ISSN 2347 - 3983

Volume 8. No. 2, February 2020

International Journal of Emerging Trends in Engineering Research

Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter45822020.pdf

https://doi.org/10.30534/ijeter/2020/45822020

A Comparative Analysis on Intrusion Detection System for SDWSN using Ensemble Classifier

Indira K[1], Sakthi U[2]

[1]Sathyabama Institute of Science and Technology, Chennai, India, indira.it@sathyabama.ac.in

[2]St.Joseph's Institute of Technology, Chennai, India, sakthi.ulaganathan@gmail.com

ABSTRACT

Security in Software Defined Wireless Sensor Network (SDWSN) is current and important area of interest amongst researchers because WSN is easily prone to vulnerabilities due to open transmission medium. Software-defined networking was described as a solution for many WSN problems relating to efficiency and reuse of resources. The SDN architecture, on the other hand, is exposed to new security threats such as false flow request attacks, false data flow forwarding attacks and false neighbor information attacks etc. These new security threats cause dramatic changes in performance metrics such as data and control packet delivery ratio, delay, throughput, energy consumption and controlpacket overhea, etc. Very essential is the development of an efficient intrusion detection and prevention system to protect the network from security threats and improve network performance. In this paper, we proposed an intrusion detection system using ensemble classification to mitigate attacks and we analyzed different ensemble classification approaches and their performance.

Key words : SDWSN, Random Forest, IDS, Ensemble Classifier.

1. INTRODUCTION

Wireless Sensor Network (WSN) is known to be a distributed, organizing-themselves network comprising a huge collection of sensor nodes and utile devices that transmit and receive data via the radio channel with one another. This network is based on a large number of small-sized sensor motes, which be composed of collecting data and processing collected data modules and a transceiver module, working together. WSN's main disadvantage is its sensor resource limitations, which are the level of memory, speed, energy, and transmission. Network resources smart management will help us address such a weakness. As a very difficult task this makes perfect network and smart operation.

In addition, the introduction of new emerging technologies such as the IoT and other invented technologies in recent years has very high demand from wireless sensor networks as well as research activities in the area of sensor based networks. Software Defined Network (SDN) is one of most popular recent network architecture built to tackle the essential drawbacks as well as insufficiency of flexibility that the existing conventional network paradigm is facing. SDN is a new network paradigm in which network controlling and administration is made simpler and also permits dynamic monitoring, alteration and management of actions via an approach named network programmability. SDN's development goals were to simplify creativity and network management operations and control programmability. SDN works especially adopt the paradigm which segregates control plane from its data plane. It works on paradigm via an interface named the OpenFlow that brings data separation and control plane through. SDN is strong and popularly used technology in recent years and has achieved universal applications.

As SDN paradigm integrates WSN, it expresses itself in the context of computer networks to a new network model called the Software Defined Wireless Sensor Network (SDWSN). It is one among the foremost popular paradigms that can be applied to enhance WSN's performance, profitability, promote interoperability with other kind of networks and play crucial role in the emerging IoT. Though, SDN extended WSN environment considerable security features, still it is having several security issues. Due to the insufficiency of important security modules, such as middleware and TLS, SDWSN is highly exposed to security threats. The centralized controller often serves a single point, so it is open to vulnerabilities like DoS attacks and other threats. Efforts, though, have been to weaken some of the possible SDWSN threats, still attacks are not thoroughly diminished. SDN and WSN may adopt some of the attacks from their own paradigms. The attacks adopted from SDN are malefic data flow, device forwarding threats, Denial of service on control layer and deficiency of trust in the middle of controller and management applications.

Thus, IDS has been among the top concerns in the research community. The focused issue is early phase identification and protection from cyber-attacks. IDS is used to find intended threat activities by creating malefic traffic that attempts to get unauthorized accessibility gain or originate misfiring in the network. The familiar measures on which the efficiency and success percentage of detection system can be computed are anomaly detection speed, accuracy and dependability. There was an effort to improve the precision of the detection based on ML techniques [9][15][21]. The current mechanisms are concentrates on Machine learning based detection system being deployed to the SDN architecture.

Thus, the security of the network can be further increased by using Machine Learning methods in the SDN paradigm [2]. Work for the convergence of machine learning methods with IDS has been completed. A Random Forest is constructed to categorize the benign and malefic traffic characteristics, and this anomaly detection approach is implemented in OpenFlow switch using a controller. The ML construct is also used to derive traffic characteristics and perform classification tasks in real time.

We propose a distinct intrusion detection method for detecting different kind of threats with high precision and performance using Random Forest as base classifier. Next, Salp Swarm Algorithm (SSA) is used to optimize the dataset by selecting important features. Nevertheless, the difference between regular and malefic traffic flow has a unresponsive impact on the accuracy of threat detection, which is distinct from classification or detection in other fields.Our approach then uses ensemble classification to improve the ADR and b ringdown the bias and variance between various training datasets, where findings from multiple classifiers are integrated based on a voting algorithm, in order to contr ol the issue of class imbalance. To increase reliability and accuracy of the detection system with minimal computational complexity and processing time complexity optimized dimensionality reduction and ensemble classification are integrated in this way. Finally try with many ensemble classifications by applying different ML techniques and made a comparative analysis among all possible ensemble combination.

The most important contributions to our research are abstrac ted as follows:

- We suggested a new approach that integrates the benefits of optimized dimensionality reduction and ensemble classification with the objective of constructing efficient and accurate attack detection.
- We implement an ensemble approach to improve classification efficiency on unbalanced datasets by integrating findings from multiple classifications (base is RF-SSO classifier and any one of the ML approach) into one. Meanwhile, a voting approach is used to construct the multi-class grouping.
- A proposal is made for a comparative performance a nalysis between several ensemble classifications on NSL-KDD dataset and CIC-IDS2017 data set.

The rest of the paper is sectioned as follows. In Section II, we study the background data on the IDSs and described our objectives. The proposed methodology is described in Section III, while in Section IV we presented the outcomes of the experiments and a comparative review is carried out on our proposal. At last, the conclusion is provided in Section V.

2. RELATED WORK

In this part, we will present a short description of the previous Random Forest work and a comparative performance analysis of the different kind of ensemble classification algorithms. In Intrusion Detection System (IDS) using Random Forest (RF) algorithm to find attacks by reducing the dimensionality and acquired cross validation to verify correctness of algorithm used. Nabila Farnaaz [3] proposed Random Forest approach is evaluated using NSL KDD data set and compared random forest modelling with j48 classier in terms of accuracy, DR, FAR and MCC. RF classifier is out performing than j48 classifier. Besides, et al. in [4] suggested a data optimization method to construct IDS, titled DO IDS. iForest is used in data sampling to sample data, and GA and RF integration is used to optimize sampling ratio. In the selection of features, the GA and RF integration is used again to choose the best subset of features. Classification is done using RF to construct IDS. DO IDS was tested using UNSW-NB15 dataset for intrusion detection.

In [5], two data sets represented the industrial environments are examined for Threat-based anomalies.SVM is used to find seven various types of attacks, and 35 various subtypes. This strategy is positive with accuracies and F1-scores of up to 92.5% and 85.2 percent respectively. Lots of missing data does not affect the performance of the algorithms in this data set. The second set of data, DS2 comprise an OPC UA-based traffic derived from hardware in real-world, a compact workstation Festo Didactic MPS PA. One class SVMs are used in this case, since there are only two instances of threats present. This results in 90.8 per cent slightly worse accuracy. Nonetheless, the f1-score performs best with 94.9 per cent due to an almost perfect recall.

Machine learning approaches, such as SVM and Random Fo rests, can increase the detection capabilities of popular indus trial IDSs. Huijun Peng (2018) et.al provides flow detection method based by SDN, develops anomaly SDN flow detection structures and performs flow classification detection for K-nearest neighboring algorithms using transductive confidence machines double P-value. The experiment demonstrate results that perhaps the presented algorithm reaches higher accuracy, a lower false positive rate and better adaptation SDN setting than other related algorithms.

In [7], they presented a hybrid aesthetic system for effective intrusion detection for service provider by utilizing the classification and optimization algorithms to enhance intrusion detection system performance. The experimental study conducted on NSL-KDD dataset found this methodology significantly increase the overall system efficiency when relative to the system performance with KNN classifier based IDS system.

3. PROPOSED METHODOLOGY

Proposed intrusion detection system is deployed in OpenFlow switch using centralized SDN controller as shown in Figure 1.

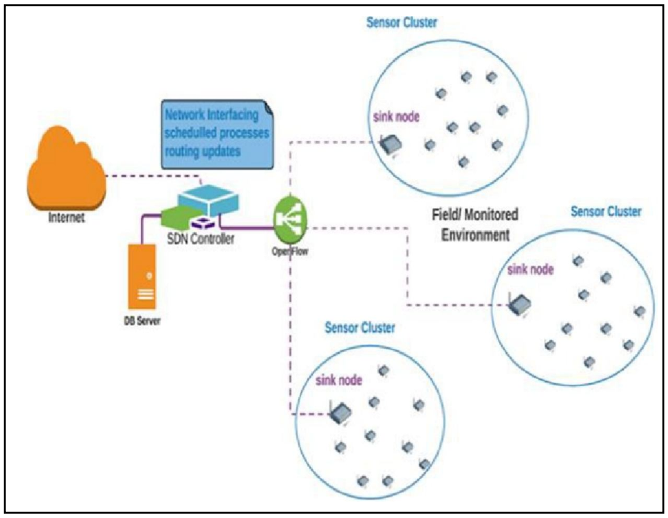

Figure 1: Architecture of SDWSN [22]

Machine learning is the method of detecting informative patterns from data and learning from those patterns without any specific instruction. All the ML algorithms are designed mathematically and optimized. In addition, ML as well as used to construct more adaptive classification based IDS. We applied a few of those algorithms to classify intrusions and made comparative analysis. Random Forest is a learning model for an ensemble that takes tree choice as a fundamental classifier. As the name suggests, with an amount of trees, this algorithm produces the forest. If more trees in the forest, it appears the more robust forest. Similarly, in forest greater amount of trees provides the outcomes of elevated precision in the random forest classification. When entering a sample to be categorized, the final outcome of classification is determined by a single decision tree's output vote. Random forest overcomes decision trees over-fitting issue, has excellent noise and anomaly values tolerance, and has excellent scalability and parallelism to the issue of high-dimensional data classification. In contrast, random forest is a data-driven, non-parametric method of classification. It trains rules of classification through sample learning and does not involve previous classification understanding.

The model of random forests is based on forests of K choice. Each tree votes on which class a specified independent variable X belongs to, and the class it deems most suitable is given only one vote. The K decision trees description is as follows:

$$\{h(X,\theta_k), k=1,2,........,k\} \qquad (1)$$

K represents amount of decision trees in random forests, among them. θ_k reflects random vectors that are autonomous and identically distributed. To enhance the IDS by Random forest detection efficiency, in [7] they presented a hybrid Intrusion Detection System (IDS). It is based upon optimized machine learning algorithm. In order to detect attack, suggested to use a hybrid IDS technique using Salp swarm optimized random forest classifier. in order to detect attack, in the paper it is suggested to use a hybrid IDS technique using Salp swarm optimized random forest classifier. An SSO technique to ensure that ideal features are selected for the intrusion detector, in addition to enhance the detection efficiency of the Random Forest (RF) classifier, SSO is used for optimization.

In random forest when constructing individual trees, to select the type of attack by split on randomization is applied. But in this work instead of randomly selecting the features we make use of Salp Swarm Optimization algorithm to find an optimal dataset with best features. First choose a set of data from the dataset and optimal features of the selected data is selected by using the SSO algorithm. The entropy of each feature is calculated by using equation (2)

$$E = \sum_{\forall c} p(c) \ln \frac{1}{p(c)} \qquad (2)$$

Assume that if there is an data i, feature j is used to define the split quality, which is stated as

$$Q(i,j) = \exp\{-(E_i + E_r)\} \qquad (3)$$

On the basis of rate of information in each feature, a feature with the most classification ability is selected among the k features to split the type of data with most important and unimportant features until the decision tree grows to the maximum. To find important features SSO algorithm initially marks top features as ‘important’ and rest of the features as ‘unimportant’ from the ranked list. Every iteration these sets of features are updated. The generated new set of data with best features θ_k , {h(X, θ_k), k=1,2,......k} are the input to the random forest classifier, random forest is used to classify new optimized dataset. To detect the attack the final categorization solution are decided by the number of votes of the tree classifiers.

3.1 Ensemble Classification

Responsibility of IDS is mandatory to detect benign and malefic traffic, and also very essential to deduce the distinct exact class of threats take place in the secured system.

The classification methods of the ensemble [24] usually combine in some way several complex, dysfunctional, and strong classificatory. Through constructing and integrating multiple independent models, these classifiers are effective in solving the same problem and jointly achieving a forecast outcome with higher consistency and precision. The traditional scenarios for the use of ensemble classifiers aproblem of representation, statistical logic and computational logic. For the first case, it is not appropriate to find the best representation in the the best representation in the hypothesis space sometimes by a single classifier.

For the second case, if the input data set isn't sufficient to tra in the learning algorithm, a single classifier may result in a weak result. In the last example, a problem can arise when a suitable hypothesis is too computationally time consuming f or an individual classifier to construct.

Bagging[10] and Boosting[11] are the two most common algorithms in ensemble learning, generating usually good classification results and being commonly chosen to construct many ensemble models. In addition, Voting [12], Bayesian parameter averaging [13], and Stacking [14] are the other well-known ensemble learning methods for improving classification efficiency. Similarly, ensemble approaches have been shown in many use cases to improve accuracy including intrusion detection. Ensemble classifiers provide tools for security professionals to assist in analyzes such as similarities to actual identified abnormal or normal activities. The ensemble classifier in this work contains two distinct classifiers. Always one of base classifier is RF-SSO is integrated with another base classifier to construct meta classifier using voting algorithm. It is proposed to enhance the forecasting performance of detection system.

Random Forest is a strong one among many algorithms explicitly for detection of attacks. Nevertheless, its performance measures can be enhanced remarkably until combined with other classifiers. Multi-Classifier [8] is a kind of intelligent hybrid system that integrates two or many algorithms of classification to generate the desired model. Additionally, Multi-classifier can be branched in to two types serial and parallel Multi-classifier.

Multi-Classifiers are used to generate any classification model in the best form. Stacking is used for the Serial Multi-Classifier implementation. Stacking combines multiple classifiers that are created by different algorithms in machine learning. It is a two-phase operation, creating a collection of base classifiers in the first step and then combining these base classifiers to create a meta-classifier in the second phase.

SVM [25] is a supervised category of machine learning algorithms that can be used to regress, identify, and detect outliers. SVM Classifier uses the hyperplanes definition for the classification of data. A hyperplane is a subspace of vector space which has one smaller than vector space dimensions. The ultimate distance between two different planes is determined by an ideal hyperplane. Support vectors are very essential data points in hyper planes. A hyperplane H in R_n can represent as in [16]:

$$H=\{x:a^{x} = b\} \tag{4}$$

Where a is an element of R_n, a!=0 and b is an element of R are given. SVM works well with linear and also nonlinear datasets. SVM needs more training time [17].

BayesNet is a structured network graph with collection of probabilities. BayesNet classifiers are mainly operates on dataset which contains missing values. Forecast the missing values based on Bayes Theorem which can be represented in [18] [23]. In that H is hypothesis, E is evidence and c is the background information. This classifier operates extremely with rough data sets.

Vote is a Meta algorithm which performs the decision process by applying several classifiers. It uses the power of different individual classifiers and applies a voting algorithm for the final decision. In this paper, applying voting algorithm to determine majority voting among number of classes and number of classifiers to make decision, where the class label is identified on the basis of highest value of the majority voting.

4. RESULTS

The performance of detection system is analyzed based on its ability of classify the traffic flow into a specific class. The Intrusion Detection System for SDWSN using Ensemble Classifier has been analyzed using the NSL-KDD and CIC-IDS2017 datasets. We have done comparative analysis of performance measures every classifier based on Accuracy, Detection Rate (DR), precision, recall and F-measure.

Accuracy is one among the major measure to describe the performance of any mechanism. It represents the degree to which an algorithm can exactly forecast the normal and abnormal activities. Calculation of accuracy, Detection Rate, Precision, Recall, and F-Measure is by the following given formula:

$$ACC = \frac{TP + FN}{TP + TN + FP + FN} \tag{5}$$

$$DR = \frac{TP}{TP + FN} \tag{6}$$

$$precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

$$F - Measure = \frac{2 * Recall * Precision}{Recall - Precision} \quad (9)$$

The above measures are calculated compared with RF-SSO alone, RF-SSO with SVM, and RF-SSO with BayesNet. Compare to all RF-SSO with SVM performing efficiently in both the datasets NSL-KDD and CIC-IDS2017 and shown Figure 2 and Figure 3.

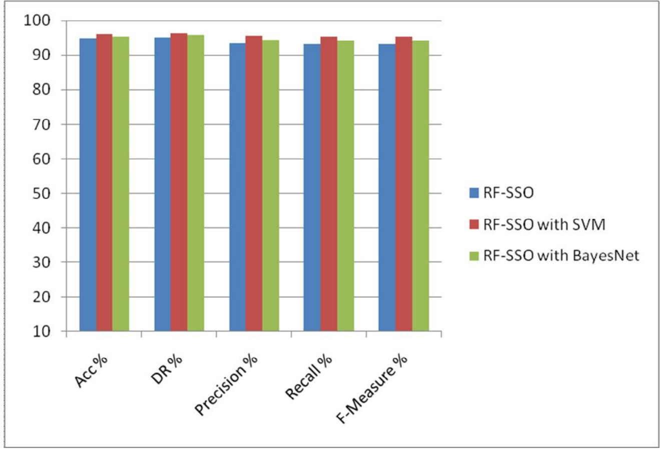

Figure 2: Comparative performence analysis on NSL-KDD

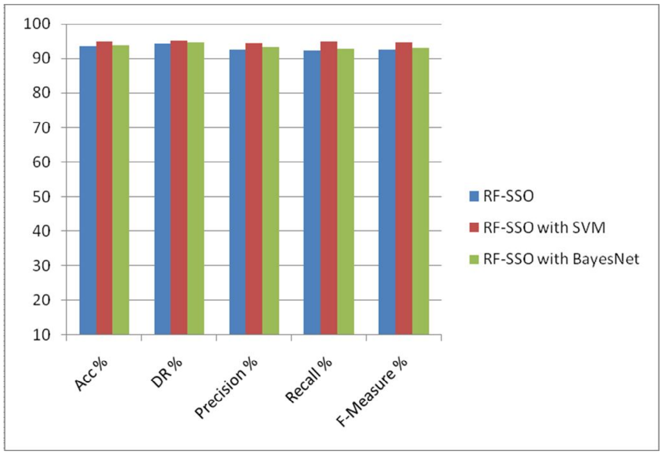

Figure 3: Comparative performence analysis on CIC-IDS2017

5. CONCLUSION

For the problem of network security, we motivated to implement a machine learning algorithms especially take Random Forest as base classifier to find attacks in the network. We have been determined major performance measures of RF-SSO and its integrated classifier. We have used NSL-KDD and CIC-IDS2017 datasets to compare the performance measures of every ensemble classifiers. The main objective of this work is to determine the good Multi-Classifier method outperforming Random Forest classification. We analyzed and determined that not all classifiers improve the performances of RF-SSO. Combination of RF-SSO and SVM is yield the better performance because RF is a best classifier to work on critical classification operations. In the future work, to improve ability of detection mechanism by finding exceptional threats in the higher volume of traffic flow.

REFERENCES

1. Gustavo A. Nunez Segura, Cintia Borges Margi and Arsenia Chorti, "**Understanding the performance of software defined wireless sensor networks under denial of service attack**", *in Open Journal of Internet of Things, Volume5, Issue1, 2019.*
2. Indira K, Ajitha P, V Reshma ; A Tamizhselvi, "**An Efficient Secured Routing Protocol for Software Defined Internet of Vehicles"** *in International Conference on Computational Intelligence in Data Science (ICCIDS),* 2019. https://doi.org/10.1109/ICCIDS.2019.8862042
3. Nabila Fanaaz an M.A. Jabbar, "**Ranom Forest Modeling for network intrusion detection system**", in *12th International Conference on Information Processing (IMCIP) 2016.*
4. Jiadon Ren, Jiawei Guo, Wang Qian, Huang Yuan, Xiaobing Hao, and Hu Jingjing, "**Building an effective intrusion detection system by using hybrid data optimization based on machine learning algorithms**", *in Journal of security and communication networks,2019.*
5. Simon D.Duque Anton, Sapna Sinha, Hans Dieter Schotten, "**Anomaly-based intrusion detection in industrial data with SVM and Random Forests**", *in 27th International Conference on Software, Telecommunication and Computer Networks (SoftCOM), IEEE, 2019.*
6. H. Peng, Z. Sun, X. Zhao, S. Tan and Z. Sun, "**A Detection Method for Anomaly Flow in Software Defined Network**", *IEEE Access, vol. 6, pp. 27809-27817, 2018.*
7. Indira K and Sakthi U, " **An efficient anonymous authentication scheme to improve security and privacy in SDN based wireless sensor networks**", *in Indian Journal of Computer Science and Engineering, 2020.*
8. M. Wozniak, M. Graa and E. Corchado, "**A survey of multiple classifier systems as hybrid systems",** *in Information Fusion, vol. 16, pp. 3-17, 2014.*
9. Indira K, Christal Joy E, "**Prevention of Spammers and Promoters in Video Social Networks using SVM-KNN**", *International Journal of Engineering and Technology, Vol 6, No.5, pp. 2024-2030, 2014.*
10. L. Breiman, "**Bagging predictors,"** *Machine learning, vol. 24, no. 2, pp.123–140, 1996.* https://doi.org/10.1007/BF00058655

11. Y. Freund, R. E. Schapire et al., "**Experiments with a new boosting algorithm**," *in ICML, vol. 96. Citeseer, 1996, pp. 148–156.*

12. J. Hu, "**An approach to eeg-based gender recognition using entropy measurement methods**", *Knowledge-Based Systems, vol. 140, pp. 134–141, 2018.*

13. K. Friston, K. Stephan, B. Li, and J. Daunizeau, "**Generalised filtering**", *Mathematical Problems in Engineering, 2010.*

14. C. Hung and J.-H. Chen, "**A selective ensemble based on expected probabilities for bankruptcy prediction**," *Expert systems with applications, vol. 36, no. 3, pp. 5297–5303, 2009.*

15. Indira K, Christal Joy E, "**Energy Efficient IDS for Cluster-Based VANETS**", *Asian Journal of Information Technology, vol 14(1), 2015, 37-41.*

16. K. Lipkowitz, "**Reviews in computational chemistry**". *Chichester: Wiley, 2007.*

17. S. Mulay, P. Devale and G. Garje, "**Intrusion Detection System Using Support Vector Machine and Decision Tree**", *in International Journal of Computer Applications, vol. 3, no. 3, pp. 40-43, 2010.*

18. D. Heckerman, "**Data Mining and Knowledge Discover**", *vol. 1, no. 1,pp. 79-119, 1997.* https://doi.org/10.1023/A:1009730122752

19. C. Catal and M. Nangir, "**A sentiment classification model based on multiple classifiers**," *Applied Soft Computing, vol. 50, pp. 135–141, 2017.*

20. Yuyang Zhou, Guang Cheng, Shanqing Jiang an Mian Dai, "**An efficient intrusion detection system based on feature selection and ensemble classifier**", *in ArXiv Journal ,2019.*

21. Indira K, Gomathi R.M, Anandhi T, Sivasangari A, Brumancia,"**Faulty node detection using RBM in underwater sensor networks**", *in International journal of advanced research in engineering and technology", Volume10, Issue5, 2019.*

22. lKgotlaetsile, MathewsModieginyane, Babedi BettyLetswamotse, RezaMalekian, and Adnan, "**Software defined wireless sensor networks application opportunities for efficient network management: A survey**", *in Journal of Computers and Electrical Engineering 2018.*

23. Hennadii Khudov , Irina Khizhnyak , Fedor Zots , Galina Misiyuk , Oleksii Serdiuk, "**The Bayes Rule of Decision Making in Joint Optimization of Search and Detection of Objects in Technical Systems",** *in International Journal of Emerging Trends in Engineering Research, Jan 2020.* https://doi.org/10.30534/ijeter/2020/02812020

24. M.Govindarajan, "**Ensemble of Classifiers in Text Categorization**", *in International Journal of Emerging Trends in Engineering Research, Jan 2020.*

25. N. Chandra Sekhar Reddy , Purna Chandra Rao Vemuri , A. Govardhan, "**An Emperical Study on Support Vector Machines for Intrusion Detection**", *in International Journal of Emerging Trends in Engineering Research, 2019.*