream and start spreading to the other organs. Often the tumor that arises in the breast leads to metastatic cancer, Like tumor growth in lymph nodes or bones. If the lymph nodes have developed cancer cells, it indicates a greater chance of spreading[3]. The cancer cells travel from the lymph system to the other organs. Since lymph node cancer can lead to metastatic cancer, diagnosing one or more tumor-lymph nodes can impact the therapy. Performing surgery to collect lymph nodes and examining it for the tumor presence will determine the cancer spread [4].

At the same time, it is also the most treatable cancer type if diagnosed early. We can detect cancer by differentiating malignant and benign tumor cells; hence, physicians require a reliable method to distinguish between malignant and benign tumors. With the advanced machine learning (ML) technology, cancer detection accuracy has also increased[5]. ML technology can help physicians for early detection of breast cancer, greatly enhancing patient survival rates.

This paper presents seven different machine learning classifying methods: Gaussian naïve Bayes (NB), Logistic Regression (LR), Support Vector Machine - Linear classifier (SVM (Lin)), Support Vector Machine - RBF classifier (SVM (RBF)), Decision Tree (DT), Random Forest (RF), and K Nearest Neighbor (KNN) algorithms, are used on the (WDBC) data, and we performed a comparative analysis of these classifiers to identify which classifier works better in the breast cancer classification. We also performed Models evaluation using classification accuracy, confusion matrix, training and testing accuracy, and receiver operating characteristic curves (ROC).

We used the Wisconsin Dataset, which comprises digitized FNA images of breast mass with their computed features and Divided the Dataset in the following order to implement the algorithms, 75% as the training and 25% as the testing Data. We improved the algorithm performance by using a min-max scalar to overcome overfitting and outliers. Moreover, after scaling the Dataset, we applied a feature selection technique (i.e., principal component analysis (PCA)) to increase the algorithm accuracy by decreasing the number of parameters (Ivančaková et al., 2018; Wu and Faisal, 2020). We manually allotted the hyper-parameters used for each classifier. All the ML mentioned above algorithms performed great, all exceeding 90% test accuracy.

2. DATA DESCRIPTION

The Data used in this project is the WDBC Dataset collected from an open-source UCI machine learning repository. The fine needle aspiration (FNA) breast cells are digitized and computed. The digitized attributes of breast cells are used in the Data to explain the features of the cell nucleus [6] shown in Table 1. The total instance of the Dataset is 569, of which 212 cases are cancerous (malignant), and 357 are non-cancerous (benign). It includes a total of 32 attributes:

1. The ID number,
2. Diagnosing result (B= Benign or M= Malignant)
3. 30 features.

fractal_dimension_mean	fractal_dimension_worst	fractal_dimension_se
texture_mean	texture_worst	texture_se
radius_mean	radius_worst	radius_se
smoothness_mean	smoothness_worst	smoothness_se
concavity_mean	concavity_worst	concavity_se
area_mean	area_worst	area_se
compactness_mean	compactness_worst	compactness_se
concave points_mran	concave points_worst	concave points_se
symmetry_mean	symmetry_worst	symmetry_se
perimeter_mean	perimeter_worst	perimeter_se

Table 1. Features of The Dataset

3. METHODOLOGY

The flowchart explains the overall workflow of the project. The Data gathered in this project was acquired under the UCI machine repository, Wisconsin Diagnostic Breast Cancer Dataset[7], and then checked for enough instances for developing Models. Then Dataset is pre-processed to increase the quality to get a precise Dataset for Modelling. Feature selection and extraction are selected based on their prediction accuracy, then used in the prediction phase. The Data pre-processing involves partitioning the Data into 75% training and 25% testing.

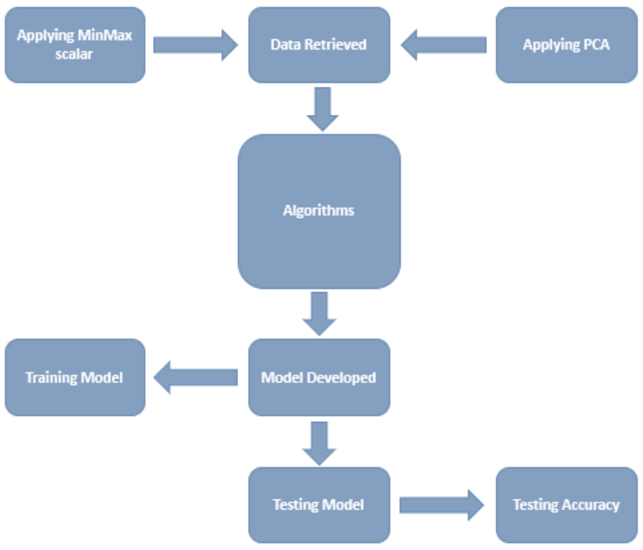


Fig. 1. Flowchart

Most widely used machine algorithms with PCA-based dimensionality reduction techniques [8], [9] were implemented to develop the Models. We used the classifiers like Gaussian naïve Bayes (NB), SVM (Lin), SVM (RBF), LR, DT, KNN, and RF. The accuracy rate of each Model is recorded and compared to get a better performance Model.

3.1 Applying MinMax Scalar to The Data:

The MinMax scalar also termed normalization, overcame the outliers and overfitting of our Dataset. This scalar unit limits the Model parameters values to avoid overfitting, achieved by imposing a magnitude-based penalty to the parameters. The value of this scalar unit lies between 0 and one. The MinMax formula is below;

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

3.2 Feature Extraction Using Principal Component Analysis:

It is an analytical technique commonly applied for Data interpretation. We used PCA for feature extraction[10]; PCA converts the given Dataset into small uncorrelated variables called PC- Principal Components [11], [12]. The PC of the first variables are the converted variables. The highest variance value is in the first fundamental variables, and it conveys Data about the relative sizes.

The feature reduction method PCA is defined as follows, A t-dimensional Data is M. The orthonormal axes are "n" principal axes' R1, R2, . . ., Rn here 1 ≤ n ≤ t. On the axes' projected space, the retaining variance is at its maximum. The sample covariance matrix n leading eigenvectors are R1, R2, . . ., Rn.,

$$C = \left(\frac{1}{L}\right) \sum_{k=1}^{L} (x_k - \bar{x})^T (x_k - \bar{x}).$$

Here $x_k \in M$, mean of samples = $\bar{x}$, number of samples = L. By this:

$$UR_k = v_k R_{K,} \quad K \in 1, \cdots n,$$

U's, kth and larger eigenvalue = v_k. Following is the PC “n” observational vector $x_k \in M$. Below, P is the n PC of x.

$$P = [p_1, p_2, p_3, \ldots\ldots, p_n] = \left[[R_1^T x, R_2^T x, \ldots\ldots\ldots, R_N^T x] = R^T x\right]$$

The principal components (PC) by the combination of attributes account for the most Dataset variance. In this paper, we performed the PCA on the retrieved Data, the n_component is 11, and the random state is zero. Fig. 2 shows binary scatterplots of the PCA between the First PC and Second PC.

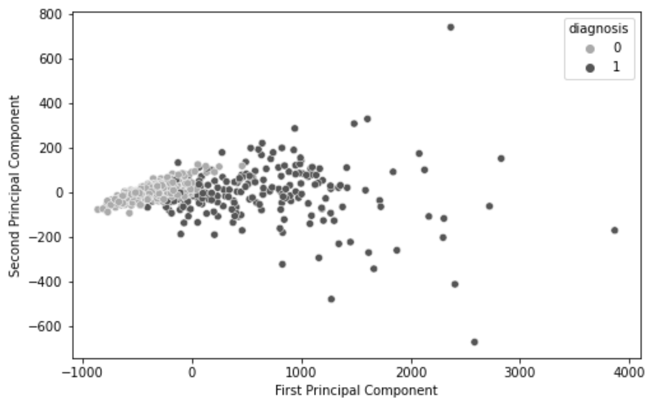


Fig. 2. PCA scatterplot

3.3 Decision Trees:

Decision trees use branches and nodes to shape/form like trees; this algorithm uses many algorithms to split a node into added sub-nodes [11]. Every node represents a feature; the purity of nodes rises with the target variable. Determining the uncertainty in the given Dataset by entropy. H(K) the entropy function defined as,

$$H(K) = \sum_{c \in C} -p(C)\log_{2} p(c)$$

The Data set is K, p(c) refers to the "no. of elements" proportionally correlated to class (c) and Data. C is the class of Dataset (i.e., "M" malignant or "B" benign).

3.4 Naïve Bayes:

It is a simple classifier for the different problems based on classification. The algorithm has the extreme ability for predicting and producing an effective result for a large Data set. The commonly used methods are conditional probability and class probability. Let the classes be R1, R2……Rn and vector be T as shown below;

$$P\left(\frac{Ri}{T}\right) = \left[p\left(\frac{T}{Ri}\right).p(Ri)\right]/p(T)$$

Conditional Probability= p(T/Ri); Priori Probability= p(Ri); Mixture Density= p(T).

In this case, it would classify the Data as benign or malignant by calculating true and false positives.

3.5 KNN Algorithm:

For regression and classification problems, KNN is applied. It saves all possible cases and initiates a new case classification based on measuring correlation (i.e., distance functions). If each of the objects is extremely close to one another, then certainly the characters among them are also very close. To measure the K nearest neighbors the distance function is used. When K = 1, the class allotted to its nearest neighbor. Given below the Euclidian distance formula;

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

3.6 Logistic Regression:

From statistics, Machine Learning adopted Logistic Regression. Problems with binary classification more frequently adapted this classifier. Logistic Regression measures the correlation of More continuous independent variables and a discrete dependent variable. The equation is below:

$$x = e^{\wedge}(b0 + b1^{*}y)/(1 + e^{\wedge}(b0 + b1^{*}y))$$

y – single input value; x – expected output

3.7 Random Forest:

Training and the learning-based algorithm is Random Forest. With the bagging technique, it trains the decision tree set. And its 'forest' is built. The increasing outcome of the learning Model is the basic idea behind the bagging technique. This algorithm creates multiple decision trees and fuses them to get more accurate and precise predictions. Fig. 3. explains how the RF algorithm works.

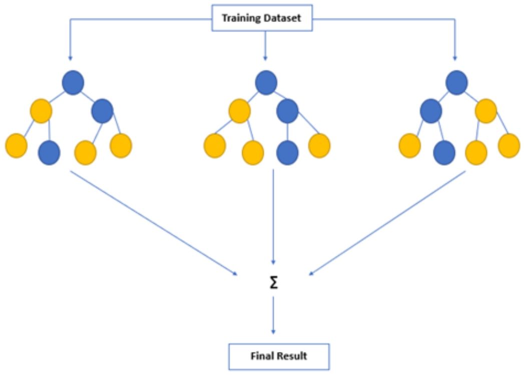


Fig. 3. Random forest

3.8 Support Vector Machine (SVM):

We used SVM for classifying problems. SVM is a maximal margin classifier. This classifier builds More hyperplanes in the high-dimensional feature space. There should be a sizeable plausible distance between the support vector and its hyperplane for the SVM classifier to work well. Misclassification is also high in the Lesser distance margin.

4. SYSTEM REQUIREMENTS

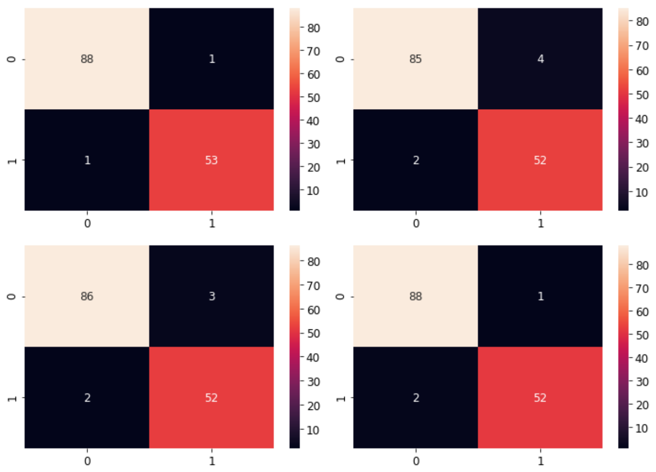
In this paper, we used tools like the anaconda software and jupyter notebook. The software Anaconda is open-source and used for Data processing, Data science, and scientific computing. We coded the Models using Python and R programming languages in Anaconda version 3 – Jupyter notebook. For Visualizing the features and Analyzing the Correlation of the Dataset, we used Jupyter notebook. Then, implemented Python-based PCA dimensionality reduction techniques for developing machine learning Models using various algorithms.

5. DISCUSSION

5.1 Evaluation Metrics

5.1.1 Confusion Matrix:

In general, a confusion matrix is a summarized form of the predicted outcome of any classification analysis. The matrix is summarized, into several correct and incorrect predictions with values. All Models undergo this classification process for their performance evaluation. It consists of (False-Negative - FN), (True-Positive - TP), (False-Positive - FP), and (TN - True-Negative) [11].



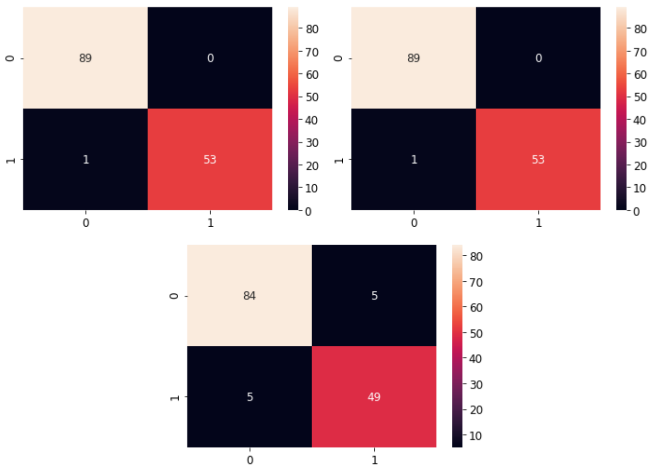


Fig. 4. Confusion matrix

5.1.2 Precision $= \frac{TP}{TP+FP}$

5.1.3 Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$

5.1.3 F1-Score $= 2 * \frac{Precision * Recall}{Precision + Recall}$

5.1.4 Recall $= TP - Rate = \frac{TP}{TP+FN}$

Algorithm	Scaling unit	Dimensionality reduction method	Testing accuracy	Precision	Recall
SVM	MinMax	PCA	99.30%	99%	99%
SVM	MinMax	NO	97.20%	97%	97%

Table 2. SVM Evaluation Metrics

Table 2 shows the comparison of obtained performance metrics between SVM with PCA and without PCA. And Fig. 5 shows a comparison of the relatively higher testing accuracy of SVM without PCA (97.20%) than with PCA (99.30%) of about 2.1 percent increase.

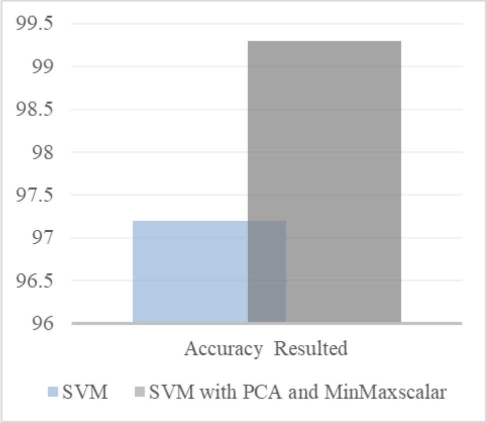


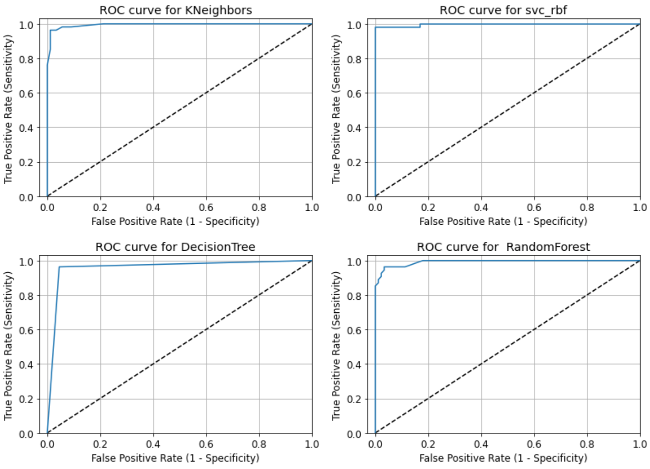
Fig. 5. Accuracy Comparison

5.1.5. Receiver Operating Characteristic (Roc) Curve

The ROC plots assess the performance of the Models. It illustrates diagnosing ability for any binary classifier by using a ROC curve through graphical plots. In the ROC graph, On the x-axis is FPR (i.e., false-positive rate) and on the y-axis is TRP (i.e., True-positive rate). The scales of the axis are between 0 and 1. The graph is obtained by plotting every possible threshold value of the classifier. We visualized the classification Model performance Using the ROC curve [10].

True-positive rate, TRP = $\frac{\text{TP}}{\text{TP+FN}}$; is the fraction of correctly classified positives divided by total positives (Hassan et al., n.d.), and FRP (False positive rate) = $\frac{\text{FP}}{\text{FP+TN}}$; is the fraction of incorrectly classified negatives divided by total negatives.

The performance of these classifiers has been compared to identify which classifier has the highest testing accuracy in the breast cancer classification. Below are the results and images of the working of the project.



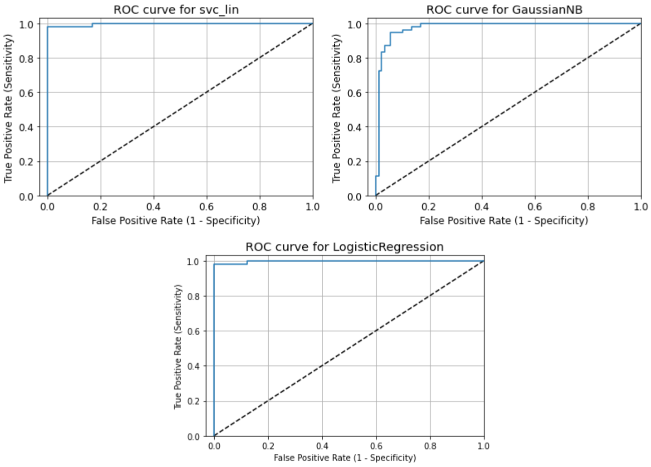


Fig. 6. Roc Curve

The area below the ROC curves is for determining and evaluating the performance of the classifiers. Fig.6 is the ROC curves of the developed models, which shows a good range of areas below the curves [6].

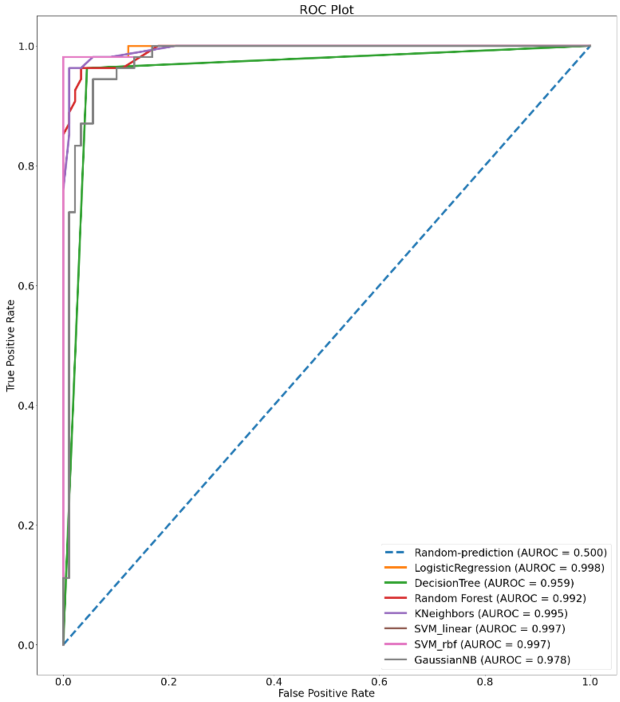


Fig. 7. Roc Curve Comparison

Fig. 7 illustrates all the Models are performing notably well since the Model curves are above or higher than the Random-prediction curve and not in the AUROC region.

Among all other Models, SVM Models performed well, as the SVM curve reaches the threshold by having a high (TPR) true-positive value close to one and a low False-positive value (FPR).

Models	Algorithm	Scaling unit	Dimensionality reduction method	Testing accuracy
1	Logistic Regression (LR) classifier	Minmax	PCA	98.60%
2	Random Forest (RF) Classifier			96.50%
3	Decision Tree (DT) Classifier			95.80%
4	KNN			97.90%
5	SVC_Linear			99.30%
6	SVM_rbf			99.30%
7	GaussianNB			93.01%

Table 3. Models Developed Using MinMax Scalar and PCA

All the Models in Table 3 used Minmax scaling features along with PCA. Each algorithm hyperparameter is tuned to obtain a high-functioning Model. The best-performed Models are SVM (lin) and SVM (RBF), achieving 99.30% testing accuracy. The low-performed Model is GAUSSIAN NB, with the lowest testing accuracy of 93.01%. All the Models are executed and monitored through performance metrics for evaluation.

6. COMPARISON WITH VARIOUS PUBLISHED MODELS

We performed a comparative study with our Models and already published Models. Most of our outputs have similar scores to the existing Models. However, SVM (lin) and SVM (RBF) Models resulted in 99.30% testing scores outperforming the average of other published outcomes [14]. we used Dimensionality reduction techniques and hyperparameter tuning to all our Models. The comparison of the available (WDBC) Models is shown in Table 4.

S.no	Algorithm	Reference	Accuracy
1.	SVM,	[15]	97.13%
	NB,		95.99%
	K-NN		95.27%
	DT		95.13%
2.	SVM	[16]	97%
	RF		96.6%
	BN		97.2%
3.	LOGISTIC REGRESSION + NN	[17]	98.50%
4.	SVM	[18]	96.71%
5.	DT	[11]	94.56%
	KNN		97.8%
	LR		96.09%
	GAUSSIAN NB		99.2%
	SVM		83.5%
6.	SVM	(Omondiage et al., 2019b)	96.47%
	SVM-LDA		98.82%
	NEURAL NETWORKS		97.06%
	NEURAL NETWORKS - PCA		97.65%
	NEURAL NETWORKS - LDA		98.82%
	NB		91.18%
	NB-LDA		98.24%
7.	DT	[20]	96.1%
	RF		95.1%
	SVM		96.1%
	NN		93.7%
	LR		95.6%
8.	K-NN	[21]	93.7%
9.	SVM	[22]	98%
	NAÏVE BAYES		95%
10.	LOGISTIC REGRESSION	[Our Models]	98.6%
	DT		95.8%
	RF		96.5%
	KNN		97.9%
	SVM LINEAR		99.3%
	SVM RBF		99.3%
	GAUSSIANNB		93.01%

Table 4. Comparison of Existing Models

7. CODE

https://github.com/Habi-naya/Improved-Breast-cancer-Diagnosis-using-Machine-learning.git

8. CONCLUSION

In this project different supervised machine learning algorithms along with dimensionality reduction method is used to analyze WDBC dataset to classify malignant and benign cancer. The main idea of this paper is to combine dimensionality reduction (i.e., PCA) with ML algorithms to develop Models. Later various performance metrics are performed to analyze the outcome. Table 4 shows SVM (lin) and SVM (RBF) obtained a very high testing accuracy of about 99.30%, followed by the logistic classifier with 98.6%. The results showed that SVM linear with PCA and SVM RBF with PCA outperforms the other ML Models. The final testing accuracy of SVM (99.30%) shows this chosen approach can be a potentially very assuring and reliable diagnosing Model.

9. REFERENCES

[1] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality

Worldwide for 36 Cancers in 185 Countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[2] M. Akram, M. Iqbal, M. Daniyal, and A. U. Khan, "Awareness and current knowledge of breast cancer," *Biological Research*, vol. 50, no. 1. BioMed Central Ltd., 2017. doi: 10.1186/s40659-017-0140-9.

[3] Y. Feng *et al.*, "Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis," *Genes and Diseases*, vol. 5, no. 2, pp. 77–106, 2018, doi: 10.1016/j.gendis.2018.05.001.

[4] A. Rechf and M.] Houlihan, "Axillary Lymph Nodes and Breast Cancer A Review."

[5] L. R. Borges, "Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection," *Proceedings of XI Workshop de Visão Computacional*, no. December, pp. 15–19, 2015, [Online]. Available: https://www.researchgate.net/publication/311950799

[6] D. A. Omondiagbe, S. Veeramani, and A. S. Sidhu, "Machine Learning Classification Techniques for Breast Cancer Diagnosis," in *IOP Conference Series: Materials Science and Engineering*, 2019, vol. 495, no. 1. doi: 10.1088/1757-899X/495/1/012033.

[7] B. Karthikeyan, S. Gollamudi, H. V. Singamsetty, P. K. Gade, and S. Y. Mekala, "Breast cancer detection using machine learning," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 2, pp. 981–984, 2020, doi: 10.30534/ijatcse/2020/12922020.

[8] H. J. Chiu, T. H. S. Li, and P. H. Kuo, "Breast cancer–detection system using PCA, multilayer perceptron, transfer learning, and support vector machine," *IEEE Access*, vol. 8, pp. 204309–204324, 2020, doi: 10.1109/ACCESS.2020.3036912.

[9] I. Journal and I. Computing, "Based on Association Rules and Pca for Detection," *International Journal of Innovative Computing*, vol. 9, no. 2, pp. 727–739, 2013.

[10] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert Systems with Applications*, vol. 36, no. 2 PART 2, pp. 3240–3247, 2009, doi: 10.1016/j.eswa.2008.01.009.

[11] Z. Mushtaq, A. Yaqub, A. Hassan, and S. F. Su, "Performance analysis of supervised classifiers using PCA based techniques on breast cancer," *2019 International Conference on Engineering and Emerging Technologies, ICEET 2019*, pp. 1–6, 2019, doi: 10.1109/CEET1.2019.8711868.

[12] M. S. Uzer, O. Inan, and N. Yilmaz, "A hybrid breast cancer detection system via neural network and feature selection based on SBS, SFS and PCA," *Neural Computing and Applications*, vol. 23, no. 3–4, pp. 719–728, Sep. 2013, doi: 10.1007/s00521-012-0982-6.

[13] M. R. Hassan, K. Ramamohanarao, C. Karmakar, M. Maruf Hossain, and J. Bailey, "A Novel Scalable Multi-class ROC for Effective Visualization and Computation."

[14] W. Wu and S. Faisal, "A data-driven principal component analysis-support vector machine approach for breast cancer diagnosis: Comparison and application," *Transactions of the Institute of Measurement and Control*, vol. 42, no. 7, pp. 1301–1312, 2020, doi: 10.1177/0142331219889221.

[15] H. Asri, H. Mousannif, H. al Moatassime, and T. Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," *Procedia Computer Science*, vol. 83, no. Fams, pp. 1064–1069, 2016, doi: 10.1016/j.procs.2016.04.224.

[16] A. al Bataineh, "A comparative analysis of nonlinear machine learning algorithms for breast cancer detection," *International Journal of Machine Learning and Computing*, vol. 9, no. 3, pp. 248–254, 2019, doi: 10.18178/ijmlc.2019.9.3.794.

[17] N. Khuriwal and N. Mishra, "Breast Cancer Diagnosis Using Deep Learning Algorithm," *Proceedings - IEEE 2018 International Conference on Advances in Computing, Communication Control and Networking, ICACCCN 2018*, pp. 98–103, 2018, doi: 10.1109/ICACCCN.2018.8748777.

[18] S. Ghosh, S. Mondal, and B. Ghosh, "A comparative study of breast cancer detection based on SVM and MLP BPN classifier," *1st International Conference on Automation, Control, Energy and Systems - 2014, ACES 2014*, pp. 1–4, 2014, doi: 10.1109/ACES.2014.6808002.

[19] D. A. Omondiagbe, S. Veeramani, and A. S. Sidhu, "Machine Learning Classification Techniques for Breast Cancer Diagnosis," *IOP Conference Series: Materials Science and Engineering*, vol. 495, no. 1, 2019, doi: 10.1088/1757-899X/495/1/012033.

[20] Y. Li, "Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction," *Applied and*

Computational Mathematics, vol. 7, no. 4, p. 212, 2018, doi: 10.11648/j.acm.20180704.15.

[21] M. R. Al-Hadidi, A. Alarabeyyat, and M. Alhanahnah, "Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm," *Proceedings - 2016 9th International Conference on Developments in eSystems Engineering, DeSE 2016*, pp. 35–39, 2017, doi: 10.1109/DeSE.2016.8.

[22] S. Nayak and D. Gope, "Comparison of supervised learning algorithms for RF-based breast cancer detection," *CEM 2017 - 2017 Computing and Electromagnetics International Workshop*, pp. 13–14, 2017, doi: 10.1109/CEM.2017.7991863.