International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Ensemble Classification Method for Credit Card Fraud Detection

Inderpreet Kaur, Mala Kalra

Abstract*: Credit card frauds are on the rise and are getting smarter with the passage of time. Usually, fraudulent transactions are conducted by stealing the credit card. When the loss of the card is not noticed by the cardholder, a huge loss can be faced by the credit card company. In the existing work, it has been found that the researchers have utilized Voting based method to identify credit card frauds. The problem with voting based method is that they are more complex and more time consuming. In this research work, a hybrid approach based on KNN and Naive Bayes for the detection of credit card frauds. KNN will be used as the base classifier and it will return predicted result. The predicted result will be provided as input to the Naive Bayes classifier which will generate the final result. The proposed model will be compared with existing techniques and the results are analyzed in terms of recall, precision, accuracy and execution time.

***Index Terms*: Credit Card Fraud Detection, Ensemble, Voting.**

I. INTRODUCTION

Credit cards are being used commonly today for buying several goods and accessing various services in our daily lives. When the physical-card based purchasing technique is applied, the card is given by the cardholder to the merchant so that a successful payment method can be performed [1]. If the credit card is used by someone else for personal reasons and the owner of the card does not have any knowledge about it, the credit card fraud is outlined. The person who is conducting the fraud will never aim to contact the owner of the card or repay the losses to the actual user. The approaches that result in causing fraud need to be perceived initially so that they can be handled in an effective manner. For committing the fraud, a variety of methods are used by credit card fraudsters [3]. The existing purchase data of the particular cardholder is the basis on which these fraud detection methods are proposed [2].

There are various issues and challenges being faced by fraud detection systems amongst which few are enlisted below [4]:

a. Defining the Type and Level of Fraud: It is very important to define the type and the level of fraud that has occurred in any transaction. However, a universal measurement approach through which all the frauds that exist today can be identified is impossible to be developed. Thus, the level of frauds is reported and then depending upon certain factors, that fraud is categorized into any of the subcategories that are predefined.

b. Fraud Departments Resides in Silos: Silos are the organizations which work individually and do not share any personal information with other groups. It is impossible for a fraud detection approach to identify the fraudster that goes online, changes [5] the account address of a genuine user and then requests for a new card for personal use, for such silo systems. It is possible for one team to just view the change is address and the other to view the transactions made from that card.

c. A requirement of Real-time Detection Techniques: The techniques available currently for detecting the frauds only work for cases where fraudulent transactions have occurred for more than one time. Also, the speed at which fraudster is moving is higher as compared to the fraud monitoring solution. Therefore, till the time that fraud is detected, the money of genuine user is already stolen [6]. So, it is important to provide real-time detection techniques for financial institutions for protecting their clients. The suspicious transaction patterns are to be monitored and recognized immediately by the tools such that actions can be made as soon as the fraudster makes any attempt. Thus, the losses can be prevented.

d. Fraud is measured a Reasonable Issue: The banks have considered fraud prevention techniques as their basic necessity since the numbers of frauds are growing with each day. Therefore, a reasonable issue being faced today is the various potential gaps, which needs to be exposed and resolved [7].

Different approaches have been applied for detecting the frauds .These approaches are explained below:

a. Artificial Neural Network: A set of interlinked nodes that are designed for imitating the working of a human brain is known as an artificial neural network (ANN). A weighted link is assigned to all the other nodes that are present in the adjacent layers of each node [8].

b. Genetic Algorithm (GA): The genetic algorithms were introduced inspiring from natural evolution. Chromosomes are the binary strings that are used to represent the populations of candidate solutions. It is based on the concept that the chances of survival and reproduction are higher for the chromosomes with higher quality i.e. having better fitness value. c. Hidden Markov Model (HMM): A double embedded stochastic process using which highly complicated stochastic processes can be generated is known as a hidden Markov model [9].

Manuscript published on 30 September 2019
* Correspondence Author
Inderpreet Kaur*,Computer Science & Engineering, National Institute of Technical Teachers Training & Research, Chandigarh, India.
Mala Kalra, Computer Science & Engineering, National Institute of Technical Teachers Training & Research, Chandigarh, India

A Markov process that has unobserved states is assumed to be available within the underlying system. The only unknown parameters are the definite transition of the states within the simpler Markov models.

d. KNN Classifier: KNN is the non-parametric algorithm used in case of regression and classification.

The input consists of K-nearest training examples in the feature space and on the other hand, the output depends upon whether KNN belongs to regression category or classification category [9].

e. Naïve Bayes: It implements the Bayesian rule on categorical data for performing classification. In comparison to other classification approaches, the performance of algorithm is known to be better and very simple.

II. LITERATURE REVIEW

This section reviews the literature for different existing techniques.

Kuldeep Randhawa et al. [10] have presented a model for detection of credit card fraud using machine learning. Initially, standard methods were used after that hybrid models came into picture which made use of majority voting and Ada Boost methods. Publically available dataset had been used to evaluate the model efficiency and another data set used from the financial institution and analyzed the fraud. Then the noise was added to the data sample through which the robustness of the algorithms could be measured. For further evaluation of the models noise of about 10% and 30% has been added to the sample data. Several voting methods have achieved a good score of 0.942 for 30% added noise. Thus, it was concluded that the voting method showed much balanced performance in the presence of noise.

Abhimanyu Roy et al. [11] proposed deep learning topologies for the detection of fraud in online money transaction.This strategy is obtained from the neural artificia l network with builtin time and memory components such as short-term long term memory. According to the efficiency of these components in fraud detection, almost 80 million online transactions through credit card were pre-labeled as fraudulent and legal. The study proposed by the researchers provides an effective guide to the sensitivity analysis of the proposed parameters as per the performance of the fraud detection. The researchers also proposed a parameter tuning framework for fraud detection of Deep Lea rning topologies. This enables financial institution to to reduce losses by avoiding fraudulent activity.

Shiyang Xuan et al. [12] have presented two types of random forests which trains an normal and abnormal transactions behavioral characteristics. The researcher compares these two random forests which are differentiated on the basis of their classifiers performance. The data used is of an e-commerce company of China which is utilized to analyze the performance of these two types of random forests model. In this paper, the author has used B2C dataset for the identification and detection of fraud from the credit cards. Therefore, the researcher concluded from the result that the proposed random forests provide good results on small dataset but there are still some problems like imbalanced data which makes it less effective than any other dataset.

Zahra Kazemi et al. [13] proposed Deep auto encoder which is used to extract the best characteristics of the information from the credit card transaction. This will further add soft max software to resolve the class labels issues. An over complete auto encoder is used for mapping the data into a high dimensional space and a sparse model was used in a descriptive manner which provides benefits for the classification of a type of fraud. Deep learning is one of the most motivated and powerful approaches used for detecting credit card fraud. These types of networks have a complex distribution of data which is very difficult to recognize. Deep auto encoder has been used in some stages to extract the best features of the data and for the classification purposes. Also, higher accuracy and low variance are achieved within these networks.

John O. Awoyemi et al. [14] proposed an investigation through which the performances of several algorithms were evaluated when they were applied on data that is highly skewed. European cardholders 284,807 transactions were used as a source to generate the dataset of credit card transactions. On skewed data, hybrid approach of oversampling and under sampling is performed. There are three different techniques applied to raw and pre-p rocessed data in Python. Based on certain parameters like precision, sensitivity, accuracy, balanced classification rate and so on, the performances of these techniques are evaluated. It is seen through the achieved results that in comparison to naïve Bayes and logistic regression approaches, the performance of k-NN is better.

Sharmistha Dutta et al. [15] presented a study on the commonly found crime within the credit card applications. There are certain issues faced when the existing non-data mining approaches are applied to avoid identity theft. A novel data mining layer of defense is proposed for solving these issues. For detecting the frauds within various applications, two algorithms named Communal Detection and Spike Detection which generate novel layer. Thus, results can be generated by the system by consuming a huge amount of time. Since the attackers do not get time to modify their behaviors with respect to the algorithms being deployed in real time, there is no true evaluation achieved even after a regular update of the algorithms. Therefore, it is not possible to properly demonstrate the concept of adaptability. These issues can be resolved by making certain enhancements in the proposed algorithm in future work.

III. RESEARCH METHODOLOGY

This research work is based on the prediction of fraud transactions of the credit card.

Following are the steps of Proposed Work:

Step 1: Input the dataset from the UCI repository for the Credit Card Fraud Detection.

Step 2: Data Preprocessing will be performed in which missing values and redundant data from the dataset will be handled.

Step 3: Apply the method of cross-validation to divide input dataset into training and test. The training dataset must be large in size as compared to the test set. The dataset is divided into 40/60 proportion of test and training sets.

Step 4: Apply KNN classifier for the prediction of the test set. The nearest neighbor point is chosen as the one which has the lowest distance.

Euclidean distance formula is used to measure the distances and select data point based on similarity. Once the value of k is selected, distance is calculated by the equation given below where y_i is the *ith* case and *y* denotes the prediction or outcome

$$y = 1/k \sum_{i=1}^{k} y_i$$

Step 5: Apply Naïve Bayes classifier on the results of KNN classifier. Naïve Bayes is based on Bayes theorem which uses conditional probability to classify the data.

$$P(C_i|X) = P(X|C_i)P(C_i)/P(X)$$

The classifier calculates the conditional probability of various classes. The class that has highest conditional probability is chosen.

Step 6: The results of hybrid model are used to analyze performance in terms of precision, recall, accuracy and execution time.

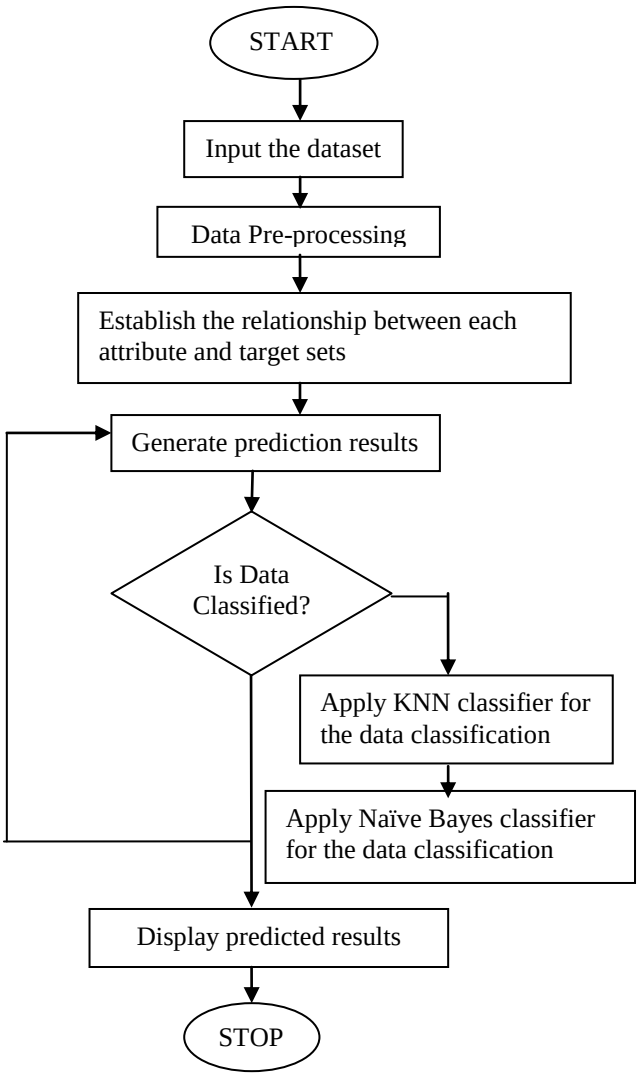

Figure 1: Flowchart of Proposed Work

IV. EXPERIMENTAL RESULTS

The proposed model is implemented using python where anaconda tool is used as a platform to perform all experiments. The anaconda has various inbuilt libraries like sklearn, pandas etc. The sklearn library has been used for implementation of machine learning algorithms in python language. The dataset for the proposed work is collected from the UCI Repository[10] and is described in Table 1.

The transactions made by credit cards in the year 2013 by European cardholders are collected to generate the dataset to be used in this research. The transactions occurring in two days were included in the dataset where out of 284,807 transactions, 492 frauds were detected. Highly unbalanced dataset is generated here. 0.172% of all the transactions are included in positive class. Only the numerical input variables that are the result of PCA transformation are included here. The principal components attained using PCA are the features denoted by V1, V2 and so on. Time and Amount are the only two features which have not been transformed with PCA. The seconds elapse between each transactions and the first transaction in the dataset are included in feature "Time". The transaction amount is known to be the feature "Amount". The response variable is known to be the feature "Class" and value 1 is taken in the case of fraud and 0 in genuine transaction.

Table 1: Dataset Description

Dataset Characteristics	Multivariate
Number of Instances	30000
Attribute Characteristics	Integer, Real
Associate Tasks	Classification
Number of Attributes	24
Missing Values	No
Area	Business
Date Donated	2016-01-26
Number of Web Hits	368365

The performance of the proposed work is validated by comparing it with existing techniques in parameters of recall, precision, accuracy, and execution time. The analysis is performed by taking different proportions of test and training sets.

Table 2: Recall Analysis

Model	**40/60 percent**	**30/70 percent**	**20/80 percent**	**10/90 percent**
Naïve Bayes	0.25	0.27	0.26	0.31
Random Forest	0.28	0.25	0.25	0.27
Ada Boost	0.2	0.18	0.18	0.21
Voting Classification	0.25	0.25	0.17	0.27
Hybrid Classification	0.33	0.32	0.31	0.36

Table 2 shows the recall achieved for the considered algorithms. Figure 2 represents the comparison of the proposed work with existing work for detecting fraud in terms of recall and graph is plotted taking different proportions of test and training sets. From the results it is clear that the recall of the proposed work is
15 -24% is better than the existing techniques.

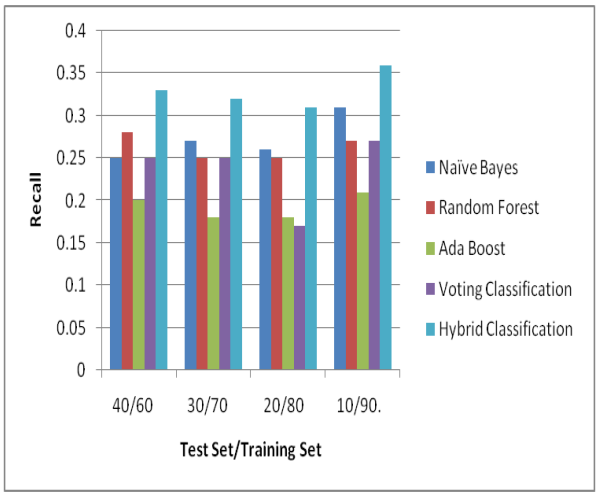

Figure 2: Recall Analysis with different proportions of test and training sets

Table 3: Precision Analysis

Model	40/60 percent	30/70 percent	20/80 percent	10/90 percent
Naïve Bayes	0.1	0.1	0.1	0.1
Random Forest	0.91	0.89	0.9	0.86
Ada Boost	0.89	0.85	0.79	0.86
Voting Classification	0.9	0.89	0.9	0.9
Hybrid Classification	1.0	1.0	1.0	1.0

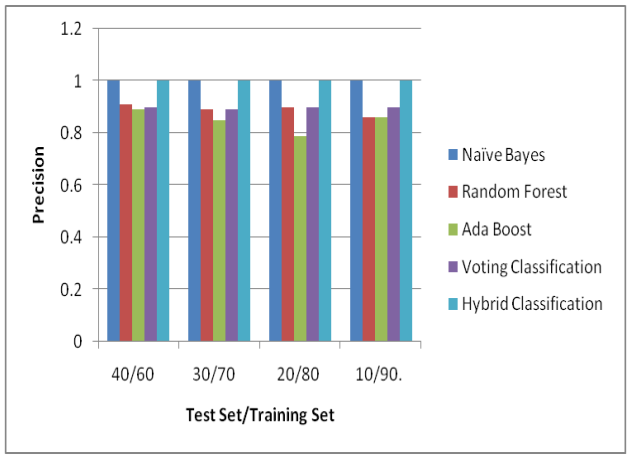

Figure 3: Precision Analysis with different proportions of test and training sets

Table 3 shows the precision achieved for the considered algorithms. Figure 3 depicts precision achieved by all algorithms for different proportions of test and training sets. From the results it is clear that the proposed work has 90-100% better value of precision than the existing techniques.

Table 4: Accuracy Analysis

Model	40/60 percent	30/70 percent	20/80 percent	10/90 percent
Naïve Bayes	97.01	96.91	96.87	96.71
Random Forest	99.59	99.6	99.59	99.6
Ada Boost	99.56	99.56	99.55	99.58
Voting Classification	99.59	99.6	99.56	99.61
Hybrid Classification	100.00	99.91	99.87	99.71

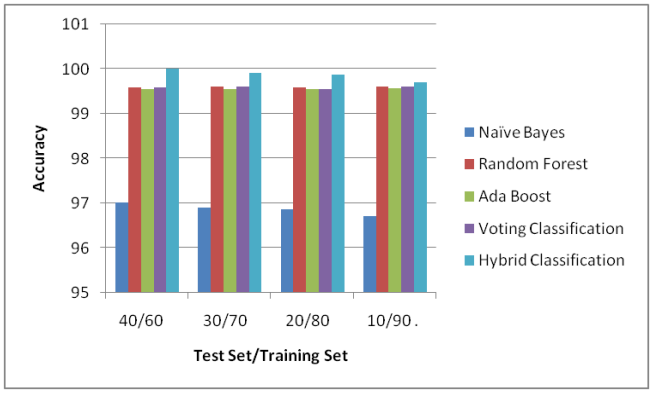

Figure 4: Accuracy Analysis with different proportions of test and training sets

Table 4 shows the accuracy achieved for the considered algorithms. Accuracy of various algorithms taking different proportions of test and training sets is shown in Figure 4. From the results it is clear that the accuracy of the proposed work is better than the existing techniques. Specifically 40/60 proportion of test and training set accuracy of 100% is achieved.

Table 5: Execution Time Analysis

Model	40-60 percent	30-70 percent	20-80 percent	10-90 percent
Naïve Bayes	2.34	2.12	2.10	2.16
Random Forest	2.67	2.34	2.56	2.78
Ada Boost	1.67	1.89	1.34	1.67
Voting Classification	1.12	1.45	1.78	1.90`
Hybrid Classification	1.10	1.30	1.34	1.56

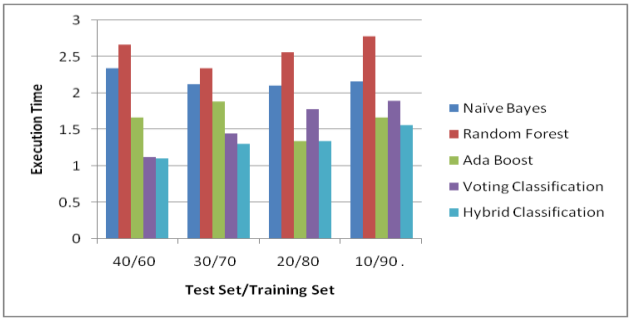

Figure 5: Execution Time Analysis with different proportions of test and training sets

Table 5 shows the execution time achieved for the considered algorithms. Figure 5 represents the comparison of the proposed work with existing work used for detecting credit card fraud terms of execution time and graph is plotted taking different proportions of test and training sets. From the results it is clear that the execution time of the proposed work is reduced up to 2 seconds,

V. CONCLUSION AND FUTURE SCOPE

Credit card frauds have become pervasive in recent years. This is not only the reason for constant increase in frauds. The main cause is the ease with which these frauds can be committed. Machine learning techniques have been proved efficient to detect credit card frauds. In order to improve the Precision and reduce complexity of the Credit Card Fraud Detection, hybrid classification approach using KNN and Naïve Bayes is proposed in this research work. The techniques proposed have three major steps Data pre-processing, feature extraction and classification. The first step is Data pre-processing is used to reduce noise and missing values. In second step features are extracted using KNN. In the third step Naïve Bayes is used to classify that fraud exists or not. Dataset used for the detection of fraud transactions is collected from publically available dataset from UCI Repository. To evaluate the performance of proposed model, different proportions of training and test sets are compared to existing work. Comparison of their outcome suggests the performance of proposed model is better than the existing techniques in terms of precision, recall, accuracy and execution time. Considering the test case of 40:60 proportion , there is an improvement of 15-24% and 90-100% for recall and precision respectively. Accuracy of the proposed model is 100% and has improved the execution time by 2 seconds. Performance achieved by the proposed model is highly efficient but still there is a scope of advancement. In Future the work can be extended by using the approach of clustering and feature simplification. This approach can cluster the dataset based on data similarity. Dataset can be simplified using PCA Algorithm. Classification approach can be applied on each cluster for prediction of fraud transactions.

REFERENCES

1. K. Modi and R. Dayma, "Review on fraud detection methods in credit card transactions, "*International Conference on Intelligent Computing and Control (I2C2)*, Coimbatore, pp. 1-5, 2017.
2. D. Pojee, S. Zulphekari, F. Rarh, and V. Shah, "Secure and quick NFC payment with data mining and intelligent fraud detection, "*2nd International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, pp. 148-152, 2017.
3. D. S. Sisodia, N. K. Reddy and S. Bhandari, "Performance evaluation of class balancing techniques for credit card fraud detection," *IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, Chennai, pp. 2747-2752, 2017.
4. L. Vergara, A. Salazar, J. Belda, G. Safont, S. Moral and S. Iglesias, "Signal processing on graphs for improving automatic credit card fraud detection," *International Carnahan Conference on Security Technology (ICCST)*, Madrid, pp. 1-6, 2017.
5. S. N. John, C. Anele, O. O. Kennedy, F. Olajide and C. G. Kennedy, "Real-time Fraud Detection in the Banking Sector Using Data Mining Techniques/Algorithm, "*International Conference on Computational Science and Computational Intelligence (CSCI)*, Las Vegas, NV, pp. 1186-1191, 2016.
6. M. S. Mahmud, P. Meesad and S. Sodsee, "An evaluation of computational intelligence in credit card fraud detection, "*International Computer Science and Engineering Conference (ICSEC)*, Chiang Mai, pp.1-6, 2016.
7. J.West and M. Bhattacharya, "An investigation on experimental issues in financial fraud mining," *IEEE 11th Conference on Industrial Electronics and Applications (ICIEA)*, Hefei, pp. 1796-180, 2016.
8. P. Wongchinsri and W. Kuratach, "A survey - data mining frameworks in credit card processing, "*13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, Chiang Mai, pp. 1-6, 2016.
9. E. M. Carneiro, L. A. Dias, A. M. Cunha and L. F. Mialaret, "Cluster Analysis and Artificial Neural Networks: A Case Study in Credit Card Fraud Detection, "*12th International Conference on Information Technology - New Generations (ITNG)*, pp. 122-126, 2015.
10. Kuldeep Randhawa, Chu Kiong Loo, Manjeevan Seera, Chee Peng Lim and Asoke K. Nandi, "Credit card fraud detection using AdaBoost and majority voting," *IEEE Access*, vol. 6, pp. 14277-14284, 2018.
11. A. Roy and J. Sun and R. Mahoney and L. Alonzi and S. Adams and P. Beling, "Deep learning detecting fraud in credit card transactions," in *Systems and Information Engineering Design Symposium (SIEDS)*, pp. 129-134, 2018.
12. Guanjun Liu, Zhenchuan Li, Lutao Zheng, Shuo Wang and Changjun Jiang Shiyang Xuan, "Random Forest for Credit Card Fraud Detection," in *IEEE 15th International Conference On Networking, Sensing and Control (ICNSC)*, pp.1-6, 2018.
13. Zarrabi, H. Kazemi, "Using deep networks for fraud detection in the credit card transaction," *IEEE 4th International Conference In Knowledge-Based Engineering and Innovation (KBEI)*, pp. 0630-0633, 2017.
14. John O., Adebayo O. Adetunmbi, and Samuel A. Oluwadaren Awoyemi, "Credit card fraud detection using machine learning techniques: A comparative analysis."*International Conference on Computing Networking and Informatics (ICCNI)*, pp. 1-9, 2017.
15. S. Dutta, A. K. Gupta and N. Narayan, "Identity Crime Detection Using Data Mining, "*3rd International Conference on Computational Intelligence and Networks (CINE)*, Odisha, pp. 1-5, 2017.

AUTHORS PROFILE

Inderpreet Kaur, received her Bachelor of Technology in Computer Science and Engineering from Chitkara University, Punjab. At present, she is pursuing Master of Technology in Computer Science and Engineering from National Institute of Technical Teacher Training and Research, Chandigarh. Her key area of interest includes Machine learning, Data Mining, Deep Learning.

Mala Kalra, received her Bachelor of Technology in Computer Science and Engineering from MDU, Rohtak, India. Master of Technology in CSE from PEC University of technology, Chandigarh, India. Ph.D in Engg. and Technology from Punjab University, India.. At present, she is working as an Assistant Professor in the Department of Computer Science and Engineering at the NITTTR.