International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-4, February 2020

Machine Learning Techniques and Extreme Learning Machine for Early Breast Cancer Prediction

Chhaya Gupta, Nasib Singh Gill

Abstract: ***Breast Cancer is one of the most deadly disease and most of the women are infected by this vital disease in many parts of the world. Medical tests conducted in hospitals for determining the disease are very much expensive as well as time-consuming. The problem can be resolved by diagnosing the problem in early spam of time and by providing results with more accuracy.In this paper, different machine learning and neural network algorithmhave been studied and compared to predict cancer in early stages so that life can be saved. The dataset available publically for Breast Cancer has been used. Different algorithms compared includeSupport Vector Machine Classification (SVM), K-Nearest NeighborClassification (KNN), Decision tree Classification (DT), Random Forest Classification (RF) and Extreme Learning Machine (ELM).All are compared on the basis of Accuracy and processing time are considered as the parameters for comparing analysis. The results reveal that extreme learning machine comes to be the better algorithm.***

Keywords : Decision tree classification (DT), Extreme Learning Machine (ELM),KNN classification, Random Forest (RF) classification, Support Vector Machine (SVM) classification.

I. INTRODUCTION

Breast Cancer has become the main reason behind the death of a lot of women all around the world.The main reason for the death of women by this disease is the process by which it is diagnosed[1].

The technology has become a major part of our lifestyles still we are lacking behind diagnosing this critical disease in early stages[1]. As the disease is not diagnosed in early stages, therefore, the mammography rate has been increased for a particular age group of concerned women[2].Breast Cancer is curable and life can be saved if it is diagnosed in early stages. Different causes have been diagnosed for this deathly disease including primarily hormonal imbalance, family histories, obesity, radiation therapies and many more. Many machine learning and deep learning algorithms are being applied to diagnosing this disease.

Revised Manuscript Received on February 28, 2020.
***** Correspondence Author
Chhaya Gupta*,Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, India. Email: chhayagupta.spm@gmail.com
Prof. Nasib Singh Gill, Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, India. Email: nasibsgill@gmail.com

Machine learning algorithms follow the several steps during classification problems[3] ,viz:

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

In this paper, we will be comparing various Machine Learning algorithms and a neural network (ELM) to find which algorithm gives the best result in terms of accuracy and processing time. Various machine learning algorithms discussed here are Support Vector Machine classification (SVM), K-Nearest Neighbourclassification (KNN), Decision Tree classfication (DT) and Random Forest classification (RF). The neural network discussed here is the Extreme Learning Machine (ELM).

II. LITERATURE SURVEY

This section illustrates previous work of different researchers with different breast cancer datasets.In [4] various machine learning classifiers like SVM classifier, Random Forest, KNN classifier and Decision Tree are compared with feature selection algorithm and results showed that Random Forest gave the best results with 93% accuracy. In [5] researchers compared different ML algorithms namely Naive Bayes, SVM, Decision TreeJ48, RandomForest, Bagging, AdaBoostandLogistic Regression over Wisconsin Breast Cancer dataset with PCA and results showed that Random forest gave the best results.

In [6] author compared SVM, KNN, Artificial Neural Network and Naïve Bayes are compared and results proved SVM gave the highest accuracy and after that neural network. ANN, SVM and Decision tree are compared in [7] and SVM was the best among all the machine learning methods with highest accuracy and lowest error rate. In [8] authors explored KNN performance with WBC(wisconsin Breast Cancer) dataset and WDBC(wisconsin diagnostic breast cancer) dataset with three iterartions, in which the initial iteration is without feature selection, second with feature selection and KNN and the last iteration consist of Chi-sqaure feature selection, all these help in getting optimal value of K and also the chi-square base feature selection with KNN classifier gives the best accuracy results.

In [9] the researchers compared single layer neural network with two layer neural networks and found that single layer neural network gave the highest accuracy of 86.5%. In [10], the dataset was taken from Iranian centre of breast cancer and compared decision tree, support vector machine and artificial neural network.

Support vector machine was proven to be the best followed by an ANN and then the DT classification model.In [11], 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 decision tree classification model, Naïve Bayes model, neural network and support vector machine with different kernels. Results showed that the neural network was best for WBC dataset and support vector machine with radial basis function (RBF) and was best for WPBC dataset.

In [12], an ANN (Artificial neural network) with Principal Component Analysis (PCA) is used.In [13], WPBC dataset is used for making comparison of different ML(Machine Learning) algorithms. The result showed that SVM, DT were among best predictors.In [14], multi-layer perceptron with backpropagationand support vector machine were used for classification of dataset. Support Vector machine was found to be the best result giving algorithm.

In [15], a signal-to-noise ratio technique was used with different classification models which are k-nearest neighbour, SVM and PNN that is a probabilistic neural network. SVM with RBF kernel was giving the best result.In [16], 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 [17], a new approach provided which was based on the neural network with feed-forward BP algorithm. A 7 hidden unit neural network was used to obtain the results.

In [18], relevance vector machine (RVM) was compared with other machine learning techniques. Linear Discriminant Analysis (LDA) method was utilised for dimension reduction. RVM gave the best results in their experiment on WBC dataset.

III. MACHINE LEARNING CLASSIFICATION MODELS USED

This section presents the machine learning classification models used in the present study.

A. Support Vector Machine Classification

This technique uses a maximal margin hyperplane to classify the dataset into different classes[3]. The technique is used in many fields like disease recognition, handwriting recognition, speech recognition, and many other fields of pattern recognition. This technique increases the gap between the classes which it creates as in Fig.1.

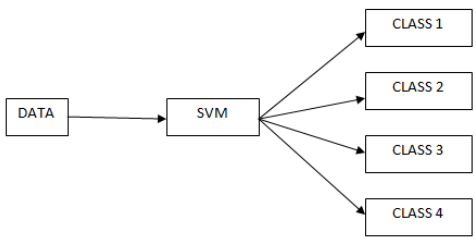

Fig. 1.Different Classes via SVM

An SVM model which uses kernel as a “Sigmoid” kernel could be considered as a neural network with two. SVM can be used with different kernels like “linear”, “poly”, “radial basis function (RBF)”etc[19].SVM is a algorithm that can classify the dataset into different classes efficiently[20]. In this, each data point is plotted in an n-dimensional space and then a hyperplane or line is determined by classification. Fig.2 beautifully distinguishes the two classes as the points in green circle class and other data in red circle class. As SVM is a multi-dimensional space therefore, each point becomes a vector here.

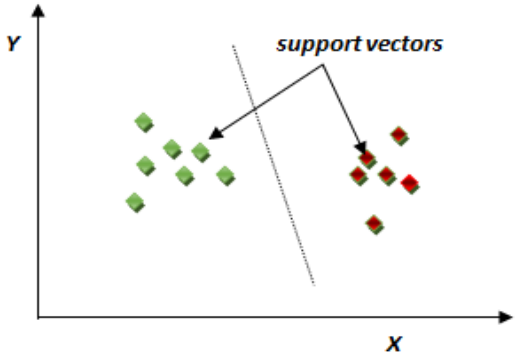

Fig. 2.Support Vector Machine Classification showing different support vectors

B. K-Nearest Neighbor Classification

This is a very effective and simple classification method which can be implemented very easily. The ideology is to find K similar samples from feature samples[21].It is measured by finding the distance between various eigenvalues which we call as Euclidean distance[21] as in Fig.3. The number of K neighbours is predetermined firstly; default value taken for K is usually 5. Then, K nearest neighbours of a new data point is taken. Among these K neighbours, data points are counted in each category and the new data point is assigned to the category for which you counted the most neighbours.

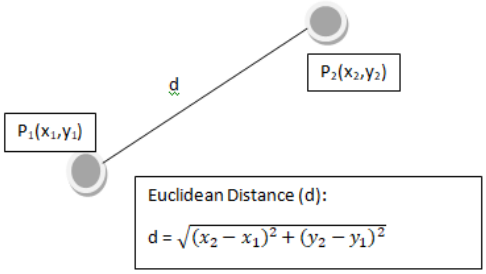

Fig. 3.K-Nearest neighbour Classification classifying Euclidean distance

The Euclidean distance is calculated as below:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

C. Decision Tree Classification

Decision Tree is a type of flow chart in which dataset is split in a manner so that every split region has a maximum number of data points as in Fig. 4.

These trees partition the inputs into cells and each cell is considered as one class[22].Partition is done according to the tests performed on the dataset. Each node gives birth to two roads either a true circumstance or a false one. It is a model that is similar to a tree. Tree leaves represent partitioned datasets. In this algorithm the best data point is root. In this algorithm, we start with root for describing the class of a record.

In this data point's attributes are compared with internal nodes of the decision tree until we reach the leaf node with predicted class.

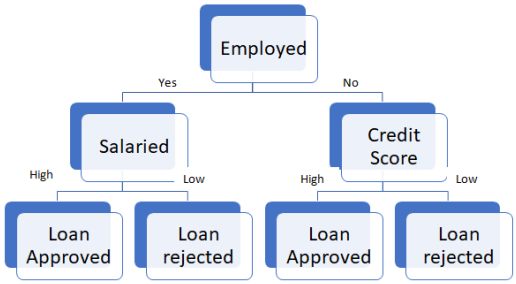

Fig. 4.Decision Tree Classification

D. Random Forest Classification

Random forest is a version of ensemble learning and it follows a bagging technique as in Fig. 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 is build for these data points. Samples are 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 best results with accuracy and in very less processing time.

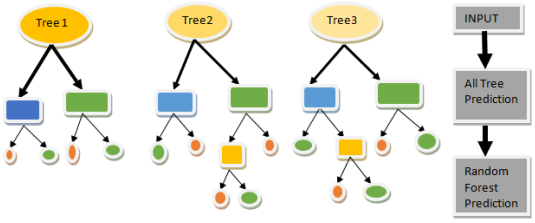

Fig. 5.Random Forest Classification with base model as Decision Tree

E. Extreme Learning Machine(ELM)

It is a technique which is used as single hidden layer feedforward neural network which chooses hidden nodes randomly and determines output weights[23] as in Fig.6. This method has one input layer which consists of input nodes, one hidden layer consisting of hidden nodes and single output layer that provides output. It is a bit different from traditional Back-Propagation algorithms.This algorithm sets number of hidden neurons and weights are assigned randomly between the input and hidden layers with a bias value, then the output layer is calculated by using Moore Penrose pseudoinverse method[24].This algorithm provides an exceptional fast processing speed and great accuracy. When ELM is compared with traditional neural network techniques it is found to be more convincing as it overcomes the overfitting problems[25]. Fig. 6 is an ELM consisting of n-input layer nodes, l hidden nodes and m output layer nodes. The algorithm for ELM is as follows:

Step1: Training sample is [X, Y] = {xi, yi} (i= 1,2,……Q) and X and Y matrices can be described as below with n = dimension of input matrix and m = dimension of output matrix

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

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

Step2: ELM then assign weights matrix for input layer as,

$$W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{l1} & w_{l2} & \cdots & w_{ln} \end{bmatrix} \quad (3)$$

Step3: Biases are assumed as,

$$\beta = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{l1} & \beta_{l2} & \cdots & \beta_{lm} \end{bmatrix} \quad (4)$$

Step4: Bias is randomly set for hidden layer neurons as,

$$B = [b_1 b_2 \ldots b_n]\ T \quad (5)$$

Step5: An activation function is chosen and according to Fig. 6 the output matrix can be expressed as,

$$T = [t_1 t_2 \ldots t_Q]m*Q \quad (6)$$

Where column vectors of T are as follows:

$$tj = \begin{bmatrix} t_{1j} \\ t_{2j} \\ \vdots \\ t_{mj} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{l} \beta_{i1} g\left(w_i x_j + b_i\right) \\ \sum_{i=1}^{l} \beta_{i2} g\left(w_i x_j + b_i\right) \\ \vdots \\ \sum_{i=1}^{l} \beta_{im} g\left(w_i x_j + b_i\right) \end{bmatrix} \quad (7)$$

(j = 1,2,3,….., Q)

Step6: Calculate the Moore-Penrose pseudoinverse of the matrix.

Step7: Calculate the output weight matrix H as,

$$H\beta = T' \quad (8)$$

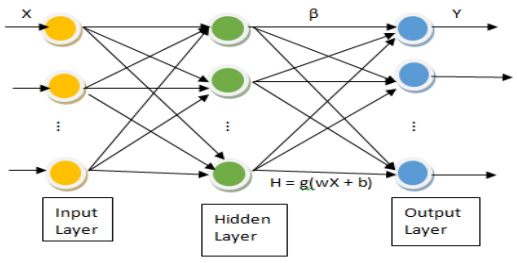

Fig. 6.ELM Neural Network

IV. METHODOLOGY USED

Above mentioned algorithms have been applied on Wisconsin Breast Cancer (WBC) dataset available publically atUCI repository. Anaconda Spyder as a platform has been used 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 technique that is Principal Component Analysis (PCA).In this paper, after reading the dataset, preprocessing of data is done by splitting the dataset into two sets namely training and testing. Ratio used for splitting the dataset is 75:25. Python API Scikit-learn is used to perform different tasks. After data is split, feature scaling is done. It is helpful in normalising the data within a range so that the algorithm speed can be increased. After normalization of data, dimensions are reduced. In this paper PCA is used for this purpose and the process is explained below.

A. Dimension Reduction

The process of reducing independent variables to principal variables is known as dimension reduction[20]. This process reduces the dimensions of the dataset so that data can be viewed better and can be utilised better. It is explained in Fig.7 below.

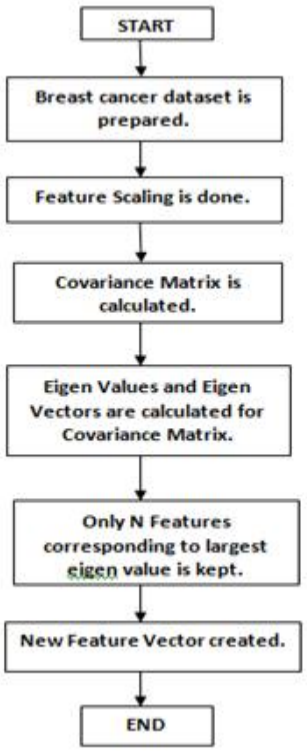

Fig.7.Principal Component Analysis Algorithm

B. Model Selection

It is the most interesting phase as in this machine learning algorithm is selected. Machine learning algorithms are categorised into two groups namely: Supervised and Unsupervised learning algorithms. In the supervised algorithm, the machine is trained on labelled data. Supervised learning algorithms are divided into regression and classification techniques. An unsupervised learning algorithm is a method in which unlabelled 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)[20]. Here classification techniques are applied. This paper compares five algorithms which are:

- K-Nearest Neighbour classification technique
- Support Vector Machine classification technique
- Decision Tree classification technique
- Random Forest classification technique
- Extreme Learning Machine neural netwrok

V. EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

Table I provides results of the experiment conductedon dataset by using various different techniques. Different techniques used here are compared on the various aspects like training and testing accuracies and training time taken on the dataset as well as testing time taken on dataset. The results clearly show that Extreme Learning Machine is the most best among others as it is giving 99% accuracy and in very less time.

Table I Performance Comparison

MODEL	ACCURACY		TIME	
	Training (%)	Testing(%)	Training(ms)	Testing(ms)
Decision Tree(DT)	0.83098	0.88811	0.046875	0.015625
K-Nearest Neighbour(KNN)	0.88967	0.8951	0.359375	0.328125
Support Vector Machine(SVM)	0.9061	0.90209	0.0625	0.015625
Random Forest(RF)	0.93192	0.93006	0.15625	0.140625
Extreme Learning Machine(ELM)	0.94366	0.993	0.046875	0.015625

[a] Table I provides experimental results.

Fig.8 shows bar chart comparison for all the models used in this paper.

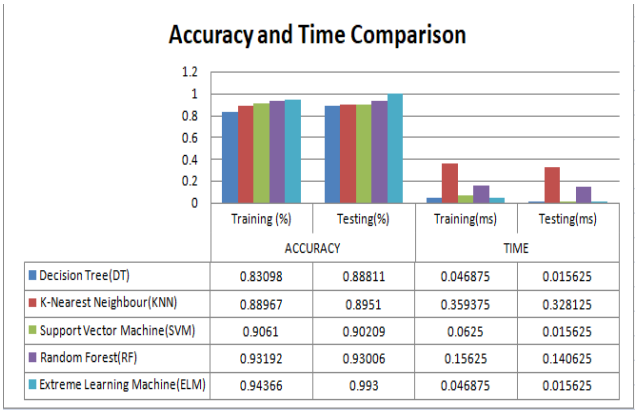

Fig.8. Accuracy and Time Comparison among various models used

VI. CONCLUSION

Extreme Learning Machine (ELM) can be used to predict Breast cancer with an approximate 99% accuracy rate after 50 epochs. This accuracy is provided with the feature selection mechanism of PCA along with ELM. This mechanism can be used in future to identify the benign and malignant cells in early stages and can be implemented as an application in mammography techniques. There is always room for improvement. The performance may certainly be enhanced by researchers and hence provides scope of further advancement.

REFERENCES

1. Y. Shen et al., "Globally-Aware Multiple Instance Classifier for Breast Cancer Screening," pp. 1–11, 2019.
2. N. E. Benzebouchi, N. Azizi, and K. Ayadi, Computational Intelligence in Data Mining, vol. 711. Springer Singapore, 2019.
3. A. 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.
4. M. F. Aljunid and D. H. Manjaiah, Data Management, Analytics and Innovation, vol. 808. 2019.
5. K. Goyal, P. Sodhi, P. Aggarwal, and M. Kumar, Comparative Analysis ofMachine Learning Algorithms for Breast Cancer Prognosis, vol. 46. Springer Singapore, 2019.
6. G. Ganapathy, N. Sivakumaran, M. Punniyamoorthy, R. Surendheran, and S. Thokala, "Comparative study of machine learning techniques for breast cancer identification/diagnosis," Int. J. Enterp. Netw. Manag., vol. 10, no. 1, pp. 44–63, 2019.
7. E. A. Bayrak, P. Kirci, and T. Ensari, "Comparison of machine learning methods for breast cancer diagnosis," 2019 Sci. Meet. Electr. Biomed. Eng. Comput. Sci. EBBT 2019, pp. 4–6, 2019.
8. Z. Mushtaq, A. Yaqub, S. Sani, and A. Khalid, "Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets," J. Chinese Inst. Eng., vol. 00, no. 00, pp. 1–13, 2019.
9. A. Osmanović, S. Halilović, L. A. Ilah, A. Fojnica, and Z. Gromilić, "Machine Learning Techniques for Classification of Breast Cancer Ahmed," pp. 731–735, 2019.
10. A. 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.
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. 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.
13. 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.
14. 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.
15. A. Osareh and B. Shadgar, "Machine learning techniques to diagnose breast cancer," 2010 5th Int. Symp. Heal. Informatics Bioinformatics, HIBIT 2010, pp. 114–120, 2010.
16. 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.
17. 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.
18. 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.
19. 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.
20. 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.
21. 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.
22. N. J., "Data Mining Techniques:a Survey Paper," Int. J. Res. Eng. Technol., vol. 02, no. 11, pp. 116–119, 2013.
23. 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.
24. D. Xiao, B. Li, and Y. Mao, "A Multiple Hidden Layers Extreme Learning Machine Method and Its Application," Math. Probl. Eng., vol. 2017, 2017.
25. 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.

AUTHORS PROFILE

Chhaya Gupta is a research scholar pursuing PhD from Maharishi Dayanand University, Rohtak. She has completed her MCA from Guru Gobind Singh Indraprastha University, Delhi in 2012. She completed her graduation from Delhi University, Delhi in 2009 in the stream of computer Science. She is a Gold Medalist from GGSIPU, Delhi. She qualified her UGC NET exam in January, 2019.

Prof. Nasib Singh Gill is Head of Department of Computer Science and Applications at Maharishi Dyanand University, Rohtak. He is Director at University Computer Centre, MDU, Rohtak. He is Post Doctorate from Brunel University, UK. He has a otal experience of 22 years in the field of academica and research. he has published 146 research papers in various renowned international and national journals. He is a recipient of Commonwealth Fellowship Award for the year 2001 - 2002. He received best paper award by Computer Society of India in 1994 for the paper titled "A New Program Complexity Measure". He has delivered various talks which have been broadcasted from All India radio.