ADVANCE CATALOGUING METHOD FOR THE MICROARRAY BREAST CANCER DETECTION

[1]AnuradhaTaluja , [2] Harish Kumar Taluja

[1]Assistant Professor, Ajay Kumar Garg Engineering College, Ghaziabad, UP, India

[2] Professor, Noida International University, Noida, UP, Ghaziabad, India

[1]talujaanuradha@akgec.ac.in, [2]harishtaluja@gmail.com

Abstract: **Cancer has been considered as a heterogeneous disease available in various forms. There is no such technique is available to predict the cancer stage, so early diagnosis and forecast of a cancer type and stage have become a today's requirement. This can be done by using data mining technique The technique can be defined as a technology by which valuable knowledge and information can be fetched out from the massive volume of data. The big patterns can be explored and analyzed using statistical and Artificial Intelligence in big databases. The technology of data mining is gaining a lot of popularity in healthcare sector. Prediction can be defined as a statement about future event on the basis of present situation. The major intend of this work is to predict the microarray cancer using machine learning (ML) algorithms. Different phases are comprised in the prediction of microarray cancer. This research makes the implementation of voting-based classification algorithm. The suggested algorithm assists in optimizing the performance up to 2% while predicting the microarray cancer.**

Keywords: **Cancer, DIP, ML,KDD, Big Data**

I. INTRODUCTION

As the most dominant cancer in women, breast cancer has always had high incidence and mortality rates. As per the most recent cancer insights, BC alone is predicted to be accounted for 25% of all new cancer diagnosis and 15% of all cancer fatalities among ladies around the world [1]. Researchers have known about the threats of BC from right off the bat, in this manner much early exploration has just been executed in the treatment of BC. Because of the endeavours of scientists and early recognition strategies, the death rate has demonstrated a consistent and declining pattern over the previous many years. As of late our capacities of both gathering and creating information have been expanding quickly. The far-reaching utilization of bar codes for most of business items, the computerization of numerous legislature and business exchanges, and the advances in information assortment devices have given us colossal measure of information.

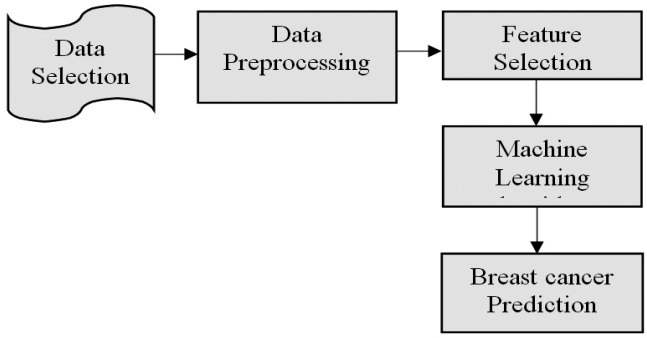

Figure 1: General process of Breast cancer prediction based on Machine Learning

This immense development in data and databases has created an earnest requirement for new strategies and devices that can keenly and naturally change the prepared information into valuable data and information. Thus, data mining has become an exploration zone with expanding significance. Data mining, which is additionally alluded to as KDD (Knowledge Discovery in Databases), implies a cycle of nontrivial extraction of certain, already obscure and conceivably helpful data, (for example, information rules, limitations, consistencies) from data stored in databases.

II. MICROARRAY CANCER PREDICTION

Breast cancer is one of the most well-known dangerous tumours in ladies around the world, and it remains the main source of malignancy demise among females in developing nations. As of late, the frequency and mortality of breast cancer have expanded step by step, which genuinely undermines the lives and wellbeing of ladies and causes incredible monetary, social and family issues? In spite of the fact that the predominance of breast cancer shows the tendency to be in more youthful females, postmenopausal ladies additionally have the danger of breast cancer. Therefore, investigating the qualities of cancer malignancy in postmenopausal ladies, and discovering important data from various information to give clinical analysis and treatment to logical dynamic and clinical exploration are of incredible importance.

Artificial intelligence and, specifically, machine learning models have an obvious history in malignancy research and useful execution [11]. The utilization of various ML models in malignancy research gives immense space to different applications. Artificial Neural Networks (ANNs) and Decision trees (DTs) have been utilized in disease prediction and detection for almost 30 years. Various models dependent on Support Vector Machine (SVM)applied to cancer prediction issues have been utilized for about a few decades. Different models for forecast of malignancy advancement and result have likewise been utilized for a few examinations.

Today, not exactly a portion of information science and bioinformatics techniques are utilized by ML-driven models with a wide scope of uses, from diagnostics to expectation and forecast in malignancy. All this examination contemplates are worried about utilizing ML strategies to recognize, classify, identify, or differentiate tumours and different malignancies, just as to foresee disease growth. Breast cancer prediction works dependent on machine learning models possess a considerable part of the contemporary examination in this domain.

a. Data Selection: In this stage, a dataset is selected for extracting information. The datasets can be openly accessible (e.g., on the web) or they may result from a joint effort among foundations and exploration groups, not accessible for the overall population. Gathered data for the most part incorporates demographic features (age, stature, weight record), physiological richness factors (time of menarche, time of menopause, age of the first pregnancy, post-menopausal hormone discharge level), illness history and hereditary variables (history of non-cancerous breast cancer sickness, family background of breast cancer), social propensities (smoking, drinking) and so forth

b. Pre-processing: Pre-processing assignments are performed to diminish noise and increment the consistency of information [13]. The pre-processing steps generally tended to in various researches are data standardization/normalization, and missing data management. Two basic methods of information pre-processing are data cleaning, standardization (Min-Max change), and normalization (Z-Score conversion). All these tasks have been explained below:

- Data cleaning: Data cleaning schedules work to "clean" the data by filling in missing values, smoothing noisy data, distinguishing or eliminating anomalies, and settling irregularities. In the event that clients accept the information are messy, they are probably not going to believe the aftereffects of any data mining that has been applied to it. Besides, filthy data can create turmoil for the mining process, bringing about untrustworthy result.
- Normalization: Normalization alludes to the feature scaling between its minimal and most extreme qualities, while standardization rescales the attributes with the goal that they keep a standard typical distribution (zero mean and unitary standard deviation). The goal of standardization/ normalization is to generate attributes with various scales and scopes of measurement (e.g., age, haemoglobin levels) practically identical, so that none has more impact than the others on classification task. Missing Data (MD) can result from a tremendous assortment of functions and speaks to a typical test in the healthcare domain.

c. Feature Selection: In data mining domain, a classification technique can profit altogether from utilizing just significant information regarding learning performance and learned outcomes, for example, improved conceivability [14]. Feature selection is a broadly applied procedure for discovering applicable information by eliminating irrelevant and superfluous information.This technique recognizes a small number most significant features and aids in result prediction. They are defined as:

- Filter method: The filter methods select the features on the basis of scores in different statistical correlations.
- Wrapper method: These methods perform feature selection using a greedy approach. These techniques evaluate all possible combination and generate the outcome for Machine learning.
- Embedded Scheme: The embedded approach merges the benefits of both abovementioned methods. The inducer possesses its specific FSA (either explicit or implicit).

d. Machine Leaning algorithms: Throughout the years, many Machine Learning (ML) algorithms have been utilized to predict the occurrence of breast cancer. Some prominent machine learning algorithms are:

i. K-Nearest Neighbours

K Nearest Neighbor algorithm is used for clustering and utilized in pattern recognition. It is generally utilized in prediction analysis. The features which may be different to large extent may have adequate impact on the interval amid data instances. In order to do classification, number is characterized by neighbours of K that is the most frequent among the K training sample. In this case, the calculation discovers K adjacent neighbours of the original data pattern. if all the data points occur in metric space, a major challenge is to measure distance [16]. If the no. of neighbours is represented by N in this algorithm, then N samples are measured with the help of a distance metric given below:

Minkowski Distance: $Dist(x,y) = (\sum_{i=1}^{n}|x_{i-}y_i|^p)^{\frac{1}{p}}$

$Dist(x,y) = (\sum_{i=1}^{n}|x_{i-}y_i|^p)^{\frac{1}{p}}$

In this expression if p=1, then it is known as Manhattan distance, p=2, represents Euclidean distance, and p=∞ denotes Chebyshev distance. Amidst various selections, Euclidean distance is universally implemented. Among these K neighbours, the calculation verifies the number of data relevant to every class, and then, it relegates the fresh data point to the classification which frames the more considerable share.

ii. Decision Trees

In Decision Trees instances are classified after sorting on the basis of feature values. A feature is illustrated in an instance for the classification using every node in a DT and a value which can be assumed by the node is signified using each branch[7][9][11]. The evaluation of Gain Ration for attribute A is presented as:

$$GainRatio(A) = \frac{Gain(A)}{Split\ Info(A)}$$

$$Gain(A) = info(D) - Info_A(D)$$

Where D is the training dataset

$$Info(D) = -\sum_{i=1}^{n} p(ci) \log_2 P(C_i)$$

iii. Support Vector Machine

The idea of SVM, which was proposed by Vapnik based on the statistical learning hypothesis, has become a fundamental part in ML strategies [18]. SVM was at first created for twofold classification, however it tends to be productively stretched out for multiclass issues with boundless use in fields of pattern recognition, handwriting recognition, text classification, and so forth. The vital component of a SVM classifier is to discover an improved decision boundary that speaks to the biggest partition (most extreme edge) amid the classes.

iv. Artificial Neural Networks (ANNs)

Artificial Neural Networks (ANN) can be characterized as a reasoning model having configuration similar to the human brain [19]. In the course of recent many years, ANNs have been utilized progressively by an ever-increasing number of analysts, and become a functioning exploration territory. Figure 2 represents a common ANN model consisting a chain of layers, i.e. input, hidden and output layers.

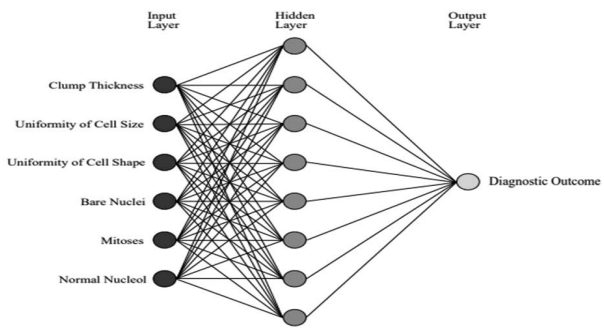

Figure 2: A typical ANN model for BC classification

As shown by the figure, layers are made out of interconnected neurons which contain an activation function for nonlinear change to fortify the nonlinear articulation capacity.The output layer generates the outcomes of classification [20]. In any case, contingent upon the issues, the way toward training an ANN may include long causal sequence of computational phases.

e. Prediction: In this step, the mapping of selected features is carried out onto the training model for classifying the given features so that the liver disease can be predicted. In order to generate predictions, a specialist doctor labels the gathered breast cancer dataset.

III. LITERATURE REVIEW

Gopal, et.al discussed that the main concern of medical community was to predict the breast cancer [1]. This research focused on predicting the breast cancer accurately. The most effectual ensemble ML model was constructed using the breast cancer Coimbra dataset that was extracted from UCI. Cath Tee, et.al analysed that the breast cancer was a disease due to which a great number deaths were occurred in every year [2]. Various algorithms were available in order to classify and predict the breast cancer such as SVM, DT, NB and KNN. Chintan, et.al described that the major cause of breast cancer was the division of abnormal cell in the breast itself due to which benign or malignant cancer was developed [3].

IV. METHODOLOGY

The smooth functioning of this part is very essential for the healthy lifestyle. If any kind of disease occurs in this organ, the other body parts are also disturbed. The researchers have often utilized the computer aided data which is extracted from the enormous databases. There are different stages included while predicting the breast cancer. These stages are mentioned as:

A. Data Acquisition: This phase includes the collection of data from distinct clinical organizations in order to carry out the experiments.
B. Data pre-processing: The entirety is accomplished and the data is analysed to deploy the Machine Learning methods and the pre-processing is performed on the data.
C. Feature selection: This phase makes the deployment of a subset having extremely unique attributes for detecting the microarray cancer. These selective attributes are related to the existing class of attributes. The Random Forest (RF) algorithm is applied in the introduced method for choosing the attributes. RF (Random Forest) is a nonparametric technique using which an ensemble model of DT (decision tree) is constructed from the random subsets of attributes and bagged samples of the training data. The performance of RF is found well even

in the presence of noise in predictive attributes. Therefore, the tree developed from this kind of bagged subspace of attributes provides lower accuracy to predict the attribute that has impact on final prediction of the Random Forest.

A training dataset is presented as

$$\mathbb{L} = \left\{(X_i, Y_i)_{i=1}^{N} \middle| X_i \in RM, Y \in \{1,2,\dots,c\}\right\}$$

$$\mathbb{L} = \left\{(X_i, Y_i)_{i=1}^{N} \middle| X_i \in RM, Y \in \{1,2,\dots,c\}\right\},$$

xi, denotes the attributes, *Y* is used to show a class response feature, *N* represents the number of training samples and the number of attributes is denoted with *M*. A RF model is defined in Algorithm 1, let *Yk* be the prediction value of tree *Tk* given input *X*. The prediction of random forest with *K* trees is expressed as:

$$\hat{Y} = majority\ vote\{\hat{Y}^k\}_1^K$$

Every tree is grown from a bagged sample set; thus, it is grown using only two-thirds of the samples in L, that is known as in-bag samples. About one-third of the samples is not utilized and these samples are known as out-of-bag (OOB) samples that assists in estimating the prediction error.

The OOB predicted value is $\hat{Y}^{OOB} = (1/\|\mathcal{O}_{i'}\|)\sum_{k\in\mathcal{O}_{i'}}\hat{Y}^k$ $\hat{Y}^{OOB} = (1/\|\mathcal{O}_{i'}\|)\sum_{k\in\mathcal{O}_{i'}}\hat{Y}^k$ in which $\mathcal{O}_{i'} = \mathbb{L}\backslash\mathcal{O}_i, i$ $\mathcal{O}_{i'} = \mathbb{L}\backslash\mathcal{O}_i, i$ and i' denote in-bag and out-of-bag sampled indices, $\|\mathcal{O}_{i'}\|$ is the size of OOBsubdataset, and the OOB prediction error is

$$\widehat{ERR}^{OOB} = \frac{1}{N_{OOB}}\sum_{i=1}^{N_{OOB}} \wp\left(Y, \hat{Y}^{OOB}\right)$$

In this $\wp(\cdot)$defines an error function and N_{OOB} is used to represent the size of OOB samples.

Measurement of Feature Importance Score from an RF: A permutation method is introduced by Breiman for quantifying the importance of attributes in the prediction that is known as an out-of-bag importance score. The other kind of feature importance measure can be obtained during the development of RF. At each node *t* in a decision tree, the minimization of node impurity Δ *R(t)* determines the separation. The gini index is defined by node impurity *R(t)*. When a sub-dataset in node *t* containing the samples from *c* classes, gini (*t*) is described as:

$$R(t) = 1 - \sum_{j=1}^{c} \hat{p}_j^2$$

In which $\hat{p}^2 j$ denotes the relative frequency of class *j* in *t*. Gini (*t*) is diminished, in case, the classes in *t* are skewed. When *t* is divided into two child nodes t_1 and t_2 with sample sizes $N_1(t)$ and $N_2(t)$, the gini index of the split data is expressed as:

$$Ginin_{split}(t) = \frac{N_1(t)}{N(t)} Gini(t_1) + \frac{N_2(t)}{N(t)} Gini(t_2)$$

The feature providing smallest $Gini_{split}(t)$ is selected for splitting the node. The importance score of features X_j in a single decision tree T_k is defined as:

$$IS_k(X_j) = \sum_{t\in T_k} \Delta R(t)$$

and *K* trees are calculated in a random forest using it and this is defined as:

$$IS(X_j) = \frac{I}{K}\sum_{k=1}^{K} IS_k(X_j)$$

A RF (random forest) is implemented in-bag samples with the objective of generating a kind of importance measure which is recognized as an in-bag importance score. This is the main difference amid the in-bag importance score and an out-of-bag measure whose creation is done with the decrease of the prediction error.

D. Classification: The mapping of chosen attributes is done for the training model to classify the provided attributes so that the microarray cancer can be predicted easily. The kind of microarray cancer is represented by each separate class. The linear Regression (LR) algorithm is implemented to perform this process. Each observation in linear regression depends on two values, one is the dependent variable and the second is the independent variable. The equation below shows how y is related to x known as regression.

$$y = \beta_0 + \beta_1 + \varepsilon$$

or equivalently

$$E(y) = \beta_0 + \beta_1 x$$

Here, ε is the error term of linear regression. The error term here uses to account the variability between both *x* and *y*, β_0 represents y-intercept, β_1 represents slope.

To get the best fit implies the difference between the actual values and predicted values should be minimum, so this minimization problem can be represented as:

$$minimize \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

$$g = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

Here, g is called a cost function, which is the root mean square of the predicted value of $y(pred_i)$ and actual y (y_i), n is the total number of data points. The microarray cancer denotes that the person have probability of occurrence of microarray cancer. The normal is utilized for the person without any possibility of microarray cancer.

V. RESULTS

In this work, the task of microarray cancer prediction has been accomplished by applying an openly available dataset called Cleveland. There are fourteen attributes included in this dataset. This work applies, and compares several classification models for predicting microarray cancer. Some of these classification models include DT, MLP (Multilayer perceptron), NB, an ensemble classifier combining random forest, and naïve bayes classifiers. The microarray cancer prediction has various phases which are dataset input, feature extraction and classification.

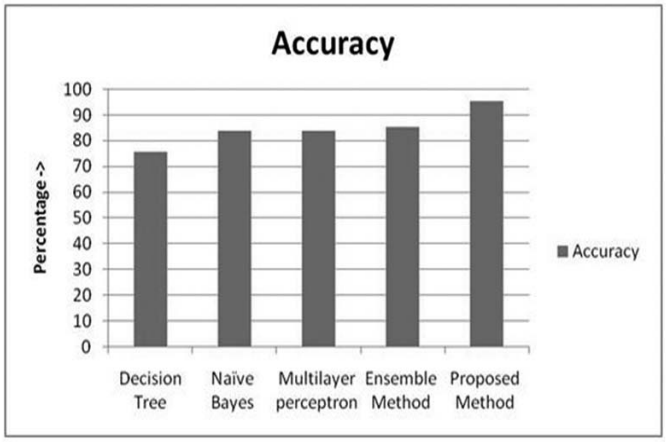

Figure 4: Accuracy Analysis

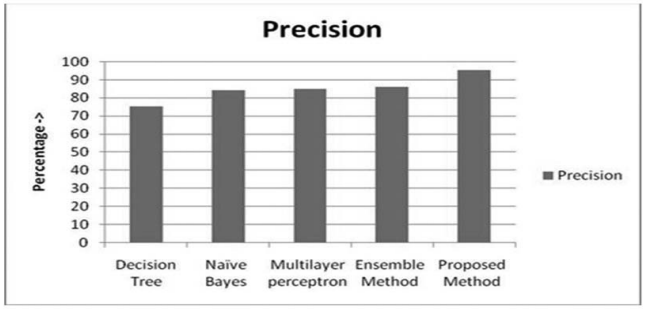

Figure 5: Precision analysis

Figure 4 depicts the accuracy-based comparison of different classification models like DT, NB, multilayer perceptron, ensemble, and proposed models. The results of the analysis depict that the introduced model outperforms other models by obtaining an accuracy rate of 95%, and proves best.

Figure 5 depicts the precision-based comparison of different classification models like DT, NB, multilayer perceptron, ensemble, and proposed models. The results of the analysis depict that the introduced model outperforms other models by obtaining a precision rate of 95%, and proves best.

REFERENCES

[1] Gopal K. Dhondalay, Dong L. Tong, Graham R. Ball, "Estrogen receptor status prediction for breast cancer using artificial neural network", International Conference on Machine Learning and Cybernetics, Volume: 2, Issue: 24, PP: 256-264, 2011

[2] Ir Cath Tee, Ali H. Gazala, "A novel breast cancer prediction system", International Symposium on Innovations in Intelligent Systems and Applications, Volume: 67, Issue: 9, PP: 984-992, 2011

[3] Chintan Shah, Anjali G. Jivani, "Comparison of data mining classification algorithms for breast cancer prediction", Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Volume 14, Issue: 30, PP: 3980-3987, 2013

[4] Hon-Yi Shi, Jinn-Tsong Tsai, Wen-Hsien Ho, Ming-Feng Hou, "Application of artificial neural networks for the prediction of quality of life in breast cancer patients", SICE Annual Conference, Volume: 53, Issue: 10, PP: 734-742, 2011

[5] Xiaoyi Xu, Ya Zhang, Liang Zou, Minghui Wang, Ao Li, "A gene signature for breast cancer prognosis using support vector machine", 5th International Conference on BioMedical Engineering and Informatics, Volume: 4, Issue: 28, PP: 2712-2720, 2012

[6] Glenn D Francis, Sandra R Stein, Glenn D Francis, "Prediction of histologic grade in breast cancer using an artificial neural network", The 2012 International Joint Conference on Neural Networks (IJCNN), Volume: 12, Issue: 2, PP: 854-861, 2012

[7] Daphne Teck Ching Lai, Jonathan M. Garibaldi, "Improving semi-supervised fuzzy c-means classification of Breast Cancer data using feature selection", IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Volume: 29, Issue: 23, PP: 1092-1100, 2013

[8] Aida Ali, SitiManyamShamsuddin, Anca L. Ralescu, "Hybrid intelligent systems in survival prediction of breast cancer", 12th International Conference on Hybrid Intelligent Systems (HIS), Volume: 80, Issue: 4, PP: 787-795, 2012

[9] Gul ShairaBanu, AmjathFareeth, NisarHundewale, "Prediction of breast cancer in mammagram image using support vector machine and fuzzy C-means", Proceedings of 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics, Volume: 27, Issue: 56, PP: 639-647, 2012

[10] Peter Adebayo Idowu, KehindeOladipo Williams, Jeremiah AdemolaBalogun and AdeniranIsholaOluwaranti, "Breast Cancer Risk Prediction Using Data Mining Classification Techniques", Transactions on Networks and Communications, Volume: 3, Issue: 42, PP: 187-194, 2015

[11] R. Preetha, S. Vinila Jinny, "A Research on Breast Cancer Prediction using Data Mining Techniques", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume: 8, Issue:11, PP: 693-701, 2019

[12] Dr. C Nalini, D.Meera, "Breast cancer prediction system using Data mining methods", International Journal of Pure and Applied Mathematics, Volume: 119, Issue: 12, PP: 10901-10911, 2018

[13] G. Ravi Kumar, Dr. G. A. Ramachandra, K.Nagamani, "An Efficient Prediction of Breast Cancer Data using Data Mining Techniques", International Journal of Innovations in Engineering and Technology (IJIET), Volume: 2, Issue:14, PP: 3677-3685, 2013

[14] Nitasha, "Review on Breast Cancer Prediction Using Data Mining Algorithms",International Journal of Computer Science Trends and Technology (IJCST), Volume: 7, Issue: 25, PP: 732-740, 2019

[15] S. Yuvarani and Dr. C. JothiVenkateswaran, "Breast Cancer Detection In Data Mining: A Review", Journal of Computer Science and Applications, Volume: 7, Issue: 1, PP:245-252, 2015

[16] A. Priyanga, Dr. S. Prakasam, "The Role of Data Mining-Based Cancer Prediction system (DMBCPS) in Cancer Awareness", International Journal of Computer Science and Engineering Communications, Volume: 10, Issue: 26, PP: 1740-1748, 2013

[17] Deneshkumar V, Manoprabha M, Senthamarai Kannan K, "Comparison of Datamining Techniques for Prediction of Breast Cancer", International Journal of Scientific & Technology Research Volume: 8, Issue: 08, PP: 268-276, 2019

[18] K. Arutchelvan, Dr. R. Periyasamy, "Cancer Prediction System using Data Mining Techniques", International Research Journal of Engineering and Technology (IRJET), Volume: 2 Issue: 68, PP: 797-804, 2015

[19] Shelly Gupta, Dharminder Kumar, Anand Sharma, "Data Mining Classification Techniques Applied for Breast Cancer Diagnosis and Prognosis", Indian Journal of Computer Science and Engineering (IJCSE), Volume: 2, Issue: 25, PP: 3782-3790, 2011

ABOUT THE AUTHORS

Anuradha is a Ph.D. Scholar in the Department of Computer Science, AKTU Lucknow. She has completed her M.Tech from GGSIPU Delhi. She has several publications in various journals of repute.

Prof. Harish Kumar Taluja received the Ph.D. in Computer Science and Engineering in 2015, with a specialization in the Neural Network Web Mining. He had published good no of papers in national and International Journals.