International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol. 9, Issue.3, March 2021 E-ISSN: 2347-2693

Fraud Detection by the Use of Correlation Based Tree Formation Approach

Shivani[1*], Harjinder Kaur[2]

[1,2]Computer Engg, Swami Sarvanand College of Engg and Technology, PTU, Dina Nagar, India

[*]*Corresponding Author: shivishivani33@gmail Tel.: 6239793998*

DOI: https://doi.org/10.26438/ijcse/v9i3.712 | Available online at: www.ijcseonline.org

Received: 10/Mar/2021, Accepted: 17/Mar/2021, Published: 31/Mar/2021

Abstract— Credit card fraud detection becomes critical due to increase in online transactions. Customers bought products online more often than not. The payment is either through debit or financials. The malicious users may attack the online information and hack credit and debit cards. Detection and prevention mechanisms thus are need of the hour. Researchers work towards achieving immunity against these attacks but perfection yet not achieved. This paper proposes similarity based decision tree approach for financial fraud detection strategy by working on state driven dataset. The objective is to detect the attack at early stage to avoid extravagant situations. The result is presented in the form of classification accuracy, precision and execution time. The result in terms of classification accuracy and execution time is improved by the factor of 10%.

Keywords- Financial fraud, similarity based decision tree, classification accuracy, precision, execution time

I. INTRODUCTION

Financial transactions are conducted using credit and debit cards. Debit/credit card is a flimsy convenient plastic card that contains ID data, for example, a mark or picture, and approves the individual named on it to charge buys or administrations to his record - charges for which he will be charged intermittently. Today, the data on the card is perused via mechanized teller machines (ATMs), store per users, bank and is likewise utilized in online web managing an account framework. They have an exceptional card number which is of most extreme significance. Its security depends on the physical security of the plastic card just as the protection of the financial number. There is a fast development in the quantity of financial exchanges which has prompted a significant ascent in fraudulent exercises. Credit card fraud is a wide-extending term for burglary and fraud submitted utilizing a financial as a fraudulent wellspring of assets in a given exchange. By and large, the factual strategies and numerous data mining calculations are utilized to take care of this fraud discovery issue. The majority of the financial fraud identification frameworks depend on man-made consciousness, Meta learning and example coordinating.

Financial fraud detection is critical in the environment where users can invest money by the use of credit and debit card. The mechanism commonly employed in most literature includes neural network based approach for fraud detection. The fraud directly or indirectly impact trust. Gathering trust in online business is difficult task and to establish trust, online business must accompanied with fraud detection strategies. In addition to neural network based approach there exist genetic approach for the fraud detection. Genetic approach uses iteration to analyse the dataset but to the worst end, convergence rate is poor.

Those e-commerce sites are more successful in which user does not have fear of being cheated. Web provides number of mechanisms in order to gain trust with the online e-commerce websites. The customer relationship will vastly improves if the trust factor is good. Without trust the development of e-commerce cannot reach its full potential. This paper is structured as follows: first of all the concept of trust is discussed from various prospective, secondly we will examine trust in e-commerce environment, A new model for trust management is presented, in the next section we will focus on the referee trust, which is examined in more detail, next we examine the methodology and the data analysis will be conducted.

A. Concept of Trust

Trust is critical for the success of online business. Establishing trust is difficult. The primary reason for the same is metric determination. Metric measure the quantity and quality associated with the online business. Metric can be direct or indirect in nature. Trust falls within indirect measure for quality. This metric depends upon several other parameters like service, delivery time, frauds etc.
To establish trust customer review about the product sold by ecommerce website is compulsory. To this end several online business organizations conduct search engine optimization.

In fact, several organization establish trust service provider to motivate the customers towards the service they provide. Trust service provider can be a digital agent that is used to divert the traffic generated towards the online business.

Frauds to the sites can hamper the trust and traffic could be deviated away from website. Fraud detection strategies thus are required in order to tackle the trust issue. The mechanism can uses mining approach for filtering the information present within large dataset. Detection strategies must incorporate pre-processing mechanism. The pre-processing mechanism eliminate the problems such as missing values within the dataset.

Once the missing values from within the dataset is tackled , segmentation is applied that divides the dataset into critical and non-critical section.

After segmentation classification is performed to determine frauds from within the dataset.
All these aspect are used in order to establish trust in online business.

Rest of the paper is organized as under: section 2 gives literature survey, section 3 gives the proposed system and methodology, section 4 gives results and last section gives conclusion and references.

II. LITERATURE SURVEY

[1]proposed data mining method for abnormal fraud patterns prediction for this purpose sequential data mining is used in order to accomplish this data preprocessing mechanism is applied. After applying preprocessing mechanism the attributes will be analyzed this will be done using passes on medical data. The first pass determines whether support for each abnormal fraud patterns is present or not at the end of this phase the frequent abnormal fraud patterns within the database will be identified, a counter will be maintained to count the occurrence of each abnormal fraud patterns within the dataset. Next phase determines the second sequence of abnormal fraud pattern present within the dataset. The overall process yields the abnormal fraud pattern which can cause the occurrence of other abnormal fraud pattern. The abnormal fraud patterns resulting in another abnormal fraud patterns are termed as candidate generation. And for declaring that it is generated from the previous level Pruning is used.

[2]proposed a sequential mining approach for early assessment of chronic abnormal fraud patterns. The clinical database is considered .A dataset of patients derived from Taiwan, it derives richest of risk fraud patterns. Data pre-processing as performed to rectify the problem if found but missing values are not considered .sequential fraud pattern mining is used to observe the risk fraud pattern and generate the result. The problem with this approach is that no precautions have been suggested. The classification accuracy is 80% further improvement in classification is needed. The chronic abnormal fraud patterns is analysed in this paper built in over the existing problem.

[3]enhancement which can be improved proposed multiclass Naïve Bayes algorithm is used for prediction of particular abnormal fraud patterns but training it on set of data before implementation. This is downloaded from UCI repository work. The proposed system can help doctors to take clinical decisions where traditional decision support system fails, J47 algorithm is also used for proving the worth of study of accuracy in heart abnormal fraud patterns, breast cancer and diabetes approaches 83% by using this approach. This accuracy requires in future.

[4]sequence fraud pattern mining is proposed in order to detect the time duration used for promotion. the sequence or fraud pattern is checked from within the database. The weight of each sequence in each database is achieved from the interval of the successive element in the sequence and the mining is performed on the basis of weight considering time interval. Time interval based fraud pattern is used in this case. In preprocessing missing values are not considered.

[5]proposed a technique that extract sequential fraud patterns from hypotensive patient groups. These fraud patterns are further utilized to inform medical decisions and randomized clinical trials. It further extended by including various clinical features and also include some sequential fraud patterns. It also does not considered missing value during the preprocessing phase.

[6]Proposed technique named ConSgen that are used to identify the contiguous sequential generator and also minimize the redundant fraud patterns, It utilizes the divide conquer technique to find the sequential generator with contiguous constraints. But it does not considered the gapped alignments and also not discovered the binding sites.

[7]it identified a problem of top –k utility based regulation fraud pattern which is used to find out meaning in biology. Firstly proposed a utility model called TU-SEQ which is used to find top –K high utility gene regulation sequential fraud patterns. It is considering the relation between the various fraud patterns and interactions in biological studies.

[8]Proposed a mining technique that are used to reduce the complexity and cost of the data storage. It divide chunks into separate parts and regression analysis are to be done to analysis the trial variable and samples dataset. But it does not considered separate chunks for feature analysis and separate storage reservoir also not utilized.

[9]Proposed an application that utilizes the data mining technique to predict the heart abnormal fraud patterns. Also it guide the patient to take treatment at early stage. But is completely dependent upon patient input and does not considered predefined dataset values. It also not utilizes the missing value that are essential to predict abnormal fraud pattern.

[10]in this paper the analysis of various fraud pattern mining techniques is done and also the features of all the algorithms. It introduced various minimizing support

counting which is used for minimizing search space. We have generated small search space which will include earlier candidate sequence pruning then database is analyzed and compression technique is used to analyze.
[11]Proposed paper presents the sequential fraud pattern mining to discover the rare abnormal fraud patterns within human body where experiments are conducted using data mining tool WEKKA. This show betterment in percentage for classification accuracy.

[12]in this paper a fraud pattern growth method is used that analyze the medical database to specify the combination of chronic abnormal fraud patterns it introduce prefix span algorithm that identify all possible fraud patterns in the images but it constrained only specific abnormal fraud patterns and can further improved for efficient search, it shows the results in terms of HTN and DP abnormal fraud pattern.

[13]In this paper analysis of various sequential fraud pattern mining algorithm are done. It discover the various challenges in these algorithm and improved the performance by proposing constraints in fraud patterns. It enhances existing CAI prefix span algorithm by introducing time constraints. The comparative results shows it is better and we can also further enhance it by applying efficient constraints.

[14]In this proposed an approach that verified and recommend clinical pathway to the patients it utilizes sequential fraud pattern mining that handles the record between various time intervals . the proposed methodology uses the actual logs of patients that would further analyze these fraud pattern using T-prefix span algorithm .but it will be necessary to introduce faster mining algorithm that are not in proposed methodology.

[15]In this paper a non-homogeneous mark over model is used to identify the chronic abnormal fraud patterns in the patients. The algorithm uses global optimization that efficiently identify the number of frequent pathway required to analyses the patient. The result shows that the proposed methodology probability is better than existing ones but this approach can be extended using admission scheduling policy.

III. METHODOLOGY

The proposed algorithm uses the prefix span algorithm for determining patterns which can be grouped together to form clusters. Pre-processing mechanism includes most probable value replacement with the missing value.

Algorithm

- Input: Dataset
- Output: Classification Accuracy, Abnormal patterns Prediction

---

- Input Dataset
  Data=$Dataset_i$
  Where I are the number of rows within the dataset
- Apply Pre-processing mechanism to resolve the missing values
  MPV=mean
  (Values(
  $Person_{id_i} = dataset(person_{id_i}))$
- Repeat while all the missing values are tackled
  If ($Missing_i$)
  $Missing_i$=MPV
  End of if
  End of loop
- Apply Similarity based random forest for pattern growth determination
- Form clusters
  Repeat until values in dataset are examined
  If($Datset_{iValue}$==$Dataset_{i+1value}$)
  $Cluster_i$=$Datset_{iValue}$
  End of if
  I=i+1
  End of loop
- Predict abnormal patterns looking at the pattern clusters
- Result: Accuracy, Abnormal patterns

The flow of the proposed work is given as under
The methodology to be followed must accommodate pre-processing mechanism. This pre-processing mechanism must eliminate noisy data. This noisy data may include missing data. In addition to missing data insignificant values must be eliminated in order to increase the execution speed. After that feature vector must be formed using Similarity based random forest approach. The proposed model that can further improve the result of similarity based approach is given in figure 8.

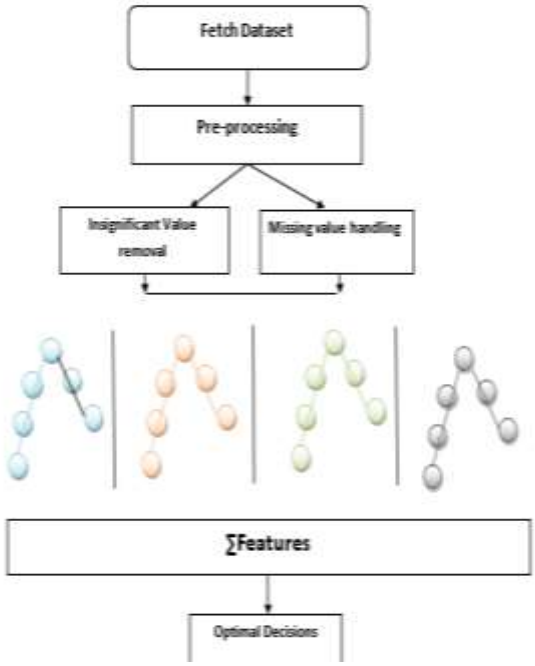

Figure 1: Proposed system to improve classification accuracy

Each tree figure 1 represents feature tree. Feature of similar types are grouped within the same cluster or group. This will form a feature vector to be compared against the test data to derive a conclusion.

The features that are extracted includes mean, median, mode, kurtosis, skewness, correlation, regression and entropy. All these features are extracted in two phases. First phase is of training and other phase is of testing. Trained features are compared against the test features and if match occurs than fraud is detected.

The working of the proposed mechanism is given in this section. (Caorsi and Lenzi 2016; C. Wang, Kennedy, and White 2017)This mechanism is most frequent in detection of financial fraud at early stage. This mechanism is applied on financial fraud detection. The mechanism used can be used in order to detect the financial fraud at early stage. Mathematical foundation for Similarity based random forest is given in this section.

A. Methametical Foundation

There are three actions associated with Similarity based random forest. Taking input, processing them and them generating output. Similarity based random forest is generally of the form given in figure 1. Convolution is generally represented with the mathematical symbol '*'. If input dataset is represented with 'X' and filter data is represented as 'F' then expression

$$Z=X*F$$

This 'Z' represents traversing of data cell by cell. The operation of convolution builds a matrix by multiplying contents with filtered matrix. Let us consider dataset of size 3x3 and filter of size 2x2.

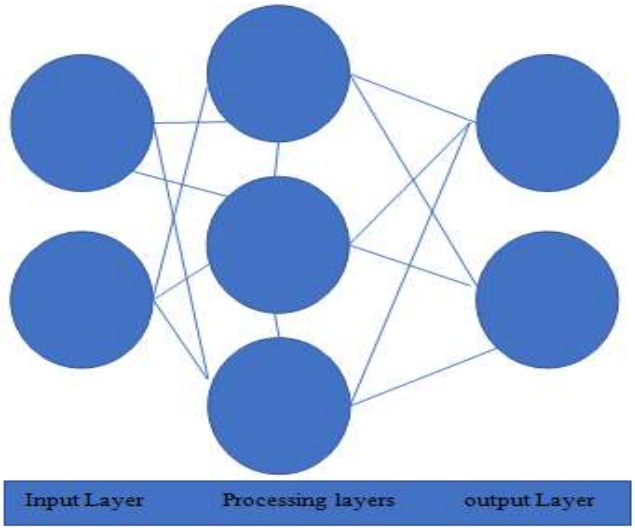

Figure 2: Convolution Neural network

Table 1: Dataset for operation

1	7	2
11	1	23
2	2	2

Table 2: Filter of size 2x2

1	1
0	1

The filter perform operation by moving through the patches of the dataset and values are summed up as given in following equations

$$(1*1+7*1+11*0+1*1)=9$$
$$(7*1+2*1+1*0+23*1)=12$$
$$(11*1+1*1+2*0+2*1)=14$$
$$(1*1+23*1+2*0+2*1)=26$$

The filter is filtering the data by considering only small chucks at a time. In case of large dataset, same operation is performed on every distinct patch. Since the considered size of the dataset is small so convergence is faster, in case size of the dataset is large then time consumption in convergence is more. Convolution layer extract useful features in 2D matrix form.

$$\begin{bmatrix} 9 \\ 12 \\ 14 \\ 26 \end{bmatrix}$$

The fully connected layer than plays its part. It takes the input that is generated from the input layer and processed by convolution layer. The linear and non linear transformation are then performed by fully connected layer.

The equation used for linear transformation is given as under

$$Z = W^T * X + b$$

'X' define input, 'W' defines weight, 'b' is a bias or constant. The size of the matrix is given as under

$$\begin{bmatrix} 9 \\ 32 \\ 14 \\ 26 \end{bmatrix} \quad \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix}$$

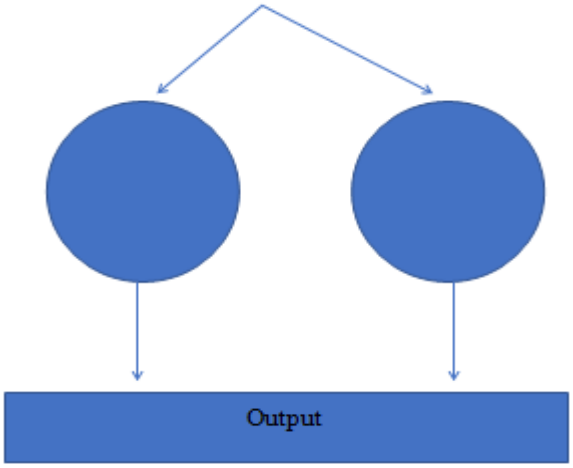

Figure 3: Size of the output generated with Similarity based random forest

The size of the matrix is equal to the amount of features extracted. There will be total of 4 features as per our example that is extracted and output layer resulted in the classification based on these features. Sigmoid function is applied and forward propagation is applied for the result generation.

In case result is not up to the mark then, back propagation is applied to adjust the weights and result is generated again. This process continues until result is up to the mark. The result is calculated on the basis of generated errors. In case error is more then back propagation becomes com [pulsar otherwise result is retained.

This mechanism first of all eliminates noisy data from the dataset. Noisy data from the dataset is eliminated by identifying missing data and then replacing it with highest frequency value. In addition, learning is based on Random Forest learning mechanism that uses the previous transaction impacts and hence converges much faster as compared to existing learning mechanism.

IV. RESULTS AND DISCUSSION

The performance of the system is analysed by the use of parameters such as accuracy, specificity and sensitivity. Accuracy is obtained by subtracting the actual result from the approximate result. In terms of predictions accuracy is obtained as

$$Accuracy = \frac{Correct_{Pre}}{Total_{Pred}}$$

Equation 1: Accuracy in terms of prediction

Sensitivity is obtained by dividing number of positive predictions to the total true positive rate.

$$\text{Sensitivity}=\frac{Correct_{Positive_{predictions}}}{Total_{Positives}}$$

Equation 2: Sensitivity evaluation formula

Specificity is another parameter used to evaluate correctness of the proposed system. It is given as under

$$\text{Specificity}=\frac{True_{Negitives}}{TP+FN}$$

Equation 3: Specificity obtaining formula

The abnormal patterns detection and prediction is given though accurate classification, result in terms of plots is given as under

Table 3: Number of abnormal patterns discovered

Dataset Size	Parameter	Base Paper	Proposed work
100 rows	Number of Abnormal Patterns	20	35
500 rows	Number of Abnormal Patterns	40	55

The result in terms of patterns is more in case of Prefix span algorithm as compared to FP growth algorithm but optimality will be tested only through the parameters such as accuracy, specificity and sensitivity.

Table 4: Comparison of result in terms of accuracy, sensitivity and specificity

Dataset Size	Parameters	Base (%)	Proposed (%)
5 Rows with 55 attributes	Accuracy	77	85
	Specificity	75	84
	Sensitivity	79	84
10 rows with 100 attributes	Accuracy	77	85
	Specificity	79	86
	Sensitivity	78	87
20 rows with 200 attributes	Accuracy	78	86
	Specificity	79	87
	Sensitivity	78	87

The plot for the above table is as under

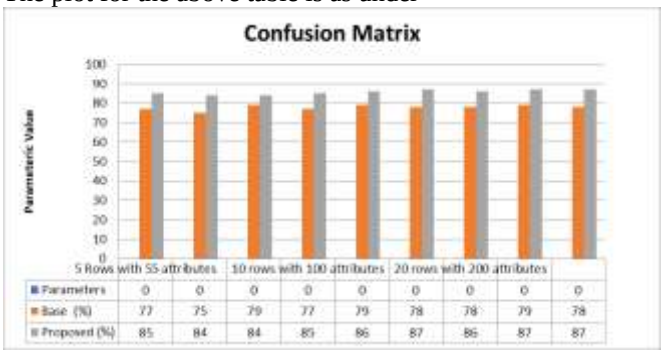

Figure 4: Plots for the accuracy, specificity and sensitivity

Classification accuracy of proposed system appears to be more as compared to existing techniques. Multiple class prediction mechanism showing higher accuracy proving the worth of study.

V. CONCLUSION AND FUTURE SCOPE

In this paper an automated system that utilizes MPV along with Similarity based random forest algorithm for detecting frauds proposed. Pre-processing phase is critical and is well defined using noise handling and resizing operation. Obtained dataset are fed into the trained network for feature extraction using Random Forest learning algorithm and classification is performed using MPV. Hybrid approach followed gives better results. The main objective of the proposed literature is creating optimized detection using Random Forest for better accuracy. Higher accuracy is achieved by the use of said literature. In future, proposed strategy can be examined against the real time datasets for better evaluation of accuracy.

REFERENCES

[1] N. Upasani and H. Om, "Evolving fuzzy min-max neural network for outlier detection," *Procedia Comput. Sci.*, **vol. 45, no. C, pp. 753–761, 2015,** doi: 10.1016/j.procs.2015.03.148.

[2] D. Wang, B. Chen, and J. Chen, "Credit card fraud detection strategies with consumer incentives," *Omega (United Kingdom)*, **vol. 88, pp. 179–195, Oct. 2019,** doi: 10.1016/j.omega.2018.07.001.

[3] R. Saia and S. Carta, "Evaluating financial transactions in the frequency domain for a proactive fraud detection approach," in *ICETE 2017 - Proceedings of the 14th International Joint Conference on e-Business and Telecommunications*, **vol. 4, pp. 335–342, 2017,** doi: 10.5220/0006425803350342.

[4] A. Pumsirirat and L. Yan, "Credit card fraud detection using deep learning based on auto-encoder and restricted Boltzmann machine," *Int. J. Adv. Comput. Sci. Appl.*, **vol. 9, no. 1, pp. 18–25, 2018,** doi: 10.14569/IJACSA.2018.090103.

[5] R. Saia, "A discrete wavelet transform approach to fraud detection," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **vol. 10394 LNCS, pp. 464–474, 2017,** doi: 10.1007/978-3-319-64701-2_34.

[6] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, **vol. 234, pp. 11–26, Apr. 2017,** doi: 10.1016/j.neucom.2016.12.038.

[7] I. Sohony, R. Pratap, and U. Nambiar, "Ensemble learning for financial fraud detection," in *ACM International Conference Proceeding Series*, **pp. 289–294, 2018,** doi: 10.1145/3152494.3156815.

[8] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for financial fraud: A comparative study," *Decis. Support Syst.*, **vol. 50, no. 3, pp. 602–613, Feb. 2011,** doi: 10.1016/j.dss.2010.08.008.

[9] E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," *Decis. Support Syst.*, **vol. 50, no. 3, pp. 559–569, 2011,** doi: 10.1016/j.dss.2010.08.006.

[10] F. Carcillo, A. Dal Pozzolo, Y. A. Le Borgne, O. Caelen, Y. Mazzer, and G. Bontempi, "SCARFF: A scalable framework for streaming financial fraud detection with spark," *Inf. Fusion*, **vol. 41, pp. 182–194, May 2018,** doi: 10.1016/j.inffus.2017.09.005.

[11] C. Wang and D. Han, "Credit card fraud forecasting model based on clustering analysis and integrated support vector machine," *Cluster Comput.*, v**ol. 22, pp. 13861–13866, Nov. 2019,** doi: 10.1007/s10586-018-2118-y.

[12] M. Zamini and G. Montazer, "Credit Card Fraud Detection using autoencoder based clustering," in *9th International Symposium on Telecommunication: With Emphasis on Information and Communication Technology, IST 2018*, **pp. 486–491, 2019,** doi: 10.1109/ISTEL.2018.8661129.

[13] E. Duman and M. H. Ozcelik, "Detecting financial fraud by genetic algorithm and scatter search," *Expert Syst. Appl.*, **vol. 38, no. 10, pp. 13057–13063, Sep. 2011,** doi: 10.1016/j.eswa.2011.04.110.

[14] N. Malini and M. Pushpa, "Analysis on financial fraud identification techniques based on KNN and outlier detection," in *Proceedings of the 3rd IEEE International Conference on Advances in Electrical and Electronics, Information, Communication and Bio-Informatics, AEEICB 2017*, **pp. 255–258, 2017,** doi: 10.1109/AEEICB.2017.7972424.

[15] I. Benchaji, S. Douzi, and B. Elouahidi, "Using Genetic Algorithm to Improve Classification of Imbalanced Datasets for Credit Card Fraud Detection," in *2018 2nd Cyber Security in Networking Conference, CSNet 2018*, **2019**, doi: 10.1109/CSNET.2018.8602972.

AUTHORS PROFILE

Miss shivani pursed Bachelor of Science from Beant college of Engineering and technology,Gurdaspur in 2018 and currently pursuing Master of Science from Swami Sarvanand Group of Institutes,DinaNagar Punjab since 2018. This is my first paper that iam publishing in international journals computer science and Engg.

Mrs Harjinder Kaur pursed Bachelor of Science from PTU campus, kapurthala and Master of Science in cse from BCET, Gurdaspur. She is currently working as Assistant Professor in Department of CSE, Swami Sarvanand Group of Institutes,DinaNagar Punjab.She has 12 years of teaching Experience in the Field of Computer Science and Department.

