Intrusion Detection using Attribute Subset Selector Bagging (ASUB) to Handle Imbalance and Noise

A.Sagaya Priya[1†] and Dr.S.Britto Ramesh Kumar[2††],
stleojohn@gmail.com
Research Scholar[1*], Assistant Professor[2]
Department of Computer Science, St. Joseph's College (Autonomous), Affiliated to Bharathidasan University,
Tiruchirappalli,Tamilnadu, India

Summary

Network intrusion detection is becoming an increasing necessity for both organizations and individuals alike. Detecting intrusions is one of the major components that aims to prevent information compromise. Automated systems have been put to use due to the voluminous nature of the domain. The major challenge for automated models is the noise and data imbalance components contained in the network transactions. This work proposes an ensemble model, Attribute Subset Selector Bagging (ASUB) that can be used to effectively handle noise and data imbalance. The proposed model performs attribute subset based bag creation, leading to reduction of the influence of the noise factor. The constructed bagging model is heterogeneous in nature, hence leading to effective imbalance handling. Experiments were conducted on the standard intrusion detection datasets KDD CUP 99, Koyoto 2006 and NSL KDD. Results show effective performances, showing the high performance of the model.

Keywords:
Intrusion detection; Ensemble modelling; Bagging; Data Imbalance; Noise

1. Introduction

Advancements in the computing technologies and the increased popularity and affordability of internet has resulted in these technologies become essential components of modern life [1]. Organizational activities like email and e-commerce transactions depend heavily on networks. Even individual users prefer online based transactions. Hence sensitive information is transmitted all over the network. Excessive dependence on the network has resulted in increased underlying security issues. Attacks on such networks causes irreparable damage [2]. Hence network security, otherwise known as cyber security has captured the attention of researchers. Firewalls, data encryption and authentication are the basic protection modes in traditional and even in the current networked systems. However, increasing sophistication of attacks signify the need for better and additional intrusion detection techniques [3].

Intrusion detection models can be generally classified into two; misuse based and anomaly based [4] models. Misuse based models captures the anomalous signatures from the networks. Traffic data are compared with these signals to identify similarities. This is used to effectively identify anomalous network transmissions. The major advantage is that the model is highly focused towards identifying anomalies, which can lead to better identification levels. However, the downside of the model is that new anomalous signatures can apparently go undetected due to the absence of similar data in the repository. Anomaly based intrusion detection models are based on creating repository with normal traffic signatures. These models can effectively identify intrusions, even new signatures. However, variation in normal traffic results in a large number of false positives (false alarms).

General intrusion detection models can be designed based on any one of these categories. These models work effectively but intrinsic issues in data complicates the prediction process and introduces bias in detection process. Major issues in data are imbalance and noise. Data is imbalanced if one of the classes exhibits a huge prominence over other existing classes. The existence of imbalance in the data results in providing good training on the majority classes. However, the minority classes, being less in number are not sufficient for training the classifier model. This results in a biased classifier and hence low prediction rates of minority classes. From a practical perspective it should be noted that minority classes correspond to anomalies or intrusions. So detecting them correctly is of higher significance in any classifier model [5]. Noise occurs due to incorrect entries [6]. Noise creates anomalies and hence introduces errors in the rule creation process. This work presents a heterogeneous ensemble model to provide effective handling of data imbalance and noise.

The remainder of this paper is structured as follows; section II presents the related works, section III presents the proposed ASUB model, section IV presents the results and section V concludes the work.

2. Related Works

Intrusion detection has been a highly researched domain since the inception of networks. However, as the size, usage and operational level of networks increase, challenges posed by the domain varies considerably. Hence this section discusses the most recent and significant works in this domain. Machine learning is one

Manuscript received May 5, 2022
Manuscript revised May 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.5.15

of the mostly used mechanisms for intrusion detection. An SVM based intrusion detection model was presented by Wang et al. [7]. It provides a framework that integrates SVM and feature augmentation as its major components for detecting intrusions in network transactions. The model is mostly based on creating high quality data for the machine learning process. A parallelized intrusion detection model was presented by Chellammal et al. [8]. This technique is designed as a real time based detection process that can also be applied on streaming data. A lazy learning algorithm for intrusion detection was presented by Chellam et al. [9]. The model modifies the lazy learning algorithm to eliminate the intrinsic complexity associated with it. A Spark based model aimed to detect intrusions in network traffic was presented by Dahiya et al. [10]. Two feature reduction methods; Canonical Correlation Analysis and Linear Discriminant Analysis were used for data preparation. Classification is performed using Naïve Bayes and Random Forest in Spark environment to provide real time predictions. Other similar models based on Big Data processing techniques include works by Marchal et al. [11] and Ahn et al. [12].

Artificial Neural Network based detection model for identifying intrusions in network traffic was presented by Shenfield et al. [13]. This work uses the deep packet inspection strategy for effective intrusion detection. A feed forward network was used for the process of packet analysis. A network ensemble based strategy for detection of intrusions was presented by Liu et al. [14]. Other similar such models include work by Wu et al. [15] that uses a regression neural network optimized by Artificial Immune System for effective detections. PCA based dimensionality reduction and ensemble model for network intrusion detection was presented by Salo et al. [16]. This model uses Information Gain based PCA for dimensionality reduction. The ensemble model has been constructed using SVM, Instance based Learning algorithm and multilayer perceptron. Other neural network based models include works by Carrasco et al. [17] and Kumar et al. [18].

Metaheuristic models are also on the raise in creating real time models for intrusion detection. A Firefly based intrusion detection model was presented by Selvakumar et al. [19]. This work is based on utilizing Firefly algorithm for feature selection and uses C 4.5 and Bayesian Networks for the intrusion detection process. A similar Firefly based model using Levy flights was presented for effective optimization by Yag et al. [20]. A swarm intelligence based strategy for intrusion detection was presented by Revathi et al. [21]. Other swarm based models for intrusion detection are PSO XGboost model by Jiang et al. [22] and SVM and PSO based model by Kalita et al. [23].

3. Attribute Subset Selector Bagging (ASUB) (Proposed Model)

Detecting intrusions is a major requirement in this highly networked world. The detection process is complex due to the intrinsic issues in the real-world data. The major issued are identified to be imbalance and noise contained in the data. Models proposed for intrusion detection should be capable of handling both these issues effectively. This work proposes a heterogeneous bagging model, the Attribute Subset Selector Bagging (ASUB). This model is composed of five major phases; data preprocessing, correlation based attribute selection, training data selection, heterogeneous bagging model creation and voting based combiner. The ASUB architecture is presented in figure 1.

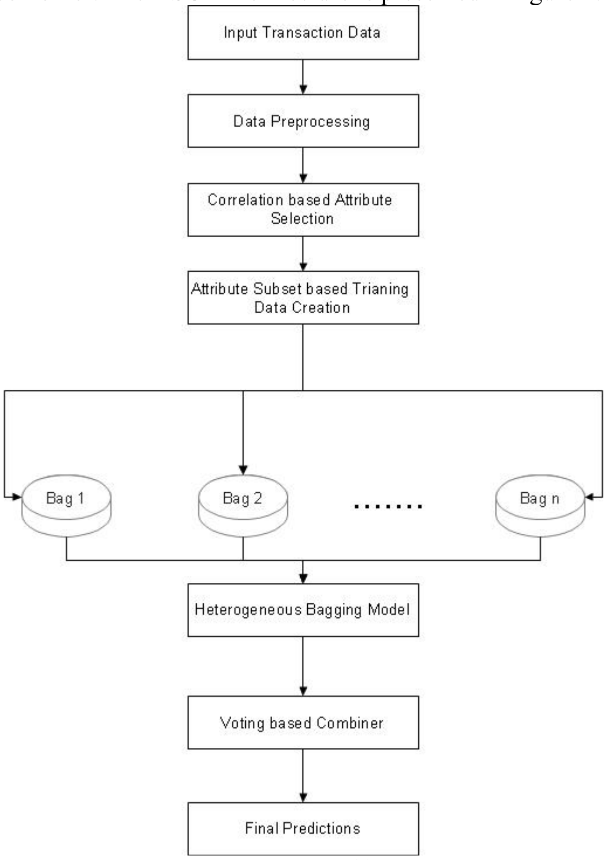
Input Transaction Data
Data Preprocessing
Correlation based Attribute Selection
Attribute Subset based Trianing Data Creation
Bag 1
Bag 2
....
Bag n
Heterogeneous Bagging Model
Voting based Combiner
Final Predictions

Figure 1: ASUB Architecture

3.1 Data Preprocessing

Data preprocessing is the initial and the mandatory phase of any machine learning model. Real time data are prone to inconsistencies. These include missing data, erroneous data etc. Such components are eliminated from the data. Apart from such types of data, most real-time data contains string and categorical attributes. Machine learning model cannot operate on such data, hence such attributes must be eliminated or encoded into numerical components. String attributes were identified and eliminated, while categorical components are encoded using One-Hot Encoding technique. Since the proposed model is generic and can be applied on any type of data, all these components are included as the preprocessing phase.

3.2 Correlation based Attribute Selection

Bagging models are based on performing instance selection as training data. This work modifies this process and uses attribute selection to create training bags. Attributes or data columns correspond to a property pertaining to the transmission. Each property has a different significance. Selection of attributes is based on the significance of the attribute. Significance is identified by the correlation of the attribute with respect to the class attribute. The data is mixed in nature and contains both numerical and categorical data. Categorical data prior to encoding is taken for this process. Correlation for numerical data is identified by Pearson Correlation Coefficient.

The correlation for an attribute pair x_i and $x_j$is given by

$$r_{x_i,x_j} = \frac{\sum_k x_{ik}, x_{jk} - n\overline{x_i}, \bar{x}_j}{\sqrt{(\sum_k x_i^2 - n\overline{x_i^2})(\sum_k x_j^2 - n\overline{x_j^2})}} \quad (1)$$

Correlation between categorical and numerical data is performed using Point Biserial correlation [24]. The process of calculating the correlation is given by

$$r_{pb} = \frac{M_1 - M_0}{s_{n-1}} \sqrt{\frac{n_1 n_0}{n(n-1)}} \quad (2)$$

Where s_{n-1} is the standard deviation for a population sample, M_1 and M_0 are the mean values for the numerical attributes under group 1 and group 0, n_1 and n_0 are the number of data points in group 1 and group 0. Imputation is performed based on the correlation between multiple attributes.

Absolute values for correlation is considered and the attributes are grouped according to the correlation levels. This work proposes three levels of grouping; low, medium and high. Low grouping considers attributes with less than 50% of correlation levels, medium grouping considers data between 50% and 75% correlation levels and high grouping considers all data above 75% correlation. Selection of attributes is based on the groups.

3.3 Training Data Selection

The ASUB model prepares training data by selecting a subset of the attributes. General bagging models considers subset of instances. Although this mode of operation is effective in handling data imbalance, noise contained in the data still prevails, as all the attributes are considered for the training process. The proposed model performs noise elimination by selecting subsets of attributes. Noise occurs when instance refers to inconsistencies in the data. Noise is classified into class noise and attribute noise. Class noise occurs when instances are provided wrong class labels. Attribute noise occurs due to corruptions in the attribute values. Class noise is rare and is almost an impossibility in manually annotated data. Since network intrusion data is labelled only after multiple references, the probability of occurrence of class noise is almost zero. However, occurrence of attribute noise is very common in data generated from automated methods. Creating subsets of attributes creates a division in the data, making a few data bags noise free. This enables the model to eliminate the noise component to a large extent.

The number of attributes to be selected is based on the dataset. However, the authors recommend a minimum of 60% of the attributes to be selected for each data bag. This enables attributes to be repeated in multiple bags, hence providing a link between multiple models. Selection of attributes is performed based on the attribute group. Attribute selection in this work is performed such that 90% of the attributes from the high group is selected, 60% of attributes in the medium group is selected and 20% attributes from the low group is selected. The percentages provided are flexible and can be varied according to the dataset, and is usually decided by the domain expert. The selection is performed in random. Hence, every selection phase picks up a different set of attributes. This enables variations in each of the models in the ensemble architecture. These variations control noise to a large extent.

3.4 Heterogeneous Bagging Model Creation

This work modifies the conventional bagging approach by introducing heterogeneous machine learning models as base learners. Conventional bagging approaches operate with homogeneous base learners. Diversity is introduced by providing varied subsets of data. However, for highly complex data with noise and data imbalance, this is not sufficient. Higher levels of diversity is required. Hence, the proposed work introduces heterogeneous base learners. Decision Tree, Random Forest and Logistic Regression are used as the base learners.

Decision Tree is a tree based model that operates by creating decision rules and branches based on conditions.

The model is effective on dynamic and imbalanced data. Random Forest is an ensemble modelling technique. The model creates multiple decision trees and combines their result to provide predictions. It is a string classifier model and enables effective reduction of variance in the prediction process. Logistic Regression is a statistical based modelling technique. This model is simple, fast, yet powerful and enables effective handling of bias contained in the data. The aggregation of such varied models enables the proposed ensemble to handle the data complexity and provide effective predictions. Further, the bagging architecture itself is capable of handling data imbalance to a large extent. Attribute based selection of data bags reduces noise to a large extent.

3.5 Voting based Combiner

The training data is passed to all the base learners. Each base learner is trained based on the partial data provided to it. As each data is composed of a different set of attribute, the trained models vary significantly from each other. During the prediction process, the test data is preprocessed and attribute segregation is performed based on the model. The preprocessed data is passed to each model. Each model provides prediction based on the training data provided to it. This results is a set of prediction for each instance. The required predictions however, needs to be a single prediction for each instance. Voting based combiner is proposed for this purpose. The prediction with highest vote is considered as the final prediction.

4. Results and discussion

The proposed ASUB model has been implemented using Python and coded using Jupyter Notebooks. Libraries in sklearn have been used for the creation of base learners. The proposed model has been analyzed using three standard intrusion detection datasets; KDD CUP 99, NSL-KDD and the Koyoto 2006 datasets.

Results are analyzed in terms of multiple performance metrics. Analysis of performance in terms of ROC curve is shown in figure 2. ROC curve is constructed with FPR in the x-axis and TPR in y-axis. Exhibiting a point in the top left quadrant is the requirement of a good classifier model. It could be observed that the proposed ASUB model exhibits low FPR levels and high TPR levels, with curves in the top left quadrant. This exhibits the high performing nature of the proposed model.

Performance of the ASUB model in terms of precision and recall is showed in the PR curve in figure 3. It could be observed that the curve is towards the top right quadrant, exhibiting high precision and high recall levels. This exhibits the high performing nature of the proposed model.

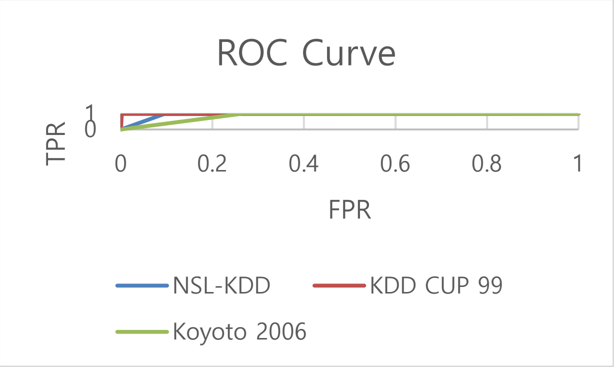

Figure 2: ROC Curve

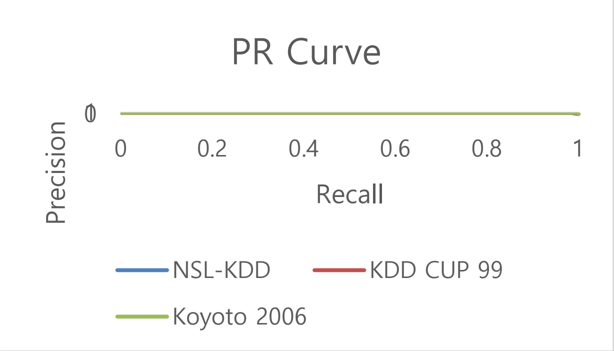

Figure 3: PR Curve

Performance in terms of other metrics is shown in table 1. It could be observed that the true prediction rate (TPR) is 98% to 99% on all the three datasets. TPR levels depict the detection level of intrusion data. Higher TPR levels indicate good detection rates. Further, the high accuracy levels also indicate effective performances.

Table 1: Performance of ASUB on KDD CUP 99, NSL-KDD and Koyoto 2006

Technique	***KDD CUP 99***	NSL-KDD	Koyoto 2006
TPR	0.990253	0.98091	0.998881
TNR	0.997947	0.905956	0.820755
Precision	0.998035	0.95702	0.979167
Recall	0.990253	0.98091	0.998881
F-Measure	0.994129	0.968818	0.988926
FPR	0.002053	0.094044	0.179245
FNR	0.009747	0.01909	0.001119
Accuracy	0.994	0.957	0.98
AUC	0.9941	0.943433	0.909818

5. Comparative Analysis

A comparison of the proposed model with the SVM based model proposed by Wang et al. [7] has been performed to identify the efficiency of the performance. Performance comparison is based on FPR, TPR and Accuracy metrics. Performance of the proposed model on KDD CUP 99 is shown in figure 4. It could be observed that the proposed model exhibits the lowest FPR levels and very high TPR and accuracy levels. This indicates high prediction on KDD CUP 99 dataset.

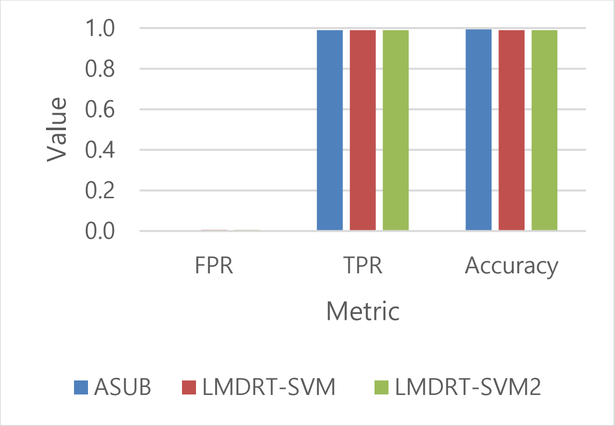

Figure 4: Performance Comparison on KDD CUP 99

Performance comparison on NSL KDD dataset is shown in figure 5. It could be observed that the proposed ASUB model exhibits slightly reduced performances at 0.08 on FPR, 0.01 on TPR and 0.03 on Accuracy. This reduction is attributed to the varied data distributions, leading to increased variance in the prediction process. However, the difference is found to be very low.

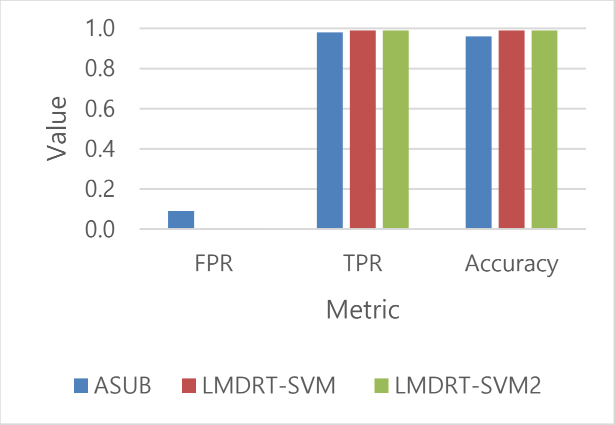

Figure 6: Performance Comparison on Koyoto 2006

Performance comparison of the ASUB model on Koyoto 2006 data is shown in figure 6. It could be observed that the proposed model exhibits similar TPR and Accuracy levels. However, an increase of 0.14 FPR level was observed in the graph. This depicts further scope for improvement in the ASUB model.

Conclusion

Detecting intrusions is one of the major needs of the today's networked world. This is especially due to the large amount of highly sensitive data passed through networks. The major issue in identifying intrusions is the imbalanced nature of the data and the large amount of noisy components contained in the data. This work presents an ensemble based bagging model, the Attribute Subset Selector Bagging (ASUB) for effective handling of noise and data imbalance. The proposed model performs attribute subset selection to create bags, enabling the model to handle noise effectively. Further, the heterogeneity introduced in the ensemble creation mechanism has resulted in providing a complex model that can handle network data. Experiments were performed using KDD CUP 99, NSL KDD and Koyoto 2006 datasets. Comparisons indicate high performance in most metrics. However, the ASUB model was observed to exhibit slight reductions in performance, which is attributed to the variation in data distributions. Future works will concentrate towards effectively handing the variations in data distributions for better predictions.

References

[1] C.F. Tsai, Y.F. Hsu, C.Y. Lin, W.Y. Lin, Intrusion detection by machine learning: a review, Expert Syst. Appl. Int. J. 36 (10) (2009) 11994–12000. http://dx.doi.org/10.1016/j.eswa.2009.05.029 .

[2] B. Luo, J. Xia, A novel intrusion detection system based on feature generation with visualization strategy, Expert Syst. Appl. 41 (9) (2014) 4139–4147. http: //dx.doi.org/10.1016/j.eswa.2013.12.048 .

[3] G.C. Tjhai, S.M. Furnell, M. Papadaki, N.L. Clarke, A preliminary two-stage alarm correlation and filtering system using som neural network and k -means algo- rithm, Comput. Security 29 (6) (2010) 712–723. http://dx.doi.org/10.1016/j.cose. 2010.02.001 .

[4] H.J. Liao, C.H.R. Lin, Y.C. Lin, K.Y. Tung, Intrusion detection system: a compre- hensive review, J. Netw. Comput. Appl. 36 (1) (2013) 16–24. http://dx.doi.org/10.1016/j.jnca.2012.09.004 .

[5] Somasundaram, A. and Reddy, U.S., 2017, June. Modelling a stable classifier for handling large scale data with noise and imbalance. In 2017 International Conference on

Computational Intelligence in Data Science (ICCIDS) (pp. 1-6). IEEE.

[6] Akila, S. and Reddy, U.S., 2016. Data imbalance: effects and solutions for classification of large and highly imbalanced data. Proceedings of ICRECT, 16, pp.28-34.

[7] An effective intrusion detection framework based on SVM with feature augmentation

[8] Chellammal, P., and Sheba Kezia PD Malarchelvi. "Real-time anomaly detection using parallelized intrusion detection architecture for streaming data." concurrency and computation-practice & experience 32, no. 4 (2020).

[9] Intrusion Detection in Computer Networks using Lazy Learning Algorithm

[10] Network Intrusion Detection in Big Dataset Using Spark

[11] Samuel Marchal, Xiuyan Jiangz, Radu State, Thomas Engel (2014) “A Big Data Architecture for Large Scale Security Monitoring” , Springer.

[12] Sung-Hwan Ahn, Nam-Uk Kim,Tai-Myoung Chung (2014) “Big Data Analysis System Concept for Detecting Unknown Attacks”, IEEE.

[13] Intelligent intrusion detection systems using artificial neural networks

[14] G. Liu, F. Hu, W. Chen, A neural network ensemble based method for detecting computer virus, in: 2010 International Conference on Computer, Mechatronics, Control and Electronic Engineering, Vol. 1, Aug 2010, pp. 391–393.

[15] J.Wu, D. Peng, Z. Li, L. Zhao, H. Ling, Network intrusion detection based on a general regression neural network optimized by an improved artificial immune algorithm, PLOS ONE 10 (3) (2015) 1–13.

[16] Dimensionality Reduction with IG-PCA and Ensemble Classifier for Network Intrusion Detection

[17] Unsupervised intrusion detection through skip-gram models of network behavior

[18] Kumar, V. D., & Radhakrishnan, S. (2014, April). Intrusion detection in MANET using self organizing map (SOM). In Recent Trends in Information Technology (ICRTIT), 2014 International Conference on (pp. 1-8). IEEE.

[19] Firefly algorithm based Feature Selection for Network Intrusion Detection

[20] X.-S. Yang,“Firefly algorithm, Levy flights and global optimization”, in: Research and Development in Intelligent Systems XXVI (Eds M. Bramer, R. Ellis, M. Petridis), Springer London, pp. 209-218 (2010)

[21] S. Revathi and A. Malathi, “Data Preprocessing for Intrusion Detection System using Swarm Intelligence Techniques,” International Journal of Computer Applications , Volume 75– No.6, August 2013

[22] Jiang, H., He, Z., Ye, G. and Zhang, H., 2020. Network Intrusion Detection Based on PSO-Xgboost Model. IEEE Access.

[23] Kalita, D.J., Singh, V.P. and Kumar, V., 2020. SVM Hyper-Parameters Optimization using Multi-PSO for Intrusion Detection. In Social Networking and Computational Intelligence (pp. 227-241). Springer, Singapore.

[24] Point Biserial Coefficient (Keith Calkins, 2005)

A. Sagaya Priya is presently pursuing her doctor of Philosophy in Department of Computer Science. St. Joseph’s College (Autonomous), Tiruchirappalli, Tamil Nadu India. She received her M.Phil Degree in St. Joseph’s College (Autonomous), Tiruchirappalli. She received her BCA and MCA degree from Holy Cross College (Autonomous), Tiruchirappalli, Tamil Nadu India. Her research interest areas are Big Data, Networks and Cloud.

Dr. S. Britto Ramesh Kumar is working as Assistant Professor in the Department of Computer Science. St. Joseph’s College (Autonomous), Tiruchirappalli, Tamil Nadu India. He has published many research articles in the national/international conferences and journals. His research interests are Software Architecture, Security, Web Service