# A COMPARATIVE STUDY BETWEEN SIMULATION OF MACHINE LEARNING AND EXTREME LEARNING TECHNIQUES ON BREAST CANCER DIAGNOSIS

Rahul Reddy Nadikattu

University of the Cumberlands Ph.D. in Information Technology Kentucky, United States

## 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 too 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. 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**: Simulation, RF,WBC, hormonal, SVM, Breast Cancer, Machine Learning.

## 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 [1]. 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 [2-3]. 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.

Various Steps followed in Machine learning algorithms are:

i) Data Collection
ii) Model selection
iii) Trained the model
iv) Prediction and accuracy check

In this research studies different Machine Learning calculations are performed and a neural

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

**Previous Work Review**

Many literary works related to breast cancer dataset have been found and studied and some of the previous work done by different researchers on different breast cancer Datasets had discussed in this section.

In a research paper by LG et al., the dataset was taken from the Iranian center of breast cancer, and the performance of the various machine learning algorithms like Decision Tree(DT), Support Vector Machine(SVM), and Artificial Neural Network(ANN) was compared. The SVM was proven to be the best machine learning algorithm followed by an ANN, and at the last DT classification model.

In each other research contemplates Huang et al., two unique datasets were taken for correlation among various ML models. The datasets were Wisconsin Prognostic Breast Cancer and Wisconsin bosom malignancy dataset. They examined different AI calculations, i.e., the choice tree characterization model, Naïve Bayes model, neural system, and bolster vector machine with various bits. Results indicated that the neural network was best for the Wisconsin bosom malignant growth dataset and support vector machine with outspread premise work (RBF) and was best for the WPBC dataset.

Xiao et.al. Used an ANN (Artificial neural network) with Principal Component Analysis (PCA) is used to distinguish between malign and benign tumor cells.

Ding et al. interoperate the WPBC dataset used for comparing the performance of different machine learning algorithms. The result represented that the support vector machine and decision tree were among the best predictors of outcomes.

A multi-layer perceptron with back-propagation neural network and support vector machine uses for the classification dataset is explained by the Yadav et al. and Support Vector machine found to be the best result giving algorithm.

The pertinence vector machine contrasts and other AI strategies. Straight Discriminant Analysis technique was utilized for measurement decrease was talked about by Nematzadeh et.al and it was discovered that RVM gave the best outcomes in their investigation on the WBC dataset.

**Brief introduction to ML Models**

Various machine learning classification models which have their own importance are discussed in subsequent sections.

**Support Vector Machine**

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[13-15]. 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.

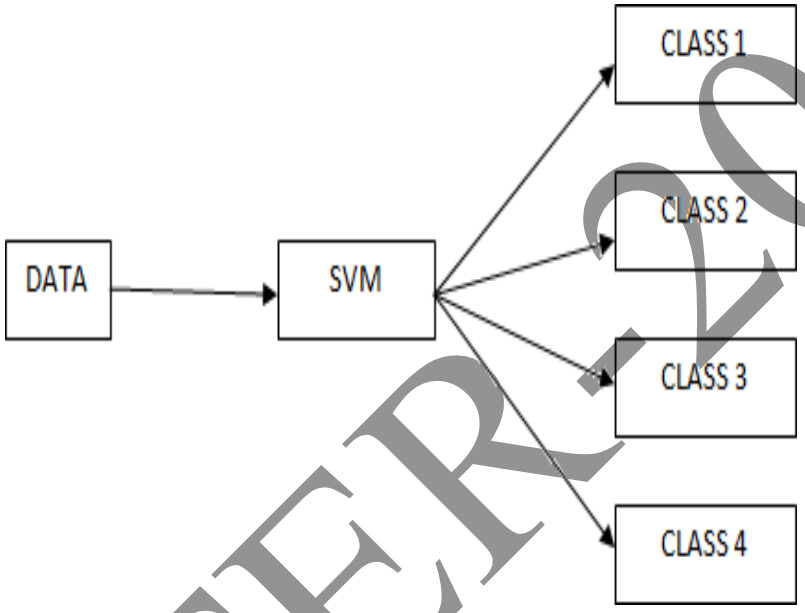


Figure 1: Support Vector Machine With different Classes

An SVM model utilizes Sigmoid function bit can be considered as a two-layer neural system. One of the major application of SVM that it can be utilized with various pieces like "direct", "poly", "spiral premise work (RBF)" etc[16-18]. SVM is a regulated AI calculation that utilizes both characterization and regression[19]. SVM use every datum plotted point as n-dimensional space, and also used a hyperplane or line dictates by grouping. Fig. 2 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 and shown in red circle class. As SVM is a multi-dimensional space and each point turns into a vector here.

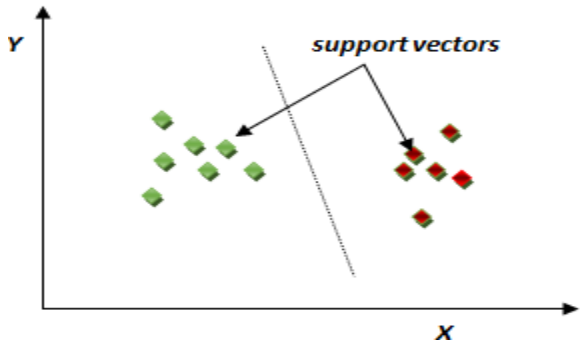


Figure 2: SVM classification example for separating two vectors

**K-Nearest Neighbor Classification**

KNN classification is an efficient, easy and straightforward classification method that can be implemented very quickly on machine learning data. It is based on the working principle of measuring distance of K's most similar samples from feature samples. KNN is measured by finding the Euclidian distance between eigenvalues as shown in Fig. 3.

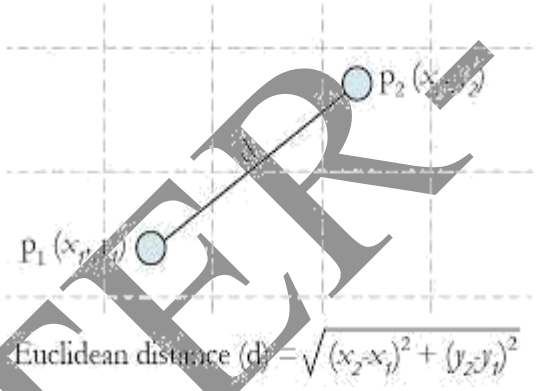


Figure 3 : Euclidean distance for KNN Algorithm

The Euclidean distance between two points P1 and P2 is calculated and shown in Eq. 1

The Euclidean distance is calculated as below:

$$Distance = \sqrt{(x_2 - x_1) + (y_2 - y_1)} \qquad (1)$$

*wh□□□* (□0, □0) □□□□□□□□□□□ □□ □□□ □□□□□ □□2 □ (□1, □1)2□□□□□□□□□□□ □□ □□*h*□□ □□□□□.

## 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 . 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.

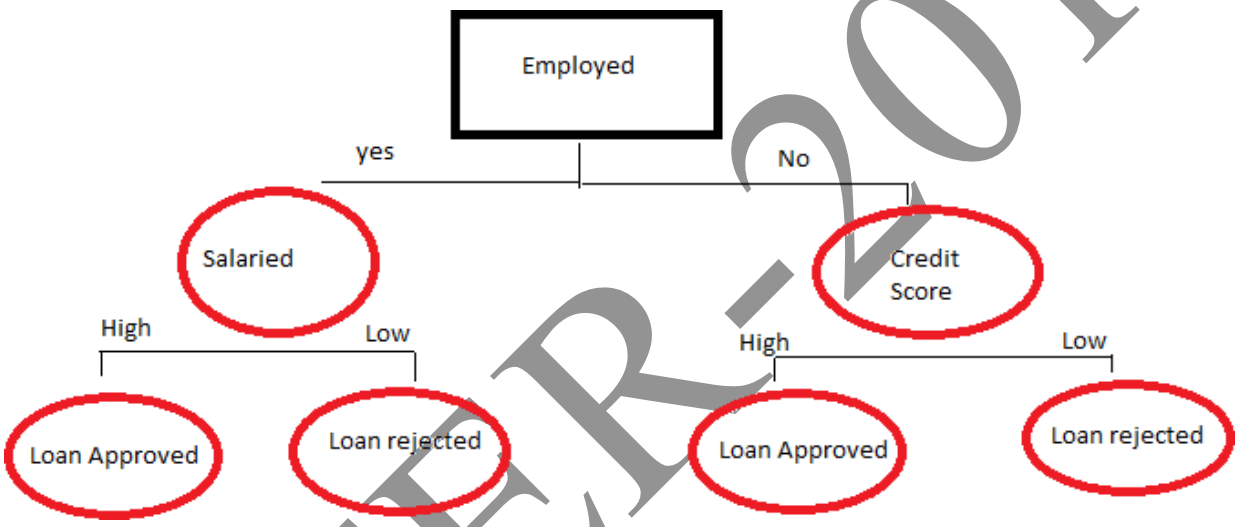


Figure 4: Example of DT Classification

## RF Classification

Random forest ( RF) is a variant of gathering learning, and it follows a packing method, It is based on the Principle of taking all inputs from all trees as shown in Fig. 5. The base model utilized in this algorithm is the choice tree. This calculation chooses information focuses haphazardly and makes numerous trees or backwoods. In this, irregular K information focuses are chosen from the informational collection, and choice trees are worked for these information focuses.

Tests were taken with a substitution, however trees are connected in such a way, so the relationship between's classifiers could be decreased. As it is random calculation, it gives the best outcomes precision and in less preparing time.

The example of Random Forest Classification is shown in the Fig. 6

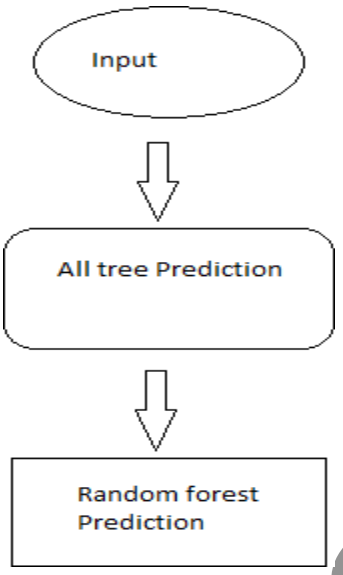


Figure 5: Random forest algorithm flow chart

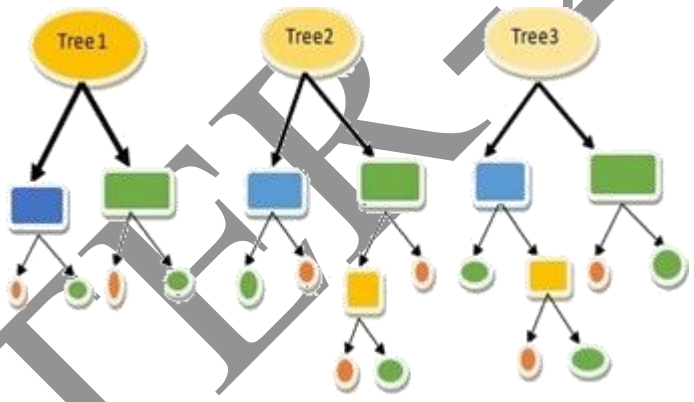


Figure 6: Example of Random Forest Classification

**Extreme Learning Machine Algorithm (ELM)**

It is also Termed as Extraordinary Learning Machine algorithm . Extraordinary Learning Machine (ELM) is a method that is utilized for a solitary concealed layer feed forward neural system that arbitrarily picks shrouded hubs and decides the yield weights[22] as in Fig. 7. This strategy just has one information layer, one shrouded layer, and one yield layer. It is somewhat not quite the same as customary Back spread calculations. ELM sets the quantity of shrouded neurons, and haphazardly loads are doled out between the information layer and concealed layers with the predisposition estimation of concealed units, at that point the yield layer is determined by utilizing the Moore Penrose pseudo opposite strategy

This calculation gives an extraordinary quick preparing velocity and incredible exactness. At the point when ELM contrasts and customary neural system methods, it saw as all the more persuading as it conquers the over fitting issues. Fig. 7 is an ELM comprising of n input layer neurons, l shrouded neurons, and m yield layer neurons. The calculation for ELM is as follow :

**Mathematical Model used for ELM**

Mathematical model for ELM algorithm with varous matrices used for computation of output matrices is shown below

Working Steps of ELM

a. Training sample is [X,Y] = {$x_i$, $y_i$} where I value of i ranges from 1,2 to Q are indicating by X and Y matrices as shown below in 2 and 3

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1Q} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nQ} \end{bmatrix} \quad (2)$$

$$Y = \begin{bmatrix} y_{11} & \cdots & y_{1Q} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nQ} \end{bmatrix} \quad (3)$$

b. Weight matrix used for ELM for the input layer is shown in 4

$$W = \begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{l1} & \cdots & w_{ln} \end{bmatrix} \quad (4)$$

c. Between the hidden layers Biases used are :

$$\beta = \begin{bmatrix} \beta_{11} & \cdots & \beta_{1n} \\ \vdots & \ddots & \vdots \\ \beta_{l1} & \cdots & \beta_{ln} \end{bmatrix} \quad (5)$$

d. The activation Function and the output matrix can be shown as 6.

$$B = [t_1 t_2 t_3 \ldots\ldots t_Q] \quad (6)$$

e. Moore-Penrose pseudo inverse of the matrix is then calculated as H as shown in 7

$$H\beta = T^t \quad (7)$$

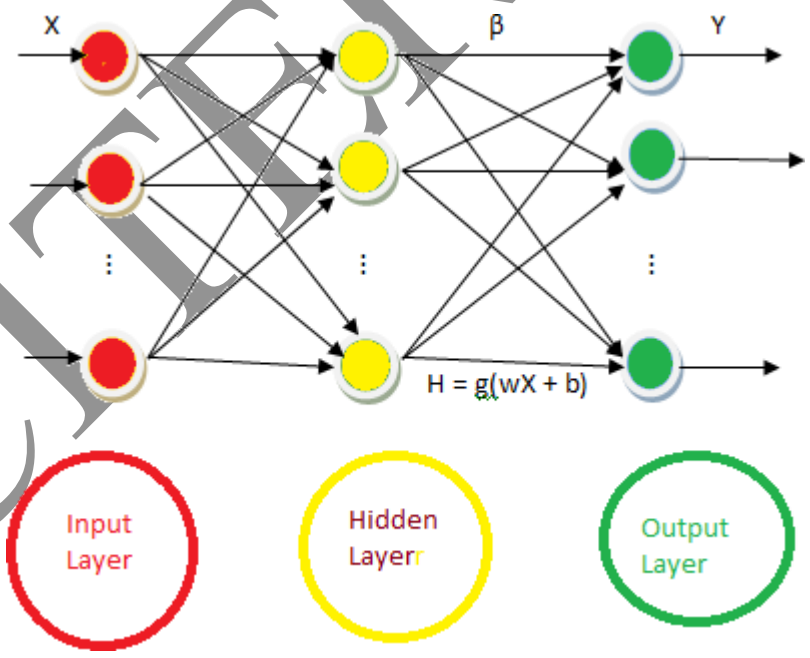


Figure 7: Extreme Learning Machine Neural Network

**Process/Methodology used**

We applied different calculations, as referenced above, on the Wisconsin Breast Cancer dataset taken from the UCI storehouse. We utilized Anaconda Spyder as a stage for coding with Python rendition 3.8. The procedure incorporates different methods like Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision tree (DT), Random Forest (RF), and Extreme Learning Machine (ELM) with measurement decrease strategies that is Principal Component Analysis (PCA).

In this paper, subsequent to perusing the dataset, the preprocessing of information is finished by parting the dataset into a preparation set and testing set. The proportion utilized for parting the dataset is 75:25. Python API Scikit-learn is utilized to perform various errands. In the wake of parting of information, include scaling would be finished. It standardizes the information inside a range with the goal that the calculation speed can increment. After standardization of information, measurements are diminished. In this paper, PCA is utilized for this reason, and the procedure had clarified underneath.

**Description of Process used**

The process of reducing the effect of independent variables on principal variables is known as dimension reduction[14]. By reducing different independent variables, data can be better viewed and utilized better. It is explained in Fig. 8 below. It comprises of the below method[1]:

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

**Step1**: the breast cancer dataset is prepared in a matrix form with all the features.

**Step2**: Features are scaled or normalized by subtracting average from each dimension to form a data, which has no meaning at all.

**Step3**: Covariance matrix is computed which describes the variance of data and

$$Cov(X, Y) = \frac{\sum_{i=1} n\ (x_i - \bar{x})\ (y_i - \bar{y})}{(n - 1)}$$

**Step4**: using above covariance, Eigenvalues and vectors are calculated which are useful in providing information about our data.

**Step5**: Eigenvalues are arranged in non-increasing order. The feature with the largest Eigenvalue becomes the principal component of the dataset.

**Step6**: A new vector forms which comprise all the principal components of the dataset.

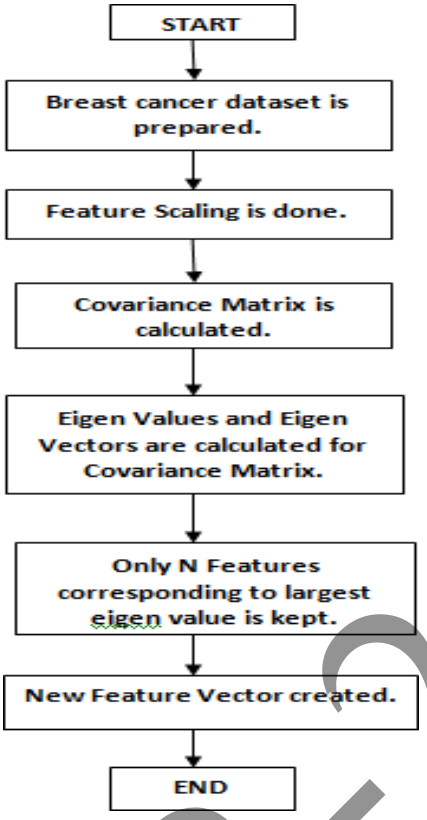


Figure 8: PCA model analysis

**Various Models Used**

It is the most energizing stage, as in this ML calculation is chosen. ML calculations are sorted into two gatherings, in particular: Supervised and Unsupervised learning calculations. In the administered calculation, the machine is prepared on marked information. Administered learning calculations are isolated into relapse and order procedures. A solo learning calculation is a technique wherein unlabeled data is given to the machine, and this data is examined with no course. In this dataset, Y is a reliant variable, which is having values either defame (1) or favorable (0)[14].

Here classification techniques are applied. In this paper, five algorithms have been chosen namely (already discussed above),

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

The Performance analysis of various algorithms are shown in table 1 below

Table 1: Performance Comparison of various models

| MODEL | Percentage accuracy of model | | Performance Timings in ms | |
|---|---|---|---|---|
| | 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 |

**Inference from table 1**

Different models are used for finding Accuracy and time of simulation and the following conclusion can be drawn trough it . For the same dataset i.e. testing and training data deployed for all the five machine learning algorithms and the same simulation tool we arrive at the following comparative results:

a) Training accuracy is found to be maximum in ELM and least in DT this indicate ELM algorithm will give more accurate result compare to the all above algorithms
b) Testing accuracy in case of ELM is Max 99% as compared to DT which is least 88%
c) Training and testing time is found same in the both ELM and DT and it is found to be least as compared to all available algorithms

These comparative results can be more visualized with the help of following bar chart for all the models on the basis of accuracy and time.

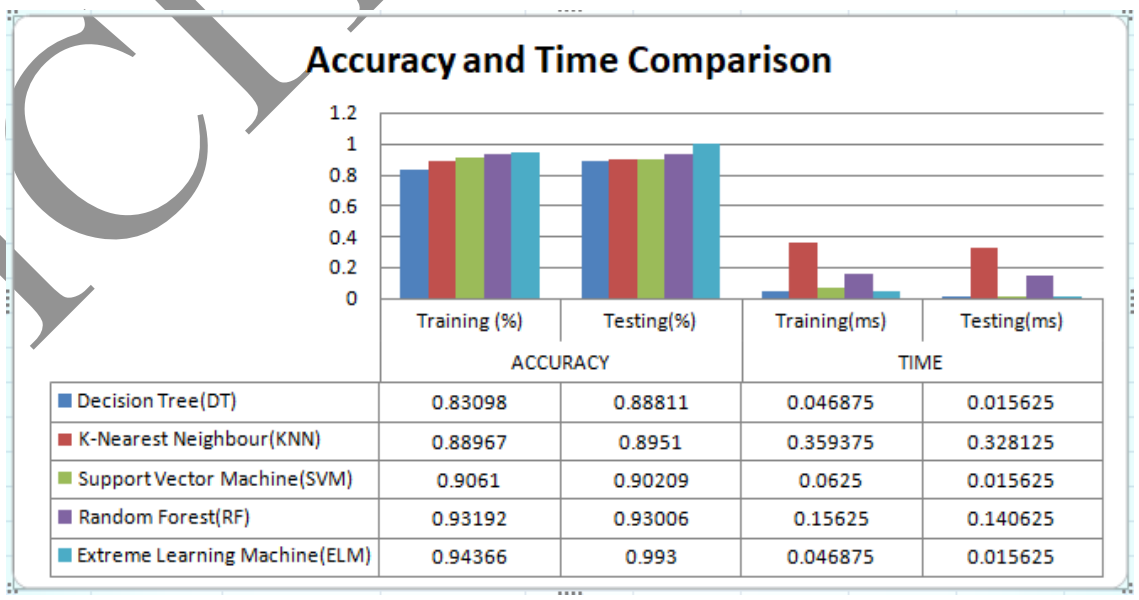


Figure 8: Accuracy and Time Comparison

**Result**

The Performance comparison of the various algorithm is shown below, which are obtained by applying different algorithms on the datasets. The accuracy and time for different models DT, KNN, RF, and ELM have been calculated and shown in Table 1. It is found that the Extreme Learning Machine (ELM) is the best among others as it gives 99.3% accuracy and least simulation time.

**Conclusion and Future Scope**

Extreme Learning Machine (ELM) will be utilized to foresee Breast malignancy with a rough 99.3% precision rate. This exactness is given with the choice instrument of PCA with this algorithm. 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. These research studies can be extended using deep learning and new development in machine learning.

**Acknowledgment**

I am profoundly grateful to Prof. Dr. Pawan Whig, Dean's research VIPS, for his valuable suggestions and helpful guidance, which help me a lot to finish this task well on time. I am also grateful to the University of the Cumberlands to allow me to pursue my Ph.D.

**RFERENCES**

1) 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.
2) 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.
3) 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
4) 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.
5) 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.
6) Xiao, B. Li, and Y. Mao, "A Multiple Hidden Layers Extreme Learning Machine Method and Its Application," Math. Probl. Eng., vol. 2017, 2017.
7) 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.
8) 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

9) 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

10) Rahul Reddy Nadikattu," Research On Data Science, Data Analytics And Big Data", International Journal of Engineering, Science and Mathematics, Vol. 9, pp.99-105, 2020.

11) Z. Nematzadeh, R. Ibrahim, and A. Selamat, "Comparative Studies on Breast Cancer Machine Learning Techniques," 2015 10th Asian Control Conf., pp. 1–6, 2015.

12) Yang Xin, Lingshuang Kong, Zhi Liu, Yuling Chen, Yanmiao Li, Hongliang Zhu, Mingcheng Gao, Haixia Hou, Chunhua Wang, "Machine Learning and Deep Learning Methods for Cybersecurity", Access IEEE, vol. 6, pp. 35365-35381, 2018.

13) R. L. Siegel, K. D. Miller and A. Jemal, "Cancer statistics 2018", CA Cancer J. Clin., vol. 68, no. 1, pp. 7-30, Jan. 2018.

14) R. Zheng, H. Zeng, S. Zhang and W. Q. Chen, "Estimates of cancer incidence and mortality in China 2013", *Chin. J. Cancer*, vol. 36, no. 1, pp. 66, Aug. 2017.

15) M. Iwase, M. Hattori, M. Sawaki, A. Yoshimura, H. Kotani, N. Gondo, et al., "Presence of small residual malignant lesions in pathologic complete response after neo-adjuvant chemotherapy in patients with breast cancer", *Breast J.*, 2019.

16) A. D. Altman, J. McGee, T. May, K. Lane, L. Lu, W. Xu, et al., "Neoadjuvant chemotherapy and chemotherapy cycle number: A national multicentre study", *Gynecol. Oncol.*, vol. 147, no. 2, pp. 257-261, Nov. 2017.

17) Rahul Reddy Nadikattu, "Implementation of New Ways of Artificial Intelligence in Sports", Journal of Xidian University, Vol. 14, Issue 5. pp. 5983-5997, 2020.

18) D. Blum, I. Liepelt-Scarfone, D. Berg, T. Gasser, C. la Fougère and M. Reimold, "Controls-based denoising a new approach for medical image analysis improves prediction of conversion to Alzheimer's disease with FDG-PET", Eur. J. Nucl. Med. Mol. Imag., vol. 46, no. 11, pp. 2370-2379, 2019.

19) M. Zhou, Y. Luo, G. Sun, G. Mai and F. Zhou, "Constraint programming based biomarker optimization", BioMed Res. Int., vol. 2015, Dec. 2014.

20) C. Xu, J. Liu, Y. Shu, Z. Wei, W. Zheng, X. Feng, et al., "An OMIC biomarker detection algorithm TriVote and its application in methylomic biomarker detection", Epigenomics, vol. 10, no. 4, pp. 335-347, Jan. 2018.

21) Chhaya Gupta, Nasib Singh Gill,' Machine Learning Techniques and Extreme Learning Machine for Early Breast Cancer Prediction", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Vol.-9 Issue-4, Feb. 2020.