Global Journal of Computer Science and Technology: H Information & Technology
Volume 20 Issue 1 Version 1.0 Year 2020
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

A Comparative Study between a Simulation of Machine Learning and Extreme Learning Machine Techniques on Breast Cancer Diagnosis

By Rahul Reddy Nadikattu
University of the Cumberlands

Abstract- Breast Cancer is a developing and most normal disease among ladies around the globe. Breast malignancy is an uncontrolled and exorbitant development of abnormal cells in the Breast because of hereditary, hormonal, and way of life factors. During the starting stages, the tumor is restricted to the Breast, and in the latter part, it can spread to lymph hubs in the armpit and different organs like the liver, bones, lungs, and cerebrum. At the point when the bosom disease spreads to different pieces of the body, it is going to metastasize. The sickness is repairable in the beginning periods, yet it is identified in later stages, which is the fundamental driver for the passing of such a large number of ladies in this entire world. Clinical tests led in medical clinics for deciding the malady are a lot of costly, just as tedious as well.

Keywords: extreme learning machine (ELM), random forest (RF), decision tree (DT), support vector machine (SVM), and KNN.

GJCST-H Classification: I.2.m

Strictly as per the compliance and regulations of:

A Comparative Study between a Simulation of Machine Learning and Extreme Learning Machine Techniques on Breast Cancer Diagnosis

Rahul Reddy Nadikattu

Abstract- Breast Cancer is a developing and most normal disease among ladies around the globe. Breast malignancy is an uncontrolled and exorbitant development of abnormal cells in the Breast because of hereditary, hormonal, and way of life factors. During the starting stages, the tumor is restricted to the Breast, and in the latter part, it can spread to lymph hubs in the armpit and different organs like the liver, bones, lungs, and cerebrum. At the point when the bosom disease spreads to different pieces of the body, it is going to metastasize. The sickness is repairable in the beginning periods, yet it is identified in later stages, which is the fundamental driver for the passing of such a large number of ladies in this entire world. Clinical tests led in medical clinics for deciding the malady are a lot of costly, just as tedious as well. The answer to counter this is by directing early and exact findings for quicker treatment, and accomplishing such exactness in a limited capacity to focus time demonstrates troublesome with existing techniques [1]. In this paper, we look at changed AI and neural system calculations to foresee malignant growth in beginning times, intending to save the patient's life. Wisconsin Breast Cancer (WBC) dataset from the UCI AI vault has been utilized. Various calculations were looked in particular Support Vector Machine Classification (SVM), K-Nearest Neighbor Classification (KNN), Decision tree Classification (DT), Random Forest Classification (RF) and Extreme Learning Machine (ELM) and they thought about based on precision and handling time taken by each. The outcomes show that an extreme learning machine gives the best outcome for both the ideal models.

Keywords: extreme learning machine (ELM), random forest (RF), decision tree (DT), support vector machine (SVM), and KNN.

I. Introduction

Breast Cancer has become the principal explanation for the passing of many ladies worldwide. The principle explanation behind the passing of ladies by this infection is the procedure by which is analyzed. The innovation has become a significant part of our ways of life; we are still missing behind diagnosing this essential ailment in early stages[2]. As the ailment isn't analyzed in beginning times, along these lines, the mammography rate has expanded for a specific age gathering of concerned women[3].

Breast Cancer is reparable, and life could be spared on the off-chance, and it would analyze in beginning times. Various causes have been analyzed for this dreadful malady, specifically, hormonal awkwardness, family ancestries, corpulence, radiation treatments, and some more. Many AI and profound learning calculations were applied to diagnose this ailment.

Machine learning algorithms follow the following steps during classification problems[1]:

- Data Collection
- Appropriate Model selection
- Modeler is trained
- Testing and prediction of results

In this paper, we analyzed different Machine Learning calculations and a neural system (ELM) to discover which calculation gives the best outcome as far as precision and preparing time. Different AI calculations examined here are Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree (DT), and Random Forest (RF). The neural system talked about here is the Extreme Learning Machine (ELM).

II. Literature Review

In this section, some of the previous work done by different researchers on different breast cancer Datasets had discussed.

In [4], the dataset was taken from the Iranian center of breast cancer, and the performance of the decision tree, support vector machine, and artificial neural network was compared. The support vector machine was proven to be the best followed by an artificial neural network, and then the decision tree classification model.

In [5], two datasets were taken for performing comparison among different machine learning models. The datasets were WPBC(Wisconsin Prognostic Breast Cancer) and Wisconsin breast cancer dataset. The comparison was between the decision tree classification model, Naïve Bayes model, neural network, and support

Author: Ph.D. in Information Technology, San Jose, United States, University of the Cumberlands. e-mail: rahulnadi40@gmail.com

vector machine with different kernels. Results showed that the neural network was best for the Wisconsin breast cancer dataset and support vector machine with radial basis function (RBF) and was best for the WPBC dataset.

In [6], an ANN (Artificial neural network) with Principal Component Analysis (PCA) is used to distinguish between malign and benign tumor cells.

In [7], the WPBC dataset is used for comparing the performance of different machine learning algorithms. The result showed that the support vector machine and decision tree were among the best predictors of results.

In [8], a multi-layer perceptron with backpropagation neural network and support vector machine uses for the classification dataset. Support Vector machine found to be the best result giving algorithm.

In [9], a signal-to-noise ratio technique is used with different classification models: k-nearest neighbor, SVM, and PNN, which is a probabilistic neural network. SVM with a radial basis function, that is RBF kernel concluded to give the best result.

In [10], a comparative study was done on the random forest classification model, Naïve Bayes model, and Support Vector Machine model to analyze the Wisconsin breast cancer dataset on the parameters of precision, accuracy, and specificity.

In [11], a new approach was provided, based on the neural network with a feed forward BP algorithm. Wisconsin breast cancer dataset used from the UCI repository. A 7 hidden unit neural network used to obtain the results.

In [12], the relevance vector machine (RVM) compares with other machine learning techniques. Linear Discriminant Analysis (LDA) method was used for dimension reduction. RVM gave the best results in their experiment on the WBC dataset.

III. Machine Learning Classification Models used

a) Support Vector Machine Classification

It is the directed AI arrangement method that separates the dataset into classes utilizing an appropriate maximal edge hyperplane, for example, the upgraded choice boundary[1]. The methods utilized in numerous fields, such as infection acknowledgment, penmanship acknowledgment, discourse acknowledgment, and numerous different fields, of example, acknowledgment. This strategy builds the gap between the classes, which it makes in figure1.

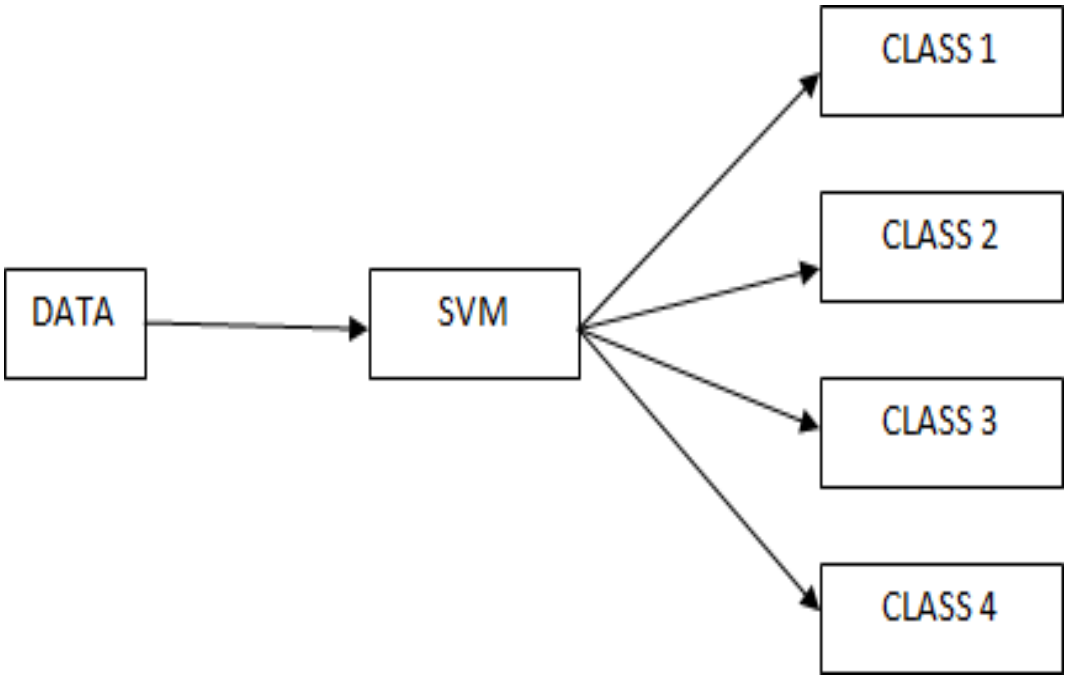

Figure 1: Different Classes via SVM

An SVM model which utilizes part as a "Sigmoid" bit can be considered as a two-layer neural system. SVM can be utilized with various pieces like "direct", "poly", "spiral premise work (RBF)" etc[13]. SVM is a regulated AI calculation that utilizes both characterization and regression[14]. In this, every datum point is plotted in n-dimensional space, and afterward, a hyperplane or line dictates by grouping. Figure2 wonderfully recognizes the two classes as the focuses on the left half of the line are in green circle class, and information focuses on the right side of line fall in red circle class. As SVM is a multi-dimensional space in this way, each point turns into a vector here.

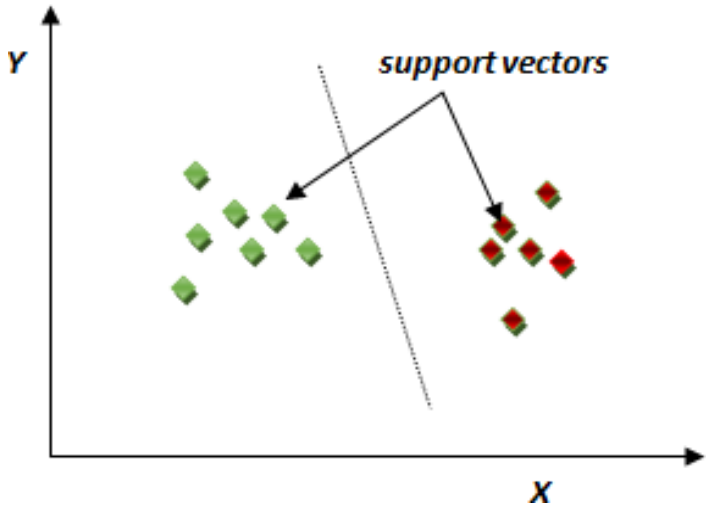

Figure 2: Support Vector Machine Classification

b) K-Nearest Neighbor Classification

KNN classification is an advantageous and straightforward classification method that can be implemented very quickly. The ideology is to find K's most similar samples from feature samples[15]. KNN is measured by finding the distance between eigenvalues, which is also known as Euclidean distance, as in figure3. The number of K neighbors is predetermined. Firstly, the default value taken for K is usually 5. Then, K nearest neighbors of a new data points are taken. Among these K neighbors, data points are counted in each category, and the new data point is assigned to the category for which you counted the most neighbors.

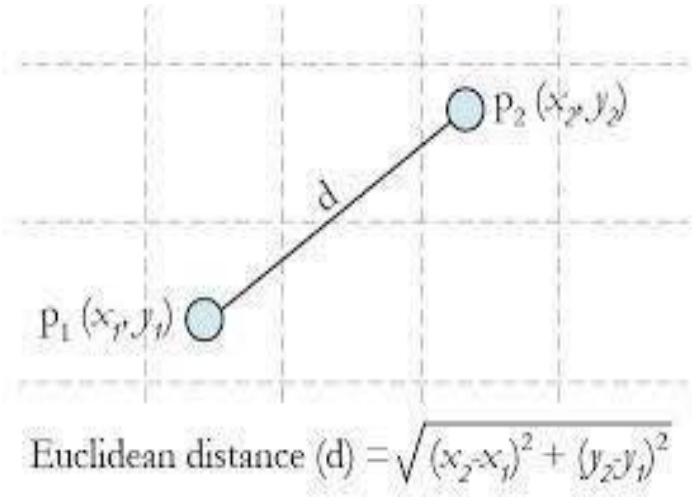

Figure 3: K-Nearest neighbour classification The Euclidean distance is calculated as below:

$$5h555\ (5_0, 5_0)\ 55555555555\ 55\ 555\ 55555\ 55^2\ 5\ (5_1, 5_1)^2 55555555555\ 55\ 55h55\ 55555.$$

c) Decision Tree Classification

Decision trees are also called a choice tree. The choice tree is a stream diagram in which the dataset is a part of the way with the goal that each part area has the most extreme number of information focuses, as in figure 4. Choice trees parcel the info space into cells where every cell has a place with one class[16]. Dividing is finished by the tests performed on the dataset. Every hub brings forth two streets, either a specific condition or a bogus one. It is a prescient model that could be viewed as a tree. Leaves of this tree speak to divided datasets. In this calculation, the best information point is root. In this calculation, we started with a pull for depicting the class of a record. In this information point's qualities are contrasted, and inward hubs of the choice tree will arrive at the leaf hub with the anticipated class.

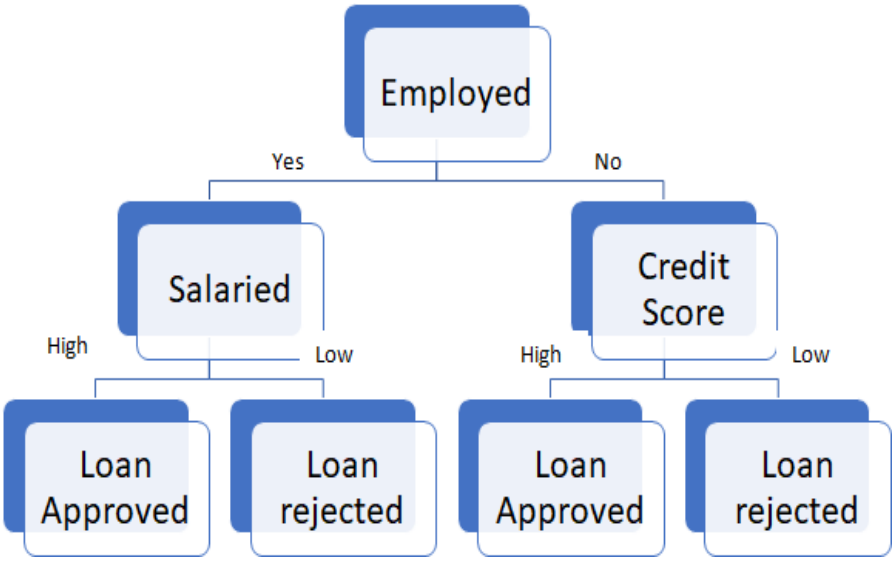

Figure 4: Decision Tree Classification

d) Random Forest Classification

Random forest is a version of ensemble learning, and it follows a bagging technique, as in figure 5. The base model used in the random forest is the decision tree. This algorithm selects data points randomly and creates multiple trees or forests. In this, random K data points are selected from the data set, and decision trees are built for these data points.

Samples were taken with a replacement, but trees are related in such a manner, so that the correlation between classifiers could be reduced. As it is an ensemble learning algorithm, it provides the best results with accuracy and in very less processing time. Fig 5. of Random Forest Classification diagram is shown below

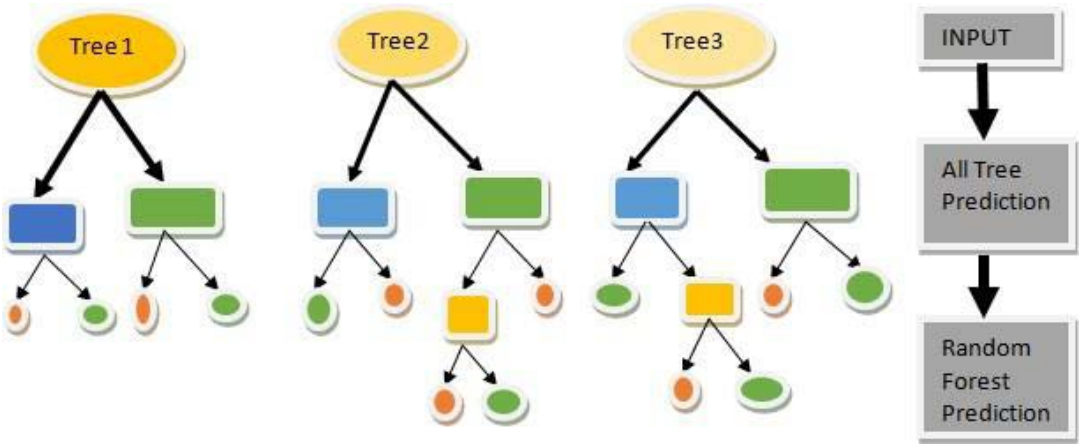

Figure 5: Random Forest Classification

e) Extreme Learning Machine

Extreme Learning Machine (ELM) is a technique that is used for a single hidden layer feed forward neural network that randomly chooses hidden nodes and determines the output weights[17] as in figure6. This method only has one input layer, one hidden layer, and one output layer. It is a bit different from traditional Back propagation algorithms.ELM sets the number of hidden neurons, and randomly weights are assigned between the input layer and hidden layers with the bias value of hidden units, then the output layer is calculated by using the Moore Penrose pseudo inverse method[18]. This algorithm provides an exceptional fast processing speed and great accuracy. When ELM compares with traditional neural network techniques, it found to be more convincing as it overcomes the overfitting problems[19]. Figure 6 is an ELM consisting of n input layer neurons, l hidden neurons, and m output layer neurons. The algorithm for ELM is as follow:

The training sample is [X,Y] = {xi, yi} (i= 1,2,……Q) and X and Y matrices can be described as

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1Q} \\ x_{21} & x_{22} & \cdots & x_{2Q} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nQ} \end{bmatrix}$$

$$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1Q} \\ y_{21} & y_{22} & \cdots & y_{2Q} \\ \vdots & \vdots & \ddots & \vdots \\ y_{m1} & y_{m2} & \cdots & y_{mQ} \end{bmatrix}$$

ELM generate the weights matrix for the input layer as

$$W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{l1} & w_{l2} & \cdots & w_{ln} \end{bmatrix}$$

Biases are assumed between the hidden layer and output layer as:

$$\beta = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{l1} & \beta_{l2} & \cdots & \beta_{lm} \end{bmatrix}$$

Bias is randomly set for hidden layer neurons as:

$$B = [b_1 b_2 \ldots b_n]^T$$

An activation function is chosen and according to figure 6 the output matrix can be expressed as

$$T = [t_1 t_2 \ldots t_Q]_{mxQ}$$

Calculate the Moore-Penrose pseudo inverse of the matrix.

Calculate the output weight matrix H as:

$$H\beta = T',$$

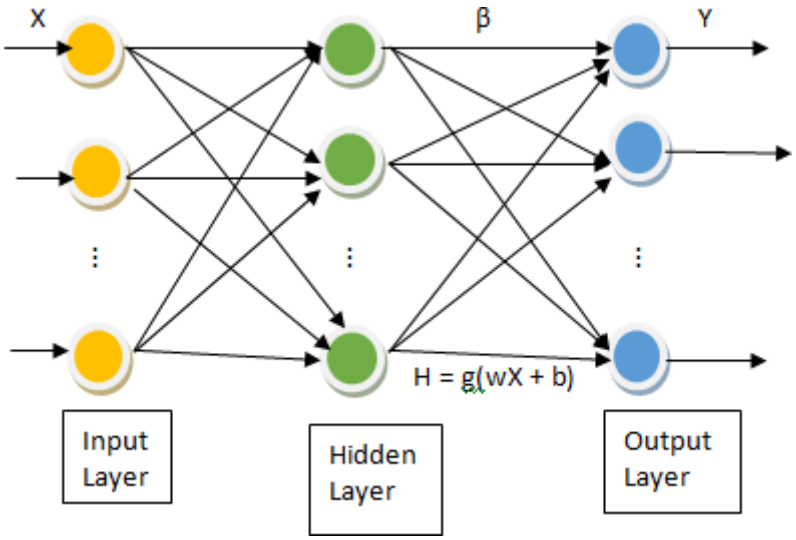

Figure 6: ELM Neural Network

IV. Methodology

We applied various algorithms, as mentioned above, on the Wisconsin Breast Cancer dataset taken from the UCI repository. We used Anaconda Spyder as a platform for coding with Python version 3.7. The methodology includes various techniques like Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision tree (DT), Random Forest (RF), and Extreme Learning Machine (ELM) with dimension reduction techniques that is Principal Component Analysis (PCA).

In this paper, after reading the dataset, the preprocessing of data is done by splitting the dataset into a training set and testing set. The ratio used for splitting the dataset is 75:25. Python API Scikit-learn is used to perform different tasks. After splitting of data, feature scaling would be done. It helps normalize the data within a range so that the algorithm speed can increase. After normalization of data, dimensions are reduced. In this paper, PCA is used for this purpose, and the process had explained below.

a) Dimension Reduction

The process of reducing independent variables to principal variables is known as dimension reduction[14]. This process reduces the dimensions of the dataset so that data can be better viewed and utilized better. It is explained in figure7 below. It comprises of the below method[1]:

Feature Selection: Finding a subset of original features by applying different methods according to the information provided is the process of finding a subset

of original features. It is a transformation in which data was compressed using linear algebra. PCA is used to reduce the dimensions of the dataset and improve the accuracy of the machine learning algorithm. The PCA algorithm, as in the figure, illustrates the entire working principle. The steps are as follows:

Step 1: the breast cancer dataset is prepared in a matrix form with all the features.
Step 2: Features are scaled or normalised by subtracting average from each dimension to form a data which has no mean at all.
Step 3: Covariance matrix is computed which describes the variance of data and

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$

Step 4: Using above covariance, Eigenvalues and vectors are calculated which are useful in providing information about our data.
Step 5: Eigenvalues are arranged in non-increasing order. The feature with the largest Eigenvalue becomes the principal component of the dataset.
Step 6: A new vector forms which comprise all the principal components of the dataset.

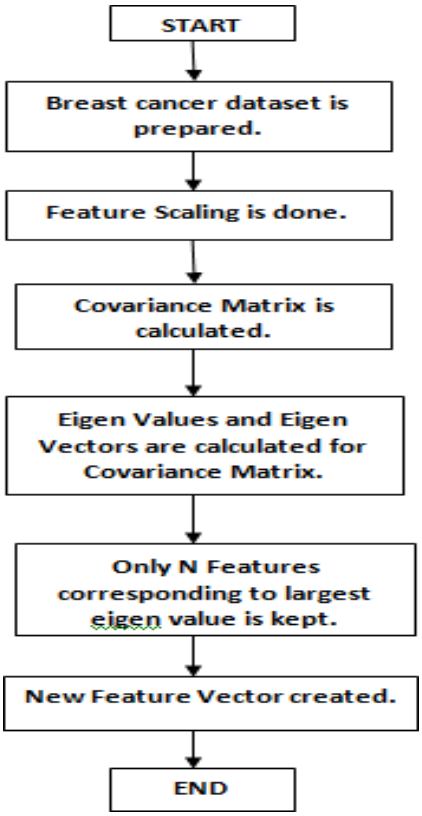

Figure 7: Principal Component Analysis Algorithm Once dimensions are reduced, model is selected for evaluating the results

b) Model Selection

It is the most exciting phase, as in this machine learning algorithm is selected. Machine learning algorithms are categorized into two groups, namely: Supervised and Unsupervised learning algorithms. In the supervised algorithm, the machine is trained on labeled data. Supervised learning algorithms are divided into regression and classification techniques. An unsupervised learning algorithm is a method in which unlabeled information is provided to the machine, and this information is analyzed without any direction. In this dataset, Y is a dependent variable, which is having values either malign (1) or benign (0)[14]. Here classification techniques are applied. In this paper, five algorithms have been chosen namely (already discussed above),

1. K-Nearest Neighbour
2. Support Vector Machine
3. Decision Tree
4. Random Forest
5. Extreme Learning Machine

V. Results

Below is a tabular format which gives results of the experiments performed on the above- discussed dataset by using various techniques. Different techniques used here are being compared based on training and testing accuracies and based on training and testing time taken on the dataset. The results clearly

show that the Extreme Learning Machine is the best among others as it gives 99% accuracy and less time.

Table 1: Performance Comparison

MODEL	ACCURACY		TIME	
	Training (%)	Testing (%)	Training (ms)	Testing (ms)
Decision Tree(DT)	83	88	0.046875	0.015625
K-Nearest Neighbour(KNN)	88	89	0.359375	0.328125
Support Vector Machine(SVM)	90	90	0.0625	0.015625
Random Forest(RF)	93	93	0.15625	0.140625
Extreme Learning Machine(ELM)	94	99	0.046875	0.015625

The below figure 8 shows a bar chart comparison for all the models on the basis of accuracy and time.

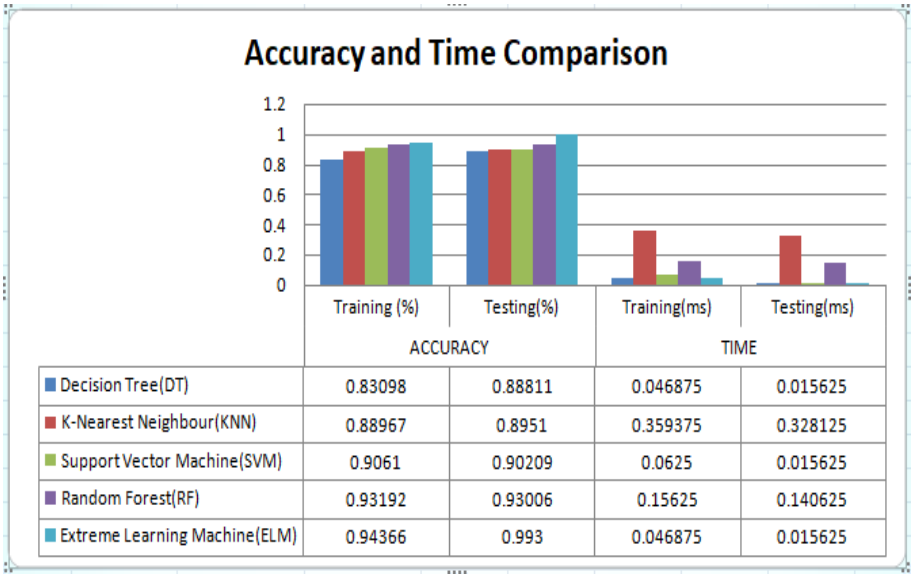

Figure 8: Accuracy and Time Comparison

VI. Conclusion

Extreme Learning Machine (ELM) will be utilized to foresee Breast malignancy with a rough 99% precision rate after 50 ages. This exactness is given with the choice instrument of PCA alongside ELM. This component can be utilized in the future to distinguish the amiable and dangerous cells in beginning periods and can be executed as an application in mammography procedures. There is consistently an opportunity to get better. This research study helps researchers working in the same field.

References Références Referencias

1. S. Anto and 2Dr.S.Chandramathi, "Supervised Machine Learning Approaches for Medical Data Set Classification - A Review," *Int. J. Comput. Sci. Technol.*, vol. 2, no. 4, pp. 234–240, 2011.
2. C. H. Shravya, K. Pravalika, and S. Subhani, "Prediction of breast cancer using supervised machine learning techniques," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 6, pp. 1106–1110, 2019.
3. W. Chenyang, W. Xiaohua, W. Jinbo, and S. Jiawei, "Research on Knowledge Classification Based on KNN and Naive Bayesian Algorithms," *J. Phys. Conf. Ser.*, vol. 1213, p. 032007, 2019.
4. . N. J., "Data Mining Techniques: a Survey Paper," *Int. J. Res. Eng. Technol.*, vol. 02, no. 11, pp. 116–119, 2013.
5. G. Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: Theory and applications," *Artif. Intell. Rev.*, vol. 44, no. 1, pp. 489–501, 2006.
6. D. Xiao, B. Li, and Y. Mao, "A Multiple Hidden Layers Extreme Learning Machine Method and Its Application," *Math. Probl. Eng.*, vol. 2017, 2017.
7. S. Ding, H. Zhao, Y. Zhang, X. Xu, and R. Nie, "Extreme learning machine: algorithm, theory and applications," *Artif. Intell. Rev.*, vol. 44, no. 1, pp. 103–115, 2015.
8. Yadav, I. Jamir, R. R. Jain, and M. Sohani, "Comparative Study of Machine Learning Algorithms for Breast Cancer Prediction - A Review,"

Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol., vol. 5, no. 2, pp. 979–985, 2019.
9. Y. Shen *et al.*, "Globally-Aware Multiple Instance Classifier for Breast Cancer Screening," pp.1–11, 2019.
10. N. E. Benzebouchi, N. Azizi, and K. Ayadi, *Computational Intelligence in Data Mining*, vol. 711. Springer Singapore, 2019.
11. LG and E. AT, "Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence," *J. Heal. Med. Informatics*, vol. 04, no. 02, pp. 2–4, 2013.
12. Z. Nematzadeh, R. Ibrahim, and A. Selamat, "Comparative Studies on Breast Cancer Machine Learning Techniques," *2015 10th Asian Control Conf.*, pp. 1–6, 2015.
13. H. Hasan and N. M. Tahir, "Feature selection of breast cancer based on Principal Component Analysis," *Proc. - CSPA 2010 2010 6th Int. Colloq. Signal Process. Its Appl.*, pp. 242–245, 2010.
14. U. Ojha and S. Goel, "A study on prediction of breast cancer recurrence using data mining techniques," *Proc. 7th Int. Conf. Conflu. 2017 Cloud Comput. Data Sci. Eng.*, pp. 527–530, 2017.
15. S. Ghosh, S. Mondal, and B. Ghosh, "A comparative study of breast cancer detection based on SVM and MLP BPN classifier," *1st Int. Conf. Autom. Control. Energy Syst. - 2014, ACES 2014*, pp. 1–4, 2014.
16. Osareh and B. Shadgar, "Machine learning techniques to diagnose breast cancer," *2010 5th Int. Symp. Heal. Informatics Bioinformatics, HIBIT 2010*, pp. 114–120, 2010.
17. D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," *Int. Conf. Electron. Devices, Syst. Appl.*, pp. 2–5, 2017.
18. M. S. B. M. Azmi and Z. C. Cob, "Breast cancer prediction based on backpropagation algorithm," *Proceeding, 2010 IEEE Student Conf. Res. Dev. - Eng. Innov. Beyond, SCOReD 2010*, no. SCOReD, pp. 164–168, 2010.
19. B. M. Gayathri and C. P. Sumathi, "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer," *2016 IEEE Int. Conf. Comput. Intell. Comput. Res. ICCIC 2016*, pp. 0–4, 2017.
20. Purva Agarwal and Pawan Whig, "Low Delay Based 4 Bit QSD Adder/Subtraction Number System by Reversible Logic Gate", 2016 8th International Conference on Computational Intelligence and Communication Networks (CICN), IEEE Xplore: 26 October 2017
21. Jacob B. Chacko ; Pawan Whig, " Low Delay Based Full Adder/Subtractor by MIG and COG Reversible Logic Gate", 2016 8th International Conference on Computational Intelligence and Communication Networks (CICN), IEEE Xplore: 26 October 2017
22. Ajay Rupani , Pawan Whig , Gajendra Sujediya and Piyush Vyas ,"A robust technique for image processing based on interfacing of Raspberry-Pi and FPGA using IoT," International Conference on Computer, Communications and Electronics (Comptelix), IEEE Xplore: 18 August 2017