Vol 6, No 2, July - December 2021, pp: 12 - 21

Faculty Publications
ISSN: 3008-1017
Title DOI: https://doi.org/10.62458/021024

Machine Learning and Deep Learning approaches in Network Intrusion Detection

Anil K. Makhija*

ABSTRACT

This paper presents multiple hybrid models for Intrusion detection systems (IDS). Some of the proposed models use combination of information-gain based feature selection followed by classification using Random Forests and Naïve Bayes algorithms. Some of the proposed model use combination of expectation-maximization based clustering, information-gain based feature selection and then feed forward neural network with the backpropagation training algorithm. NSL-KDD dataset has been used to train and validate the model and NSL-KDD Test dataset is used to test the accuracy, precision, recall and F1-score of each of the proposed model. Performance of the proposed models is also compared with performance of Random Forests and Naïve Bayes based classification. The experimental results on the model that uses combination of expectation-maximization based clustering, information-gain based feature selection and then feed forward neural network showed promising results on detecting the intrusion when tested on NSL-KDD Test dataset.

Keywords – Network Intrusion Detection Systems, Deep Learning, Artificial Intelligence, NSL-KDD, NIDS, Artificial Intelligence in Network Security

1. INTRODUCTION

Pervasiveness of technology into our daily lives has transformed the world into a connected-world. Technology has increased connectivity and ease of doing business, helping create global reach both of organizations as well as customers. The same technology, that brings this paradigm of ever connected world is also being leveraged by users with malicious intent to attack both businesses and individuals. This threat of attack is managed by installing intrusion detection systems in the organizational networks. The nature and characteristics of attack is evolving continuously, and thus there is a need to create evolving and self-learning and self-improving intrusion detection systems to protect both businesses and individuals. Advances in machine learning and deep learning have created interest amongst researchers worldwide to design and model self-learning and self-evolving intrusion detection systems that provide improved detection rates and reduced false alarm rates. Many of these intrusion detection systems are based on machine learning approaches, with some of the recent researchers applying deep learning models as well. One of the limitations of these models is the dataset used for training and testing the model. This research proposes multiple hybrid models – some of the proposed models use combination of information-gain based feature selection followed by Random Forests and Naïve Bayes based classification and some of the proposed models use expectation-maximization clustering, followed by information gain feature selection, followed by deeply connected feed-forward neural network layers. The models are trained and validated on NSL-KDD train dataset and tested on NSL-KDD test dataset. Performance of each of the model is compared. The model that leverages expectation-maximization clustering, followed by information gain feature selection, followed by feed-forward neural network layers provides the highest level of accuracy, precision, recall and F1-score on NSL-KDD test dataset.

2. RELATED WORK

Cyberspace Challenges:

Technology has become an integral part of our day-to-day life. Technology helps increase both the speed of information retrieval and the reliability (ŞAHİN, 2018). Now-a-days, almost all businesses are leveraging technology to enable efficiency and effectiveness. Technology, coupled with the power of internet and cloud computing, is helping businesses

* Anil K Makhija, B.E., PGDIM, MBA. Lecturer, CamEd Business School.
Email: anil@cam-ed.com@cam-ed.com

create new value for themselves as well as for their customers (Avram, 2014). Almost every business today leverages some sort of online presence, be it through their proprietary platforms or platforms that are provided by a third party, including but not limited to social media platforms. Consumers have also become more technology savvy and they expect an interactive, collaborative, and personalized experience. Rapid proliferation of social media, and business platforms created on those social media has provided new mode of communication and interaction. This is helping organizations deliver new value to customers with a personalized experience (Baumöl & Jung, 2016).

This growth of technology, and pervasiveness of internet has propelled many businesses worldwide to provide online commerce channels to their customers, as part of their omnichannel strategy. Internet has helped businesses increase their customer reach as well as helped reduce the time it takes to reach the customers.

Technology, while a great business and customer experience enabler, opens new set of risks as well for the businesses. In addition to users who utilize technology to avail services, there are users with malicious intent as well. Cyberspace, defined as interconnected network of information technology and information on these networks, is subject to both malicious use and multiple types of destructive attacks (Bou-Harb et al., 2014). These users exploit vulnerabilities of systems to launch attack to cause both non-financial and financial harm to the organizations / businesses (Bawany et al., 2017). In any system, software bugs, configuration defects and design flaws create vulnerabilities. Those vulnerabilities or weaknesses are exploited by attackers to gain unauthorized access. The users with malicious intent use multiple techniques to gain unauthorized access. Some of those techniques are, for example, gaining or elevating access through network sniffing, session hijacking, DNS cache poisoning, buffer overflows, password cracking, and worm infection, application-level attacks, cross-site scripting, and denying service to users by flooding the target servers with an enormous amount external communications requests to gain access to business organizations systems or sometime disrupt the availability of those systems to the genuine customers (Liu & Cheng, 2009).

Cyber-attacks landscape:

Following are some of the attacks that are prevalent in cyberspace:

Denial of Service (DoS) attack: Denial of Service attacks involve flooding the victim system with huge number of useless packets or exploiting its vulnerabilities and impose huge number of computational tasks, drastically reducing its availability and many a times forcing the victim system out of service (Tan et al., 2014).

Distributed Denial of Service (DDoS) attack: When multiple, remotely controlled and widely dispersed network nodes, known as zombies (collectively also known as botnets), are used to launch an attack, its known as Distributed Denial of Service attack. These attacks could result in bandwidth depletion or the resource depletion. In bandwidth depletion attacks, target or victim system is flooded with unwanted traffic in order to stop the legitimate traffic from reaching the victim network. Resource depletion attack involves exhausting the system resources of victim systems and it leads to legitimate users not getting services (Deshmukh & Devadkar, 2015).

Advanced Persistent Threat (APT): APT attacks involve attacking to extract intellectual property by carrying out a targeted attack on government units or organizations. These attacks use multiple attack vectors such as cyber, physical and deception, to extract information or to disrupt a program or mission within the target organization or to disrupt the entire organization. These attacks are carried out over an extended period of time (Chen et al., 2014).

Intrusion & Intrusion Detection Systems:

Any attempt to compromise confidentiality, availability or integrity of a computer or network is known as **intrusion**. Monitoring and analyzing the events happening in a network or computer system with an objective of identifying the attempts to intrude is known as **intrusion detection**. A system that carries tasks such as monitoring and analysis if network events to identify the intrusion is known as intrusion detection system. **Intrusion detection systems** supplement traditional firewalls and identify malicious network traffic that doesn’t get identified by them (Liao et al., 2013).

Designing intrusion detection systems is a challenging task, due to continuous evolution of malwares. Attacks on networks and computer systems are becoming sophisticated and malware creators are using newer approaches and techniques to avoid being detected by intrusion detection systems (Khraiset et al., 2019).

Intrusion detection systems that examine network traffic and monitor multiple hosts real time are known as **Network intrusion detection systems (NIDS)**. Network intrusion detection systems monitor multiple hosts and analyze all packets moving across network. The other type of intrusion detection system involves monitoring logs, system configuration, and activities of various applications running on organization's network to identify unexpected changes and generating alerts. This is known as **Host intrusion detection system** (HIDS) (Rahul-Vigneswaran et al., 2020).

Based on detection techniques, intrusion detection systems are divided into two broad categories of **signature based** and **anomaly based**. Intrusion detection systems that identify attacks making use of signatures of known attacks and vulnerabilities are known as signature based or misuse-based intrusion detection system (Samrin & Vasumathi, 2017). Evolution of malware has made identification of attacks significantly difficult. This has led to an increase in focus on anomaly-based intrusion detection systems. Anomaly based detection systems work by identifying any parameter in the network that is different from the normal behavior. It identifies intrusion by examining the network behavior and comparing it with the normal behavior and creating an alert if the network behavior is significantly different than the normal behavior (Aljawarneh et al., 2018).

Machine learning is finding significant use in anomaly-based intrusion detection systems. Machine learning involves making machines learn based on data and then improve automatically through experience. Machine learning can be both supervised learning and unsupervised learning. Supervised learning systems use learnt mapping from labelled data to make predictions. Unsupervised machine learning deals with unlabeled data discovers information and patterns on its own and uses it to make predictions. When algorithms don't use labeled data and they utilize the concept of layers of artificial neural networks the algorithms are using an approach known as deep-learning (Jordan & Mitchell, 2015). Most of the upcoming anomaly-based intrusion detection systems are using some machine learning algorithm or deep learning algorithms in identifying the anomalies.

Application of machine learning & emerging deep learning approaches to design of intrusion detection systems:

Network intrusion detection systems have been an excellent application area for machine learning algorithms. Several categories of machine learning and deep learning algorithms are being researched to create robust anomaly-based network intrusion detection systems. Algorithms such as K-means, Random Forests and Naive Bayes, Convolutional neural networks and others have been applied to design intrusion detection systems.

- **Naive Bayes** is a probabilistic classifier and it applies Bayes theorem with assumption of strong independence. Its usefulness emanates from scenarios where data is limited and its also one of the classifiers that can be trained very quickly (Aziz et al., 2017).
- **Clustering** involves identifying patterns in the data and then dividing data into groups based on the patterns. K-means is a distance or centroid based algorithm that works on minimizing the sum of distances between the points in cluster and their centroid. **K-means clustering** works by randomly choosing k centroids initially from within the dataset and then iteratively recalculate the centroids to minimize the sum of distances between the points in cluster and the respective centroid (Jianliang et al., 2009).
- **Random Forests** algorithms are based on the concept of decision tree. Decision trees can help in both classification and regression. Random forests consist of large number of decision trees that are independent of each other and they operate as ensemble. In intrusion detection, it involves building the pattern of network services by the random forests algorithm and then outlier detection algorithms within it detect outliers to flag the anomaly (Zhang et al., 2008).
- **Convolutional neural networks** or CNN were initially studied for image processing and are deep learning algorithms and involve performing convolution operations to generate feature maps which are then pooled and which is subsequently used to predict the output (Vinaykumar et al., 2017).

There have been several researches done in recent times to identify which of the algorithms are more reliable and accurate in detecting the network intrusion. Effectiveness of those algorithms are assessed by several metrics. The key metrics that are used are Precision, Recall, and F1-score (Apruzzese et al., 2018).

- **Precision** is defined as ratio of true positives to total positives (including both true and false positives). In this true positive happens when a sample is malicious and it is correctly detected as malicious by the algorithm. It reflects the correct classification. When a normal network traffic packet is classified as negative, it is known as true negative. False positive represents a scenario where sample that is not malicious but is detected as malicious. It is incorrect classification as normal network traffic is incorrectly classified as attack. Precision is used to reflect performance.
- **Recall** helps identify the detection rate and is defined as ratio of true positives to sum of true positives and false negatives. False negative represents a scenario where sample that is malicious but is not detected or flagged as malicious.
- **F1-score** is harmonic mean of precision and recall and its value is 1 at a perfect precision and prefect recall (Almseidin et al., 2018).

Performance results of machine learning & emerging deep learning approaches in intrusion detection systems:

We looked at various researches that have been done to design intrusion detection systems, based on machine learning, deep learning algorithms and some modifications in the base algorithms to model the network intrusion systems. One of the limitations faced by all such researches and intrusion detection systems thus created is lack of reliable learning and test data. Following are the key algorithms evaluated:

One dimensional CNN with normalization on imbalanced data – it proposes a deep learning approach for developing intrusion detection system using one dimensional convolution neural network and LSTM (long short-term memory). It uses an approach of serializing the Transmission Control Protocol/Internet Protocol (TCP/IP) packets. Normal and anomalous traffic is collected from different sources and labelled as normal and abnormal traffic. Deep learning techniques are used to create good feature representation from this data. This model uses UNSW_NB15 (using both imbalanced data and balanced data) , which is recent intrusion dataset. The results from this model indicate a precision of 86.15%, detection rate of 95.15% and F-score of 90.43%. This research establishes the applicability of deep learning approaches to design of intrusion detection systems, giving results comparable to those obtained by machine learning based intrusion detection systems (Azizjon et al., 2020).

K-MEANS algorithm based on information Entropy – This research models and detects network anomalies using K-MEANS algorithm. This approach involves filtering the outliers initially to reduce the negative impact of outliers and isolated points. Then Identification of initial cluster centroids is done using information entropy. These centers are then used to classify the records into different clusters iteratively. The aim of this research is to increase the detection rate and reduce the false alarm rate. This research is done using KDDCUP99 dataset and it achieves a detection rate of 98.1% and false alarm rate of 2.3% (Han, 2012).

Hybrid Random Forests and weighted K-means – This research establishes, in their experiment, that anomaly detection method (using k-means clustering) achieves high detection rate reaching up to 99% and bad false positive rate reaching up to 12.6%. It further establishes that misuse detection methods (using random forests algorithm) achieve lower detection rate (close to 92.7%) but extremely good (low) false positive rate, of 0.54%. It therefore creates a mix-and-match model leveraging strength of both random forests in misuse detection and K-means clustering in anomaly detection. It creates a hybrid framework where it uses output from misuse detection part to be fed as input to the weighted k-means algorithm. It uses KDDCUP99 dataset and using 10% of the dataset, it achieves (for the hybrid framework) detection rate of 98.3% and false positive rate of 1.6% (Elbasiony et al., 2012).

Intrusion detection based on K-Means clustering and Naïve Bayes classification – This research proposes a hybrid approach where K-means clustering is used to group similar data instances. This is done as part of pre-classification. Resulting clusters from K-means clustering are then further classified to determine if they are in attack class using Naïve Bayes. This second stage classification (using Naïve Bayes) also corrects any misclassification done in first stage. This research uses KDDCUP99 dataset, and using the proposed hybrid approach, it achieves a precision of 99.5% and recall of 99.8% (Muda et al., 2011).

Table: Summary of Precision, Recall and F1-score achieved by Machine Learning and Deep Learning Methods as applied to model network intrusion detection

Machine Learning / Deep Learning Method	Dataset used	Precision	Recall	F1-Score
One dimensional CNN with normalization on imbalanced data	UNSW_NB15	86.15	95.15	0.9043
K-MEANS algorithm based on information Entropy	KD-DCUP99	97.7%	98.1%	0.979
Hybrid Random Forests and weighted K-means	KD-DCUP99	98.4%	98.3%	0.984
Intrusion detection based on K-Means clustering and Naïve Bayes classification	KD-DCUP99	99.5%	99.8%	.996

Current limitations & next steps:

Most of the models for intrusion detection that have been created using machine learning and deep learning are based on KDDCUP 99 dataset of network intrusion. This dataset is old and has limitations in terms of duplication of records and it being significantly old whereas network attacks and intrusions have evolved significantly (Cao et al., 2013). Evaluation of effectiveness of various machine learning and deep learning algorithms is significantly dependent on quality of both learning dataset and test dataset. While there have been attempts to create network intrusion dataset that represents more recent and realistic scenarios of network attack, this dimension still is a limitation in evaluating the effectiveness of such algorithms. Further, while hybrid approaches in applying machine learning algorithms have given a F1-score of greater than 0.95, the deep learning F1-scores are less than 0.95, despite using the more recent dataset than KDDCUP99. Another significant limitation of past research in this area is that high F1-scores and high accuracy are based on either KDDCUP99 dataset, which is outdated or the validation is done using a proportion of training dataset itself (validation accuracy) instead of testing done on a completely independent set of test data (test accuracy). Some of the research on NSL-KDD data performance shows that using all the features of NSL-KDD dataset, the validation accuracy is as high as 0.998 whereas test accuracy is in the range of 0.76 to 0.80 (Rawat et al., 2020). When testing has been done using a test data that is completely independent of training data, the F1 scores are There have been attempts to leverage hybrid approaches but they are either based on deep learning or based on machine learning but not their combination.

Reviewing the literature leads back to the question: How to design a robust and effective network intrusion detection system that uses latest datasets of cyber-attacks and gives high accuracy and F1-score using independent test data. This research aims to design a network intrusion detection system based on hybrid (or multistage application) of multiple deep learning algorithms (models) or combination of deep learning and machine learning algorithms using the latest cyber-attacks datasets and evaluating their precision, recall and F1-score.

3. RESEARCH METHODOLOGY & PROPOSED MODELS

Research Methodology and Data

This research uses the NSL-KDD dataset, which is an improved version of KDDCUP99 dataset. It uses three machine learning algorithms such as Random Forest (RF) decision tree and Naïve Bayes (NB) and trains them using “NSL-KDD Train” dataset. NSL-KDD dataset contains five types of data which is categorized as Normal, DOS, Probe, R2L, and U2R. NSL-KDD Train dataset consists of 125973 instances out of which 67343 are normal class and 58630 are anomaly class. This is divided into 80% for training and 20% for validation of the trained model. Trained model is then tested using “NSL KDD Test” dataset. NSL-KDD Test dataset consists of 22544 instances, out of which 9711 are normal class and 12833 are anomaly class. Accuracy, Precision, Recall and F1-scores are recorded for both validations done using 20% of the “NSL-KDD Train” dataset and independent testing done using “NSL-KDD Test” dataset. These models are built using all 41 attributes of NSL-KDD (class is the 42nd attribute) and WEKA is used to build, train, validate and test the machine learning models (Gao et al, 2019; Paulauskas & Auskalnis, 2017).

Proposed Model(s) Details

To answer the given research question, multiple models have been proposed in this research paper. Some of the hybrid models involve clustering in stage 1(Min et al., 2018). Then in second stage, those clusters have been further classified whether they are in attack class or normal class using another set of deep learning algorithm. Some of the models involve information-gain based attribute / feature selection to reduce the training time and to eliminate the features that are not influencing the outcome / prediction. Those algorithms have been trained and tested based on comprehensive set of cyberattack dataset from sources such as NSL-KDD (Divekar et al., 2018). Results of performance of above model will then be compared with published results of existing designs of network intrusion detection systems. There are four new hybrid models that will be built, trained, validated and tested in this research. They will be used to classify the data into normal and anomaly class (binary classification)

- Two of these models use information-gain based attribute selection in stage 1, followed by Random Forests and Naïve Bayes algorithms for classification in stage 2.
- Third model uses information-gain based attribute selection in stage 1, followed by DL4JMLPClassifier from Deeplearning4j for classification in stage 2 (Lang et al., 2019).
- The fourth model uses EM clustering from Deeplearning4j in stage 1 to identify normal and anomaly clusters, followed by information-gain based attribute selection for the normal cluster identified in stage 1 [this is stage 2], followed by followed by DL4JMLPClassifier from Deeplearning4j for classification (to identify any anomalies in the current cluster which was identified as normal cluster by EM algorithm in stage 1) [this is stage 3].

Random Forests and Naïve Bayes approaches are defined earlier in this research paper. Information Gain involves calculating the information gain or entropy for each feature and selecting the ones with highest information gain after defining a threshold value for the information gain (Lei, 2012). EM or expectation maximization is a clustering algorithm that models datasets as linear combinations of multivariate normal distributions. The quality of results is measured by log likelihood. The clustering results are created in such a was so as to set distribution parameters to maximize the log likelihood. Its ability to deal with noisy data, and also accepting number of clusters as inputs make it suitable for application to NSL-KDD dataset to identify two desired clusters of normal traffic and anomalous traffic (Abbas, 2008).

Model 1 (IG-RF):

(a) Stage 1: Attribute selection based on Information Gain

(b) Stage 2: Use Random Forests Decision Tree algorithm for classification utilizing the attributes selected in previous step

Model 2 (IG-NB):

(a) **Stage 1**: Attribute selection based on Information Gain

(b) **Stage 2**: Use Naïve Bayes' algorithm for classification utilizing the attributes selected in previous step

Model 3 (IG-DL3DL):

(a) **Stage 1**: Attribute selection based on Information Gain

(b) **Stage 2**: Use DL4JMLPClassifier algorithm (using 3 Dense Layers) for classification utilizing the attributes selected in previous step

Model 4 (EM-IG-DL2DL):

(a) **Stage 1**: EM algorithm based clustering

(b) **Stage 2**: Attribute selection based on Information Gain

(c) **Stage 2**: Use DL4JMLPClassifier algorithm (using 2 Dense Layers) for classification utilizing the attributes selected in previous step

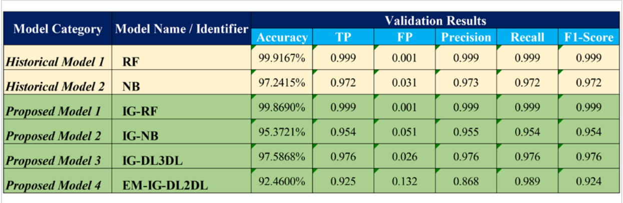
All these models are built, trained, validated and tested using WEKA (Lang et al., 2019).

4. EXPERIMENTAL RESULTS & ANALYSIS

Experiment Design

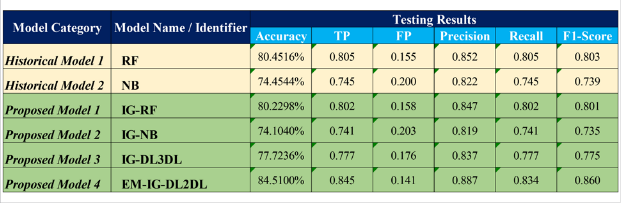
The proposed models were built and trained using "NSL-KDD Train" data. Out of total 125973 instances in the dataset, first 80% (100778 instances) were used to train the model, whereas remaining 20% (25195 instances) were used to validate the models. Independent testing of the model was done using "NSL-KDD Test" data consisting of 22544 instances. Validation and testing of models was done using WEKA. Results of the validation and testing are summarized below:

Validation Results



Model Category	Model Name / Identifier	Validation Results					
		Accuracy	TP	FP	Precision	Recall	F1-Score
Historical Model 1	RF	99.9167%	0.999	0.001	0.999	0.999	0.999
Historical Model 2	NB	97.2415%	0.972	0.031	0.973	0.972	0.972
Proposed Model 1	IG-RF	99.8690%	0.999	0.001	0.999	0.999	0.999
Proposed Model 2	IG-NB	95.3721%	0.954	0.051	0.955	0.954	0.954
Proposed Model 3	IG-DL3DL	97.5868%	0.976	0.026	0.976	0.976	0.976
Proposed Model 4	EM-IG-DL2DL	92.4600%	0.925	0.132	0.868	0.989	0.924

Testing Results



Model Category	Model Name / Identifier	Testing Results					
		Accuracy	TP	FP	Precision	Recall	F1-Score
Historical Model 1	**RF**	80.4516%	0.805	0.155	0.852	0.805	0.803
Historical Model 2	**NB**	74.4544%	0.745	0.200	0.822	0.745	0.739
Proposed Model 1	**IG-RF**	80.2298%	0.802	0.158	0.847	0.802	0.801
Proposed Model 2	**IG-NB**	74.1040%	0.741	0.203	0.819	0.741	0.735
Proposed Model 3	**IG-DL3DL**	77.7236%	0.777	0.176	0.837	0.777	0.775
Proposed Model 4	**EM-IG-DL2DL**	84.5100%	0.845	0.141	0.887	0.834	0.860

Results Analysis

Accuracy: Model validation results indicate that the Random Forests based classification generates a model that gives highest level of accuracy (99.9%+) for the validation dataset, amongst the models evaluated. However, for the same model of Random Forests based classification gives 80.45% accuracy for the test dataset, as shown in Figure 1. The proposed model of EM-IG-DL2DL gives highest level of accuracy (84.51%) for the test dataset.

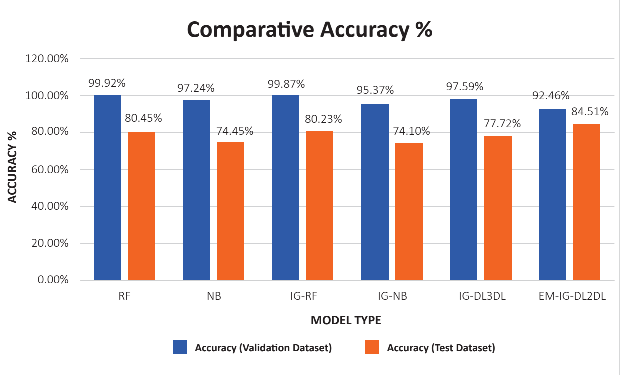


Figure 1: Comparison of Accuracy for validation dataset and test dataset

Precision: Model validation results indicate that the Random Forests based classification generates a model that gives highest level of precision (0.999) for the validation dataset. However, for the same model of Random Forests based classification gives 0.852 precision value for test dataset, as shown in Figure 2. The proposed model of EM-IG-DL2DL gives highest level of precision (0.887) on test dataset. Other models, using Naïve Bayes and combination of information gain with Random Forests, Naïve Bayes also give high level of precision for validation dataset. However, precision value is highest for the proposed model of EM-IG-DL2DL for the test dataset.

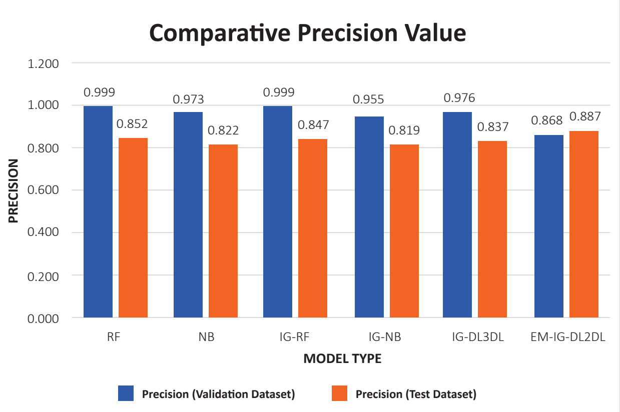


Figure 2: Comparison of Precision for validation dataset and test dataset

Recall: As shown in figure 3 below, proposed model EM-IG-DL2DL gives the highest level of recall values (0.834) for the NSL-KDD test data.

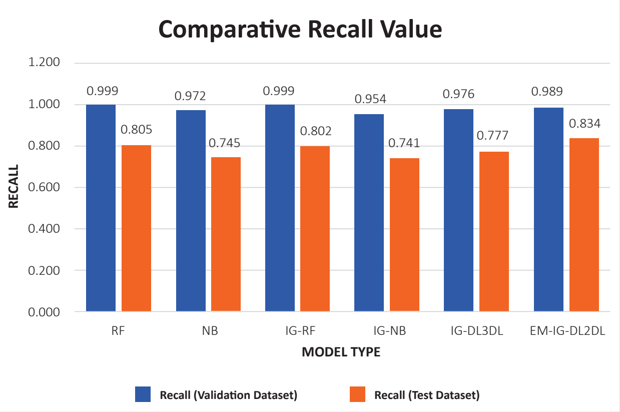


Figure 3: Comparison of Recall for validation dataset and test dataset

F-1 Score: Model validation results indicate that the Random Forests based classification gives highest level of F1-score (0.999) for the test dataset. However, the same model of Random Forests based classification gives 0.803 F1-score value, as shown in Figure 4. The proposed model of EM-IG-DL2DL gives the highest level of F1-score (0.860) for the test dataset. Other models, using Naïve Bayes and combination of information gain with Random Forests, Naïve Bayes also give high level of F1-score for validation. However, F1-score is highest for the proposed model of EM-IG-DL2DL for the test dataset.

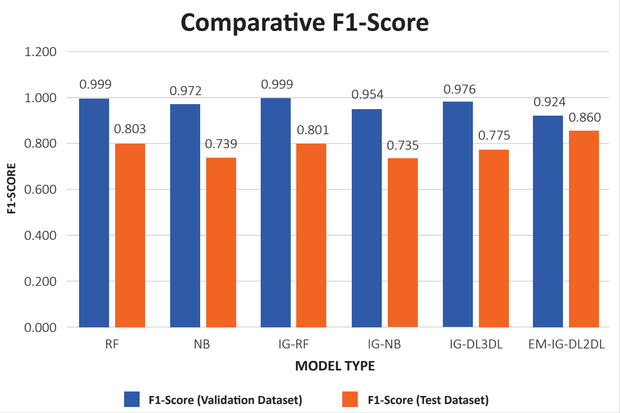


Figure 4: Comparison of F1-score for validation dataset and test dataset

5. CONCLUSION & FUTURE WORK

It is evident that in order to achieve high level of accuracy in identification of intrusion attempts, deep learning approaches can be leveraged in combination with other machine learning approaches. Out of the four approaches (models) proposed and tested in this research, the one using combination of EM clustering, information-gain based feature selection combined with dense layer based deep learning gave promising results, with accuracy of 84.51% and F1-score of 0.860 on NSL-KDD Test dataset. One of the possible limitations of this research work is that time taken to build and train the deep learning models was too high. This was high even when information-gain based feature selection was used to reduce the number of features used to predict whether the data is normal or anomalous. Hence more research shall be done to identify the approaches that reduce the time to build and train the models.

6. REFERENCES

Aziz, A. S., Hanafi, S. E.-O., & Hassanien, A. E. (2017). Comparison of classification techniques applied for network intrusion detection and classification. Journal of Applied Logic, 24, 109–118. doi:10.1016/j.jal.2016.11.018

Divekar, M. Parekh, V. Savla, R. Mishra and M. Shirole, “Benchmarking datasets for Anomaly-based Network Intrusion Detection: KDD CUP 99 alternatives,” 2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS), Kathmandu, 2018, pp. 1-8, doi: 10.1109/CCCS.2018.8586840.

Abbas OA. Comparisons Between Data Clustering Algorithms. Int Arab J Inf Technol. 2008; 5(3):320–325

Aljawarneh, S., Aldwairi, M., & Yassein, M. B. (2018). Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model. Journal of Computational Science, 25, 152–160.doi:10.1016/j.jocs.2017.03.006

Avram, M. G. (2014). Advantages and Challenges of Adopting Cloud Computing from an Enterprise Perspective. Procedia Technology, 12, 529–534. doi:10.1016/j.protcy.2013.12.525

Baumöl, U., Hollebeek, L. & Jung, R. Dynamics of customer interaction on social media platforms. Electron Markets 26, 199–202 (2016). https://doi.org/10.1007/s12525-016-0227-0

Bawany, N.Z., Shamsi, J.A. & Salah, K. DDoS Attack Detection and Mitigation Using SDN: Methods, Practices, and Solutions. Arab J Sci Eng 42, 425–441 (2017). https://doi.org/10.1007/s13369-017-2414-5

Chen, P., Desmet, L., & Huygens, C. (2014). A Study on Advanced Persistent Threats. Lecture Notes in Computer Science, 63–72. doi:10.1007/978-3-662-44885-4_5

Deshmukh, R. V., & Devadkar, K. K. (2015). Understanding DDoS Attack & its Effect in Cloud Environment. Procedia Computer Science, 49, 202–210. doi:10.1016/j.procs.2015.04.245

E. Bou-Harb, M. Debbabi and C. Assi, “Cyber Scanning: A Comprehensive Survey,” in IEEE Communications Surveys & Tutorials, vol. 16, no. 3, pp. 1496-1519, Third Quarter 2014, doi: 10.1109/SURV.2013.102913.00020.

E. Min, X. Guo, Q. Liu, G. Zhang, J. Cui and J. Long, “A Survey of Clustering With Deep Learning: From the Perspective of Network Architecture,” in IEEE Access, vol. 6, pp. 39501-39514, 2018, doi: 10.1109/ACCESS.2018.2855437.

Elbasiony, R. M., Sallam, E. A., Eltobely, T. E., & Fahmy, M. M. (2013). A hybrid network intrusion detection framework based on random forests and weighted k-means. Ain Shams Engineering Journal, 4(4), 753–762. doi:10.1016/j.asej.2013.01.003

G. Apruzzese, M. Colajanni, L. Ferretti, A. Guido and M. Marchetti, “On the effectiveness of machine and deep learning for cyber security,” 2018 10th International Conference on Cyber Conflict (CyCon), Tallinn, 2018, pp. 371-390, doi: 10.23919/CYCON.2018.8405026.

Gao, X., Shan, C., Hu, C., Niu, Z., & Liu, Z. (2019). An Adaptive Ensemble Machine Learning Model for Intrusion Detection. IEEE Access, 1–1. doi:10.1109/access.2019.2923640

J. Zhang, M. Zulkernine and A. Haque, “Random-Forests-Based Network Intrusion Detection Systems,” in IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 38, no. 5, pp. 649-659, Sept. 2008, doi: 10.1109/TSMCC.2008.923876.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255–260. doi:10.1126/science.aaa8415

Khraisat, A., Gondal, I., Vamplew, P. et al. Survey of intrusion detection systems: techniques, datasets and challenges. Cybersecur 2, 20 (2019). https://doi.org/10.1186/s42400-019-0038-7

L. Han, “Research of K-MEANS Algorithm Based on Information Entropy in Anomaly Detection,” 2012 Fourth International Conference on Multimedia Information Networking and Security, Nanjing, 2012, pp. 71-74, doi: 10.1109/MINES.2012.169

Lang, S., Bravo-Marquez, F., Beckham, C., Hall, M., & Frank, E. (2019). WekaDeeplearning4j: A deep learning package for weka based on Deeplearning4j.Knowledge-BasedSystems. doi:10.1016/j.knosys.2019.04.013

20. Lei, S. (2012). A Feature Selection Method Based on Information Gain and Genetic Algorithm. 2012 International Conference on Computer Science and Electronics Engineering. doi:10.1109/iccsee.2012.97

Liao, H.-J., Richard Lin, C.-H., Lin, Y.-C., & Tung, K.-Y. (2013). Intrusion detection system: A comprehensive review. Journal of Network and Computer Applications, 36(1), 16–24. doi:10.1016/j.jnca.2012.09.004

M. Almseidin, M. Alzubi, S. Kovacs and M. Alkasassbeh, “Evaluation of machine learning algorithms for intrusion detection system,” 2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY), Subotica, 2017, pp. 000277-000282, doi: 10.1109/SISY.2017.8080566.

M. Azizjon, A. Jumabek and W. Kim, “1D CNN based network intrusion detection with normalization on imbalanced data,” 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Fukuoka, Japan, 2020, pp. 218-224, doi: 10.1109/ICAIIC48513.2020.9064976.

M. Jianliang, S. Haikun and B. Ling, “The Application on Intrusion Detection Based on K-means Cluster Algorithm,” 2009 International Forum on Information Technology and Applications, Chengdu, 2009, pp. 150-152, doi: 10.1109/IFITA.2009.34.

Muda, Z., Yassin, W., Sulaiman, M. N., & Udzir, N. I. (2011). Intrusion detection based on K-Means clustering and Naïve Bayes classification. 2011 7th International Conference on Information Technology in Asia. doi:10.1109/cita.2011.5999520

Paulauskas, N., & Auskalnis, J. (2017). Analysis of data pre-processing influence on intrusion detection using NSL-KDD dataset. 2017 Open Conference of Electrical, Electronic and Information Sciences (eStream). doi:10.1109/estream.2017.7950325

Phyu, T. Z., & Oo, N. N. (2016). Performance comparison of feature selection methods. In P. Plapper, M. Guo, & S. Suhag (Eds.), 2015 3rd International Conference on Control, Mechatronics And Automation (Vol. 42). Cedex A, France: EDP Sciences

R. Samrin and D. Vasumathi, “Review on anomaly based network intrusion detection system,” 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), Mysuru, 2017, pp. 141-147, doi: 10.1109/ICEECCOT.2017.8284655.

R. Vinayakumar, K. P. Soman and P. Poornachandran, "Applying convolutional neural network for network intrusion detection," 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Udupi, 2017, pp. 1222-1228, doi: 10.1109/ICACCI.2017.8126009.

Rahul-Vigneswaran K., Poornachandran P., Soman K. (2020) A Compendium on Network and Host Based Intrusion Detection Systems. In: Kumar A., Paprzycki M., Gunjan V. (eds) ICDSMLA 2019. Lecture Notes in Electrical Engineering, vol 601. Springer, Singapore

Rawat, S., Srinivasan, A., Ravi, V., & Ghosh, U. (2020). Intrusion detection systems using classical machine learning techniques vs integrated unsupervised feature learning and deep neural network. Internet Technology Letters. doi:10.1002/itl2.232

S. Liu and B. Cheng, "Cyberattacks: Why, What, Who, and How," in IT Professional, vol. 11, no. 3, pp. 14-21, May-June 2009, doi: 10.1109/MITP.2009.46.

ŞAHİN, H., TOPAL, B. (2018). Impact of Information Technology on Business Performance: Integrated Structural Equation Modeling and Artificial Neural Network Approach. Scientia Iranica, 25(3), 1272-1280. doi: 10.24200/sci.2018.20526

V. L. Cao, V. T. Hoang and Q. U. Nguyen, "A scheme for building a dataset for intrusion detection systems," 2013 Third World Congress on Information and Communication Technologies (WICT 2013), Hanoi, 2013, pp. 280-284, doi: 10.1109/WICT.2013.7113149

Z. Tan, A. Jamdagni, X. He, P. Nanda and R. P. Liu, "A System for Denial-of-Service Attack Detection Based on Multivariate Correlation Analysis," in IEEE Transactions on Parallel and Distributed Systems, vol. 25, no. 2, pp. 447-456, Feb. 2014, doi: 10.1109/TPDS.2013.146