Booster In High Dimensional Data Classification Using Cnn And Decision Tree Algorithm

Aruljothi R, Maya Eapen

y

Abstract*: Classification problems in high dimensional data with small number of observations are becoming more common especially in microarray data. The performance in terms of accuracy is essential while handling sensitive data particularly in medical field. For this the stability of the selected features must be evaluated. Therefore, this paper proposes a new evaluation measure that incorporates the stability of the selected feature subsets and accuracy of the prediction. Booster in feature selection algorithm helps to achieve the same. The proposed work resolves both structured and unstructured data using convolution neural network based multimodal disease prediction and decision tree algorithm respectively. The algorithm is tested on heart disease dataset retrieved from UCI repository and the analysis shows the improved prediction accuracy.

Index Terms*: Feature Selection, Micro Array, Structured Data, Un-structured Data.

I. INTRODUCTION

High dimensional data are data characterized by few dozen to many thousands of dimensions. Microarrays may be used to measure gene expression in many ways. Gene expression microarray experiments have generated large amounts of data that are collected in public repositories. A microarray database is a source database which includes microarray gene expression data. It is typically a glass slide on to which DNA molecules are fixed in an orderly manner at specific locations called spots or features. It may contain thousands of spots and each spot may contain a few million copies of identical DNA molecules that uniquely correspond to a gene. As high dimensional data contains a greater number of features, processing such data is complex. Feature selection is very important as data is created constantly and at an ever increasing rate, it helps to reduce the high dimensionality of some problems. Feature selection helps to reduce redundant data, removes unrelated data thereby increasing the accuracy of prediction.

Some of the machine learning algorithm which deals with feature selection are FAST, FCBF. The high dimensional data can be structured and/or unstructured. The existing methods does not deal with these cases. Algorithms which can handle these cases are convolutional neural network and decision tree. Convolutional Neural Networks (CNN) are deep artificial neural networks which primarily adopted to stratify images. It groups them by similarity, and further perform object recognition. They are algorithms that can identify faces, individuals, street signs, tumours, platypuses and many other aspects of visual data. Decision Tree (DT) is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node. Therefore, the proposed model make use of these two algorithms to handle high dimensional data. The model uses booster during feature selection process which helps to improve the prediction accuracy.

The contributions of the proposed prediction model are,

- Feature selection to handle structured data
- Feature selection to handle unstructured data

The rest of the paper is organized as follows. Section II brief about related work. The proposed prediction model and its description in elaborated in section III. In section IV experimental evaluation is carried out. Finally, section V concludes the paper followed by future work.

II. RELATED WORK

In [1], the author specifies many big-data systems, large amounts of information are recorded and stored for analytics purposes. In most of the cases, this huge volume of data does not offer any further benefits for effective decision making. In many cases, it may be rather confusing and or very costly for gathering, storage, and processing. For example, let us consider, tumour classification using high-throughput microarray data. This is challenging due to large number of noisy features present. These don’t contribute to the reduction of classification errors. In these cases, limited number of genes that highly differentiate among the classes are found. Thus, in this paper, author report about a particular class of machine learning. This is also known as the problem of feature selection within support vector machine classification. It deals with finding an accurate binary classifier which uses a minimal number of features. Author introduce a new approach based on iteratively adjusting a bound on the l1-norm of the classifier vector in order to force the number of selected features to converge towards the desired maximum limit. Also analysed two real-life classification problems with high dimensional features. The first example deals with the medical diagnosis of tumours using microarray data, in which generic approach for cancer classification based on gene expression is presented. The second case deals with sentiment classification of on-line reviews from Amazon, Yelp, and IMDb. The results show that the approach is quite simple which is computationally

Revised Manuscript Received on July 22, 2019.

Aruljothi R, PG Scholar, Department of Computer Science and Engineering, Jerusalem College of Engineering, Chennai, India.

Maya Eapen, Assistant Professor, Department of Computer Science and Engineering, Jerusalem College of Engineering, Chennai, India.

tractable. Further, it results in low error rates. This low error rate is the main feature for the construction of advanced decision-support systems. The disadvantage is it is a dimensionality reduction method; it does not select features, it just tries to find linear combination of features.

In [2], the author specifies about classification problems in high dimensional data with small number of observations are becoming more common especially in microarray data. Over the last 20 years, so many classification models and feature selection (FS) algorithms which are effective have been adopted for higher prediction accuracies. But, in case of high dimensional data, the final result of an FS algorithm based on the prediction accuracy will be unstable over the variations in the training set. Q-statistic, a new evaluation measure is suggested in this paper to incorporate the stability of the particular feature subset in addition to the prediction accuracy. Then the Booster of an FS algorithm is used that boosts the value of the Q-statistic of the algorithm applied. If the data set is not intrinsically difficult to predict with the provided algorithm, empirical studies using synthetic data and 14 microarray data sets illustrate that Booster boosts the value of the Q-statistic. Further, the prediction accuracy of the algorithm applied is also enhanced. But it is Hard to implement in real time platform.

In [3], the author defines about the Classification problems in high dimensional data with small number of observations are becoming more common particularly in microarray data. To predict or the feature selection (FS) algorithm applied is not efficient with the accurate data set. The LP Boost maximizes a margin between training samples of dissimilar classes and therefore also belongs to the class of margin-maximizing supervised classification algorithms. Therefore, Booster can also be used as a criterion to estimate the act of an FS algorithm or to estimate the complexity of a data set for classification. LPBoost iteratively optimizes double misclassification costs and vigorously generates pathetic hypotheses to build new LP columns. But It does not provide better predictive performance.

In [4], the author defines about the Grouping issues in high dimensional data with few perceptions are ending up more typical particularly in microarray data. Assume that the learner can access all the features of training instances, and the goal is to efficiently identify a fixed number of relevant features for accurate prediction. The learner is allowed to access a fixed small number of features for each training instance to identify the subset of relevant features. This work proposes a new estimation measure Q-statistic that includes the solidity of the selected feature subset in addition to the estimate accuracy. Then it proposes the Booster of an FS algorithm that boosts the value of the Q-statistic of the algorithm applied. Empirical studies based on synthetic data and 14 microarray data sets show that Booster boosts the estimate accuracy of the algorithm. Here, the Complexity of the classification increases.

In [5], the author describes about the high-dimensional data that bring great challenges in terms of computational complexity and classification performance. Therefore, it is important to efficiently compress in a low-dimensional feature space from high dimensional feature space. The best aspect of Feature selection is it preserves the original features which can be combined feature subset that has better explanatory ability. It is very important to study the reduction of dimension and understand the practical problems involved in mechanism of microarray data. In applications involved in medical field, such as high dimensional cancer microarray data, the important step is dimension reduction. In this research, a new Maximal Information-based Nonparametric Exploration method is proposed for the dimension reduction of the microarray data. The MIC (Maximal Information Coefficient) plays the important role to show the relation between the data in MINE method. The paper focused on improving the performance in terms of accuracy, relevance, and redundancy, after comparing the performance of MINE method and Total PLS algorithm on data. It is Time and computation expensive.

In [6], the author focused on more accurate prediction of presence of heart disease with reduced number of attributes. Thirteen attributes were involved originally in forecasting the heart disease. Then these thirteen attributes are reduced to 11 attributes in this approach. Three classifiers like Naive Bayes, Decision Tree and Bagging algorithm are used to predict the diagnosis of patients with the high accuracy. Author uses 10-fold cross validation method to measure the unbiased estimate of these prediction models. But the major drawback is Dependencies exist among variables.

III. PROPOSED PREDICTION MODEL

The proposed prediction model is shown in fig 1. It consists of two major components, one is to predict structured data and other is for unstructured data prediction.

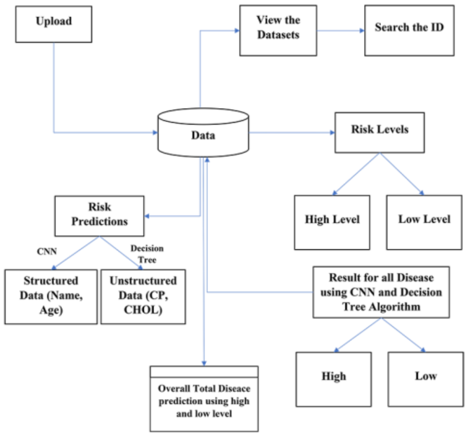

Upload
View the Datasets
Search the ID
Data
Risk Levels
High Level
Low Level
Risk Predictions
CNN
Decision Tree
Structured Data (Name, Age)
Unstructured Data (CP, CHOL)
Result for all Disease using CNN and Decision Tree Algorithm
High
Low
Overall Total Diseace prediction using high and low level

Fig. 1. Architecture of the proposed prediction model

To start with the high dimensional microarray data is collected from standard repository [7]. This collected data is stored in the database. Then the two major prediction process mentioned above are carried out and risk levels are identified.

We combine the structured and unstructured data in healthcare field to assess the risk of Heart disease. First, we use Decision tree to reconstruct the missing data from the medical records collected from a hospital in central China. we could determine the major chronic diseases in the region using statistical knowledge, which is the Second reason. Third, hospital experts are to be consulted to extract useful features and to handle structured data [2].

A. Convolution Neural Network prediction approach

A Convolutional Neural Network (CNN, or ConvNet) is a class of deep, feed-forward artificial neural networks and it has been successfully used for analysing visual imagery, in machine learning. The CNN has an excellent performance in machine learning problems[8] CNNs use a variation of multilayer preceptors designed to require minimal pre-processing. It uses relatively little pre-processing compared to another image classification.

A neural network is obtained by connecting many neurons together. We focus on feedforward networks, formally defined by a directed acyclic graph G = (V, E).

Input nodes: nodes with no incoming edges

Output nodes: nodes without out going edges

weights: w: E→ R Calculation using breadth-first-search (BFS),

where each neuron (node) receives as

input: $a[v] = Xu \rightarrow v \in E\, w[u \rightarrow v] o[u]$ and

output $o[v] = \sigma(a[v])$

The working principle of CNN is diagrammatically represented in figure 2. It consists of one input, one output layer and more than one hidden layer. In these hidden layers feature sets are mapped.

For unstructured text data, we select the features automatically using CNN algorithm. Finally, we propose a novel CNN-based multimodal disease risk prediction algorithm for unstructured data.

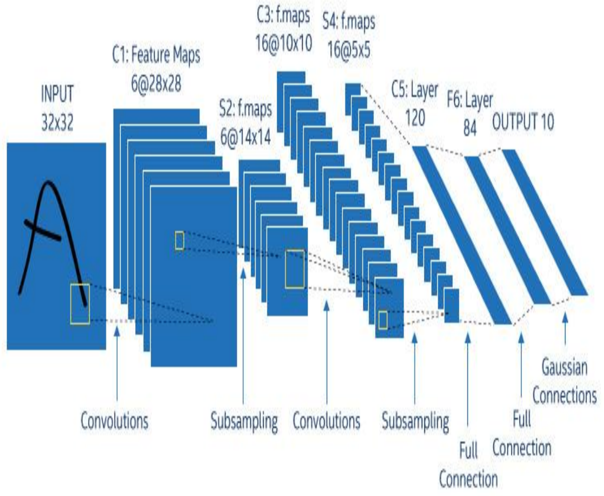


Fig. 2 Working principle of CNN

B. Decision Tree Based Prediction Approach

Decision tree algorithm is a kind of data mining model to make induction learning algorithm.[9].Decision Tree algorithm is part of the family of supervised learning algorithms. For classification and regression, decision trees are usually used which are tree-structured models.

The decision tree learning algorithm recursively learns the tree as follows:

1) Assign all training instances to the root of the tree. Set current node to root node.
2) For each attribute
   a) Partition all data instances at the node by the value of the attribute.
   b) Compute the information gain ratio from the partitioning.
3) Identify feature that results in the greatest information gain ratio. Set this feature to be the splitting criterion at the current node.
   c) If the best information gain ratio is 0, tag the current node as a leaf and return.
4) Partition all instances according to attribute value of the best feature.
5) Denote each partition as a child node of the current node.
6) For each child node:
   d) If the child node is “pure” (has instances from only one class) tag it as a leaf and return.
   e) If not set the child node as the current node and recurse to step 2

IV. EXPERIMENTAL EVALUATION

For experimental evaluation, the data is collected from UCI repository [7]. Experiment is carried out in Pentium IV 2.4 GHz processor, using Java language and MySQL is used for storage of data. Only Authorised persons needs to login with username and password. If both matches, then he/she will be considered as valid person. After login, Admin has to upload disease datasets which has 13 attributes of the patients. The uploaded dataset can be viewed by the admin and he can edit the missing fields in the dataset. Admin has the capability to maintain all user details.

A. Data Analysis

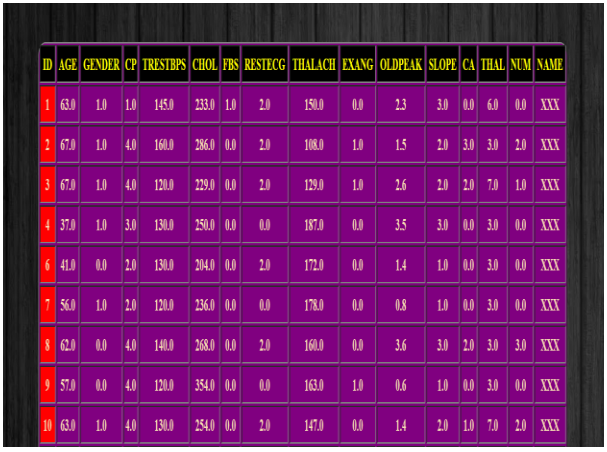
Data analysis is carried out for prediction of structured and unstructured micro array data. The Structured data is based upon the Laboratory report. And the Unstructured data is retrieved from the dataset.

1) Classification of Structured Data: The structured data includes laboratory data and the patient’s basic information such as the patient’s age, gender and life habits, etc. Structured data (S-data): uses the patient’s analysed data to predict whether the patient is at high-risk of heart disease or Low-risk of heart disease. The number of structured text data is extracted by using Decision tree algorithm. Unlike other supervised learning algorithms, decision Tree algorithm is a part of family of supervised learning algorithms which can be used for solving regression and classification problems also. The general motive of using Decision Tree is to create a training model. This can also be used to predict class or value of target variables by learning decision rules inferred from training data (prior data).
2) Classification of Unstructured Data: While the unstructured text data (chol, fbs, restecg, thalach, exang) includes the patient’s narration of his/her illness, the doctor’s interrogation records and diagnosis, etc. Unstructured data (U-data): uses the patient’s analysed text data to predict whether the patient is at high-risk or Low-risk of Heart disease. The number of unstructured text data extracted by using CNN algorithm. The Risk level of the unstructured data can be predicted using CNN algorithm. They are also known as

shift invariant or Space Invariant Artificial Neural Networks (SIANN). This is based on translation invariance characteristics and their shared-weights architectural structure.

B. B.Results

The details of the dataset contain id, age ,gender, cholesterol and other heart related attributes are shown in Fig.3.



ID	AGE	GENDER	CP	TRESTBPS	CHOL	FBS	RESTECG	THALACH	EXANG	OLDPEAK	SLOPE	CA	THAL	NUM	NAME
1	63.0	1.0	1.0	145.0	233.0	1.0	2.0	150.0	0.0	2.3	3.0	0.0	6.0	0.0	XXX
2	67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	1.5	2.0	3.0	3.0	2.0	XXX
3	67.0	1.0	4.0	120.0	229.0	0.0	2.0	129.0	1.0	2.6	2.0	2.0	7.0	1.0	XXX
4	37.0	1.0	3.0	130.0	250.0	0.0	0.0	187.0	0.0	3.5	3.0	0.0	3.0	0.0	XXX
6	41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0	0.0	1.4	1.0	0.0	3.0	0.0	XXX
7	56.0	1.0	2.0	120.0	236.0	0.0	0.0	178.0	0.0	0.8	1.0	0.0	3.0	0.0	XXX
8	62.0	0.0	4.0	140.0	268.0	0.0	2.0	160.0	0.0	3.6	3.0	2.0	3.0	3.0	XXX
9	57.0	0.0	4.0	120.0	354.0	0.0	0.0	163.0	1.0	0.6	1.0	0.0	3.0	0.0	XXX
10	63.0	1.0	4.0	130.0	254.0	0.0	2.0	147.0	0.0	1.4	2.0	1.0	7.0	2.0	XXX

Fig.3 Sample Dataset

To retrieve the details of patients unique id has to be provided as shown in Fig.4. When the unique id is entered the patients details are displayed.

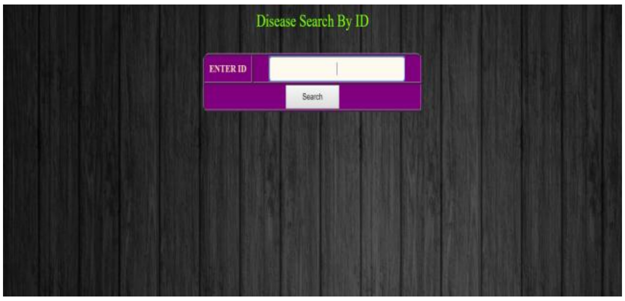
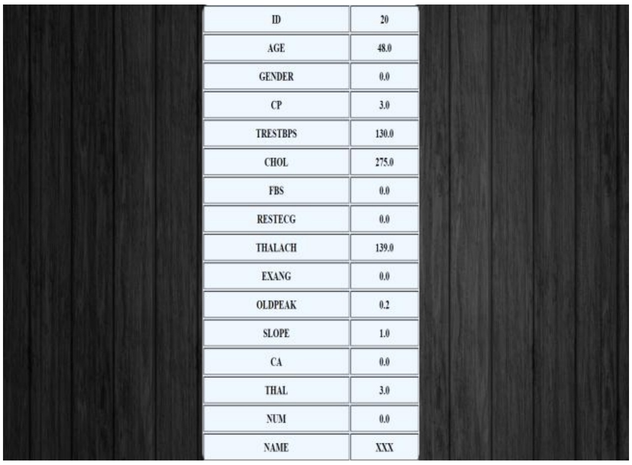


Fig.4 Snapshot of searching patient’s details

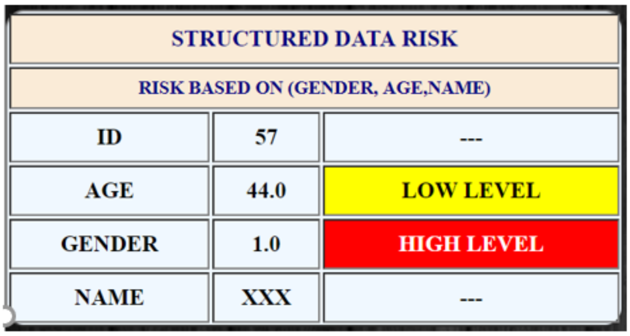
The Fig.5 shows the patient’s particulars retrieved and used in prediction process according to the unique id that is given.



ID	20
AGE	48.0
GENDER	0.0
CP	3.0
TRESTBPS	130.0
CHOL	275.0
FBS	0.0
RESTECG	0.0
THALACH	139.0
EXANG	0.0
OLDPEAK	0.2
SLOPE	1.0
CA	0.0
THAL	3.0
NUM	0.0
NAME	XXX

Fig. 5 Patient’s particulars

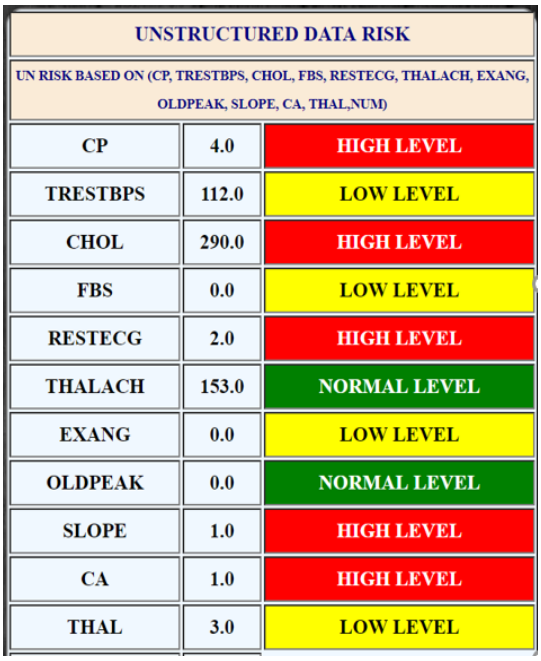
Structured data such as id, age, gender, name is classified from the given data as shown in Fig.6.



STRUCTURED DATA RISK		
RISK BASED ON (GENDER, AGE,NAME)		
ID	57	---
AGE	44.0	LOW LEVEL
GENDER	1.0	HIGH LEVEL
NAME	XXX	---

Fig.6 Classification of structured data

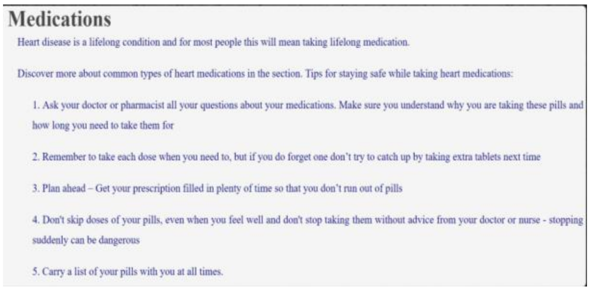
Unstructured data such as CP, CHOL, FBS, etc. are classified from the given data as high level, low level, and normal level as shown in Fig.7.



UNSTRUCTURED DATA RISK		
UN RISK BASED ON (CP, TRESTBPS, CHOL, FBS, RESTECG, THALACH, EXANG, OLDPEAK, SLOPE, CA, THAL,NUM)		
CP	4.0	HIGH LEVEL
TRESTBPS	112.0	LOW LEVEL
CHOL	290.0	HIGH LEVEL
FBS	0.0	LOW LEVEL
RESTECG	2.0	HIGH LEVEL
THALACH	153.0	NORMAL LEVEL
EXANG	0.0	LOW LEVEL
OLDPEAK	0.0	NORMAL LEVEL
SLOPE	1.0	HIGH LEVEL
CA	1.0	HIGH LEVEL
THAL	3.0	LOW LEVEL

Fig. 7 Classification of Unstructured data

Medications are suggested according to the level of risk after the analysis of risk level as shown in Fig 8.



Medications

Heart disease is a lifelong condition and for most people this will mean taking lifelong medication.

Discover more about common types of heart medications in the section. Tips for staying safe while taking heart medications:

1. Ask your doctor or pharmacist all your questions about your medications. Make sure you understand why you are taking these pills and how long you need to take them for
2. Remember to take each dose when you need to, but if you do forget one don’t try to catch up by taking extra tablets next time
3. Plan ahead – Get your prescription filled in plenty of time so that you don’t run out of pills
4. Don't skip doses of your pills, even when you feel well and don't stop taking them without advice from your doctor or nurse - stopping suddenly can be dangerous
5. Carry a list of your pills with you at all times.

Fig. 8 Medications based on Risk level

After analysing the structured and unstructured decision models, the accuracy of 95% is achieved.

V. CONCLUSION AND FUTURE WORK

The disease risk model contains both structured and unstructured features. Through the experiment, we draw a conclusion that the performance of CNN-MDPR is better than other existing methods. We leverage not only the structured data but also the text data of patients based on the proposed CNN and Decision Tree algorithm.

The future work is, to incorporate and analyse the graph structured data.

REFERENCES

1. BissanGhaddar, JoeNaoum-Sawaya, “High dimensional data classification and feature selection using support vector machines”.
2. HyunJi Kim, Byong Su Choi and Moon Yul Huh, “Booster in high dimensional data classification.”.
3. E. Kokilamani, Dr. R. Gunavathi, “A Survey on Boosting High dimensional Feature Selection Classification”.
4. K. Sai Sravani, Dr. P. Kiran Sree, “Efficient Technique for Classifying High-Dimensional Data”.
5. Shubhangi N. Katole; Swapnili P. Karmore, “A new approach of microarray data dimension reduction for medical applications”
6. Vikas Chaurasia, Saurabh Pal, “Data Mining Approach to Detect Heart Diseases”.
7. https://archive.ics.uci.edu/ml/datasets/%20heart+Disease
8. Saad Albawi ; Tareq Abed Mohammed ; Saad Al-Zawi,Understanding of a convolutional neural network
9. Linna Li ; Xuemin Zhang,Study of data mining algorithm based on decision tree

AUTHORS PROFILE

Aruljothi R, completed Bachelor’s Degree in Computer Science and Engineering in the year 2017. Presently she is pursuing Post Graduation in Computer Science and Engineering. Her area of interest are Data Mining and Artificial Intelligence.

Dr Maya Eapen, is working as Assistant professor in the department of Computer Science and Engineering at Jerusalem college of Engineering, Chennai, India. She obtained her Ph D degree from Anna University, Chennai in 2017. Her main research interests include medical Image processing, computer vision, Image mining and visualization