Intelligent Human Systems Integration (IHSI 2024), Vol. 119, 2024, 237–247
https://doi.org/10.54941/ahfe1004488

IDS With Hybrid Sampling Technique: Combination Over and Under-Sampling Technique and Comparison With Deep Convolutional Approach

Ghislain Vlavonou[1], Ischyros Gangbo[1], Thierry Nsabimana[1], Christian Bimenyimana[2], Perpetus Jacques Houngbo[1], Joël T. Hounsou[1], and Fulvio Frati[3]

[1]Institute of Mathematics and Physics of University of Abomey Calavi, UAC, Benin
[2]Università des Grands Lacs (UGL), Burundi
[3]Università degli Studi di Milano, Italy

ABSTRACT

Digital is constantly evolving with the appearance of connected objects and on top of the popularization today of artificial intelligence. One of the direct inductions remains the excessive proliferation of various kinds of attacks in computer systems. Hackers exploit these vulnerabilities to break in and attack systems with increasingly complex attacks. The consequences of intrusions are destructive and ruinous for businesses and organizations such as electronic ransom ware, data alteration and loss, financial and brand image loss. It is important for those involved in computer systems to equip any computer centre with adequate tools to prevent malicious individuals from accessing the systems. To remedy these setbacks, several IT tools are developed including IDS intrusion detection systems. IDS intrusion detection systems are devices designed to monitor a computer system, give alerts and trigger real-time counterattacks in the event of attacks. These intelligent systems use several detection approaches and various algorithms. The performance of the IDS is increased when the features dimensionality are reduced significantly. This study proposed feature dimensionality reduction techniques such as Principal Component Analysis (PCA) and Auto-Encoder (AE). The output from the reduced dimensional features are used to build machine Learning algorithms. The performance results is evaluated on the CSECICIDS2018 datasets. The proposed public intrusion data sets suffer from the Imbalance class. In order to handle this issue, we propose hybrid sampling technique by combining Over and undersampling technique. The performance results from the reduced features in terms of true positive, False positve, recall, precision, F-Measure, ROC Area, PRC Area show the better performance. In addition, the obtained results are compared with deep convolutional approach.

Keywords: Machine learning, Principal component analysis, Intrusion detection system, Artificial neural network, Deep learning, CSECICIDS2018 datasets, Hybrid sampling technique

INTRODUCTION

Intrusion detection is today an important area of computer security given the new technological developments and especially because of the ingenuity

of hackers. These hackers exploit the new vulnerabilities inherent in technological development with increasingly sophisticated attacks. All this allows them to take control of devices, divert them from their normal activities or steal sensitive information. To detect intrusions, two techniques can be used by IDS. Signature- based detection suffers from a fundamental inadequacy because if an attack is not listed in the signature database, it goes unnoticed while hackers are increasingly ingenious especially if they pursue a specific objective. Anomaly detection is based on the triggering of an alert in the event of a significant deviation from normal behaviour.

Although several anomaly detection algorithms exist, shortcomings persist in the detection of intrusions. The idea is to have as few false positives as possible while having a very good level of accuracy and efficiency of the intrusion detection model. Public intrusion datasets suffer from the bivalence class. To solve this problem, we propose a hybrid sampling technique by combining the oversampling and subsampling technique. Similarly, we propose to compare and combine several learning algorithms. The performance of the reduced characteristics of these systems are analyzed to determine a high-performance system. In this article, we will successively address the following points:

1) provide an overview of the basic concepts of intrusion detection by machine learning;
2) prepare datasets and the hybrid approach model;
3) compare the results obtained with a deep convolution approach.

BASIC CONCEPTS

Intrusion detection has always been a major concern for researchers especially because of its disastrous consequences on organizations. Traditional techniques have many limitations due to new developments in IT and the complexity of new attacks. Thus, several new detection techniques are developed by many researchers to secure systems. These techniques are essentially based on artificial intelligence. The first definition of machine learning was produced in 1959 by Arthur Samuel, who defined machine learning as "a field of study that gives the computer the ability to learn without being explicitly programmed" [SAMUEL (1959)]. After this first definition, several other definitions were proposed by other researcher sincluding:

- Machine learning (or artificial learning) is, the study of algorithms and methods that allow programs to improve automatically by experience [MITCHELL et al. (1983)].
- A machine learning system builds models from data. The process of building these models is called learning or training. The data used in the training are therefore called learning data [WEIMER et al. (2010)].

There are two ways for a system to learn: 1) the system modifies itself to exploit its own knowledge more effectively; 2) the system acquires new knowledge through external sources.

TYPES OF MACHINE LEARNING

Three machine fundamental learning techniques exist: supervised learning, unsupervised learning and reinforcement learning. Supervised learning is a technique that uses labeled datasets and algorithms to accurately classify data or predict results. Unsupervised learning is used in cases where the information used to train the model is neither labeled nor classified. As for reinforcement learning, it consists in iteratively optimizing an algorithm from the rewards awarded in case of a good answer.

By combining these methods, it is possible to design another method of machine learning which is semi-supervised learning. Semi-supervised learning is a compromise between the types of supervised learning and unsupervised learning. In this case, all the data does not need to be labelled. In the case of intrusion detection, machine learning is used to learn, from a dataset with network traffic, to perform classification as normal traffic or attacks. The following graph shows the different machine learning techniques.

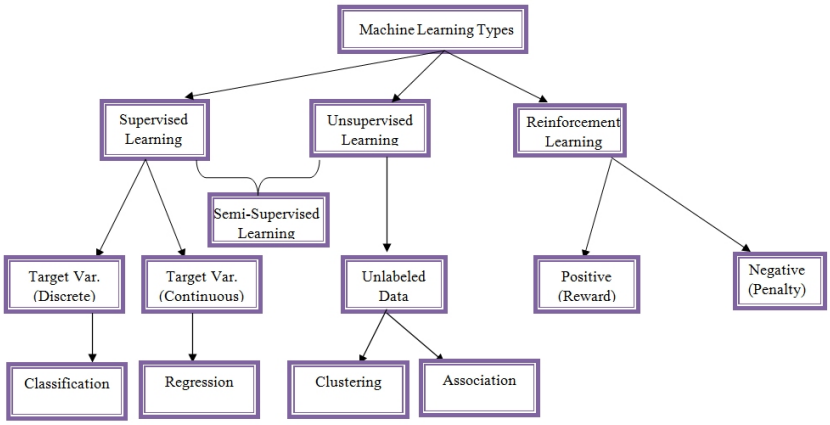

Figure 1: Machine learning types.

IMPLEMENTATION AND ANALYSIS OF RESULTS

For research, several public datasets exist including the KDD Cup 1999, the NSL-KDD, the DARPA Intrusion Detection Data, Kyoto 2006+, CSE-CIC-IDS. For our work, we will use CSE-CIC-IDS2018 for machine learning. This dataset is generated by the collaboration between Communication Security Establishment (CSE) and the Canadian Institute for Cybersecurity (CIC). It is a dynamically generated dataset that reflects real traffic consisting of normal traffic and intrusions. The final data set includes seven different attack scenarios: Brute-force, Heartbleed, Botnet, DoS, DDoS, web attacks and network infiltration from within.

Corpus Formation

We performed a number of operations to clean the corpus of non-numeric values in the numeric fields. Similarly, some fields contained some infinite values. This negligible amount of data was deleted. In the case of our dataset,

each instance of attacks is associated with one of the 14 attack labels. The output value can have 15 possible values considering the 14 attack labels plus normal traffic. In addition, the CIC-IDS2018 CSE is a very unbalanced corpus because normal traffic accounts for about 80% of total traffic and some attacks like SQL Injection, SSH-Bruteforce and FTP-BruteForce are largely underrepresented. To correct this imbalance, we have processed this corpus in order to have a less unbalanced database. At the end of this work, the database obtained with work includes about 89083 data. We then randomly divided this base into two subsets. These two subsets allowed us to have an 80% training subset and a 20% test subset. The dataset contains fourteen types of attacks and normal traffic. This table shows the traffic statistics in the three data sets as well as the percentage of each traffic over the total of that set.

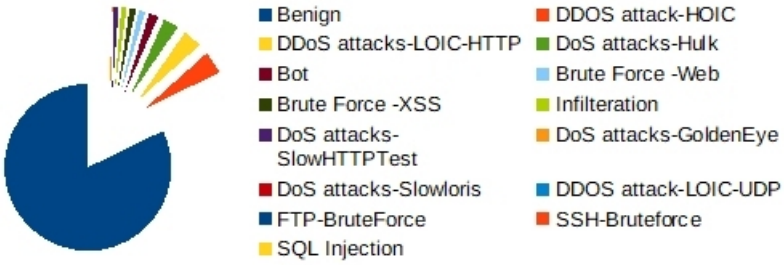

Figure 2: Characteristics of the corpus CSE-CIC-IDS2018.

Table 1. Characteristics of data sets.

Classes	Work Matrix		Training		Testing	
	Number	%	Number	%	Number	%
Benign	18805	21.11%	15055	21.13%	3750	21.05%
Bot	4930	5.53%	3946	5.54%	984	5.52%
Brute Force –Web	611	0.69%	493	0.69%	118	0.66%
Brute Force –XSS	230	0.26%	185	0.26%	45	0.25%
DDOS attack-HOIC	8753	9.83%	6985	9.80%	1768	9.92%
DDOS attack-LOIC-UDP	1730	1.94%	1354	1.90%	376	2.11%
DDoS attacks-LOIC-HTTP	9400	10.55%	7530	10.57%	1870	10.50%
DoS attacks-GoldenEye	7405	8.31%	5917	8.30%	1488	8.35%
DoS attacks-Hulk	7113	7.98%	5721	8.03%	1392	7.81%
DoS attacks-SlowHTTPTest	4618	5.18%	3691	5.18%	927	5.20%
DoS attacks-Slowloris	6310	7.08%	5040	7.07%	1270	7.13%
FTP-BruteForce	5111	5.74%	4117	5.78%	994	5.58%
Infilteration	13411	15.05%	10700	15.01%	2711	15.22%
SQL Injection	87	0.10%	67	0.09%	20	0.11%
SSH-Bruteforce	569	0.64%	465	0.65%	104	0.58%
TOTAL	**89083**	**100 %**	**71266**	**100 %**	**17817**	**100 %**

Our system is based on four main modules namely reduction, class rebalancing, processing module and detection module Fig. 3 represents the diagram of our system.

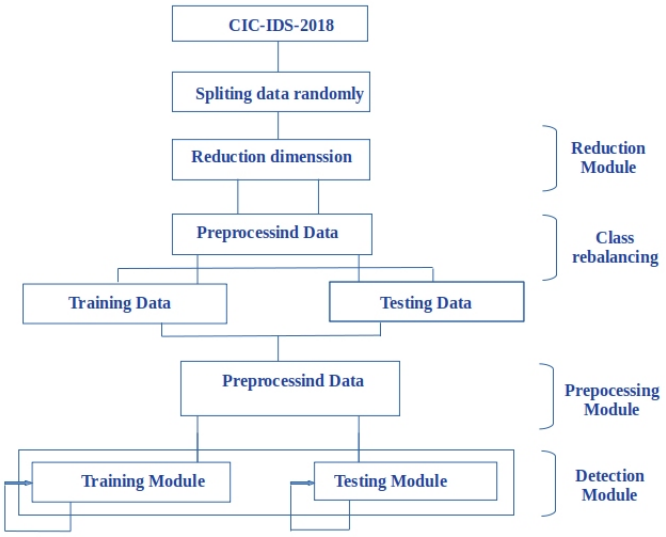

Figure 3: Diagram of the model.

Reduction Module

We used techniques to reduce the dimensionality of the characteristics namely: PCA and auto-encoder.

Principal Component Analysis (PCA): it is a method of dimension reduction. It is a method that allows to make predictive models with very little loss of information. It transforms a group of correlated variables and finds the underlying group of orthogonal variables with the greatest variance. Since PCA is a dimensional reduction method, it aims to identify the schema-forming parameters within the data.

It is a matter of summarizing the information in a number of synthetic variables called: Main components. The first main component is the direction that maximizes variance within the data, while the second main component also maximizes variance but is orthogonal to the first. Each main component added will be orthogonal to the previous component with the greatest variance. We must explicitly center and reduce the variables.

As shown in the next diagram, the first thirteen components capture over 85% of the available information. These components will be used in the subsequent overs and undersampling process

Auto-Encoder: The goal of auto-encoder networks (AE) is to learn to reproduce its inputs at the output through a minimal representation. Its structure is symmetrical with the same number of inputs and outputs, a decrease in the number of nodes per layer, followed by an increase as shown in the following figure.

The first part realizes an encoder by seeking a representation of the input data on a set of reduced dimensions. As for the second, it performs the reverse operation by transforming the encoded data to the original dimension set. In this paper, we just used the encoder part of the auto encoder to perform data reduction.

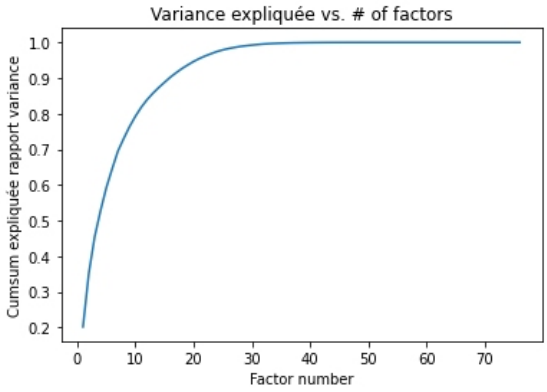

Figure 4: cumulative variance graph based on the number of factors applied to the corpus.

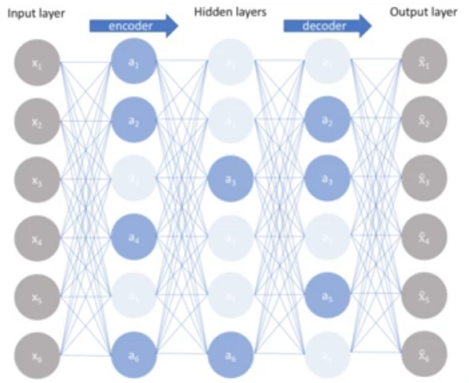

Figure 5: Principes of the auto encoder.

The first part realizes an encoder by seeking a representation of the input data on a set of reduced dimensions. As for the second, it performs the reverse operation by transforming the encoded data to the original dimension set. It's used for example in non-supervised learning, to reduce the size of the input set, and therefore the number of characteristics. When the reconstruction error between the inputs and outputs is low, the center layer is of sufficient size to contain useful information of the inputs characteristics. All input data can then be converted by the encoder before use at the input of a classification system and thus reduce its complexity. In semi-supervised learning, the AE can be trained to reproduce a particular class, corresponding for example to the nominal functioning of a system. Below are the emulations of the treatment on our corpus

Since the corpus suffers from an imbalance problem, we proposed a hybrid sampling technique by combining the technique of over-sampling and sub-sampling. The algorithm used is Adaptive synthetic sampling (ADASYN) which improves learning compared to data distributions in two ways: it reduces the bias introduced by class imbalance and adaptively shifts the classification limit to hard-to-learn examples. [HE et al. (2008)]. After

preprocessing we train an automatic classification model (Random Forest Classifier and K-NN) to predict the results on all test data. In the first phase of our study, we choose the random forest classification algorithm because it performs better and the second phase we combined the two algorithms using a voting system. The evaluation of our proposed hybrid system is based on a confusion matrix using 15 indicators.

Table 2.a. PCA and random Forest

3308	0	1	0	0	0	0	3	1	1	0	435	1	0	0
0	984	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	110	0	0	0	0	0	0	0	0	0	4	0	0
0	0	4	40	0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	1768	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	376	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1870	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1486	2	0	0	0	0	0	0
1	0	0	0	0	0	0	0	1391	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1270	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	882	0	0	0	112
636	0	0	0	0	0	0	0	0	1	0	2074	0	0	0
0	0	3	1	0	0	0	0	0	0	0	0	16	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	104	0
0	0	0	0	0	0	0	0	0	0	423	0	0	0	504

Table 2.b. Encoder and random Forest

3253	0	1	0	0	0	0	4	0	1	0	490	1	0	0
0	984	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	106	3	0	0	0	0	0	0	0	0	4	0	0
0	0	6	36	0	0	0	0	0	0	0	0	3	0	0
0	0	0	0	1768	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	376	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1870	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1486	2	0	0	0	0	0	0
1	0	0	0	0	0	0	0	1391	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1270	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	882	0	0	0	112
612	0	0	0	0	0	0	0	0	1	0	2098	0	0	0
1	0	4	1	0	0	0	0	0	0	0	0	14	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	104	0
0	0	0	0	0	0	0	0	0	0	423	0	0	0	504

Table 2.c. PCA and combine classification methods

3111	8	8	1	1	1	0	17	44	4	12	5	497	33	8
1	965	0	0	2	0	0	0	16	0	0	0	0	0	0
15	0	77	3	1	0	0	0	5	0	0	0	2	15	0
9	1	4	22	0	0	0	0	4	0	0	0	3	2	0
0	4	0	0	1596	0	0	0	168	0	0	0	0	0	0
0	0	0	0	0	376	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1870	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	1463	2	0	4	0	0	12	0
32	36	1	1	147	0	0	0	1171	0	0	0	1	3	0
0	0	0	0	0	0	0	0	0	916	0	10	1	0	0
10	0	0	0	0	0	0	0	0	0	1250	0	4	6	0
0	0	0	0	0	0	0	0	0	500	0	494	0	0	0
750	4	5	1	3	0	0	19	25	1	2	0	1883	17	1
1	0	4	1	0	0	0	2	1	0	0	0	0	11	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	103

Table 2.d. Encoder and combine classification

3289	0	0	0	0	0	0	6	0	2	3	0	446	3	1
0	984	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	107	3	0	0	0	0	0	0	0	0	0	4	0
0	0	5	36	0	0	0	0	0	0	0	0	0	4	0
0	0	0	0	1768	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	376	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1870	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1486	2	0	0	0	0	0	0
1	0	0	0	0	0	0	0	1391	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	922	0	5	0	0	0
0	0	0	0	0	0	0	0	0	0	1270	0	0	0	0
0	0	0	0	0	0	0	0	0	960	0	34	0	0	0
661	0	0	0	0	0	0	1	0	0	1	0	2048	0	0
1	0	4	0	0	0	0	0	0	0	0	0	0	15	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	104

Performance based on (04) indicators, namely false positives, true positives, false negatives and true negatives, allows us to calculate metrics for our model: precision score, recall score, accuracy score and F1 scores of the classification. The performance results of the reduced characteristics in terms of true positive, false positive, recall, accuracy, measurement F, ROC zone, PRC zone show the best performance compared to the hybrid algorithm using the PCA data reduction technique on autoencoder.

Precision Score = TP / (FP + TP)
Recall Score = TP / (FN + TP)
Accuracy Score = (TP + TN)/ (TP + FN + TN + FP)
F1 Score = 2* Precision Score * Recall Score/ (Precision Score + Recall Score/)

Table 3a. PCA and random forest.

Precision Score :	0.9082898355503171
Recall Score :	0.9082898355503171
f1 Score :	0.9082898355503171
accuracy Score :	0.9082898355503171

Table 3b. Encoder and random forest.

Precision Score :	0.90598866251333
Recall Score :	0.90598866251333
f1 Score :	0.90598866251333
accuracy Score :	0.90598866251333

Table 3c. PCA and combine classifier.

Precision Score :	0.8591794353707134
Recall Score :	0.8591794353707134
f1 Score :	0.8591794353707134
accuracy Score :	0.8591794353707134

Table 3d. Encoder and combine classifier.

Precision Score :	0.8811808946511759
Recall Score :	0.8811808946511759
f1 Score :	0.881180894651176
accuracy Score :	0.8811808946511759

Below are the graphs for the metric ROC of the test and learning data according to the reduction algorithm**:**

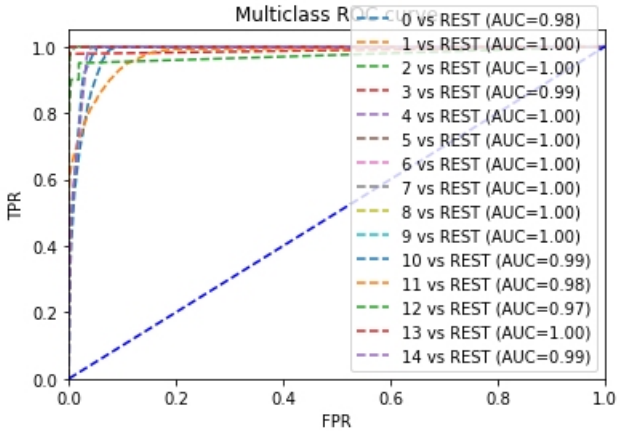

Figure 6a: ROC PCA and random forest.

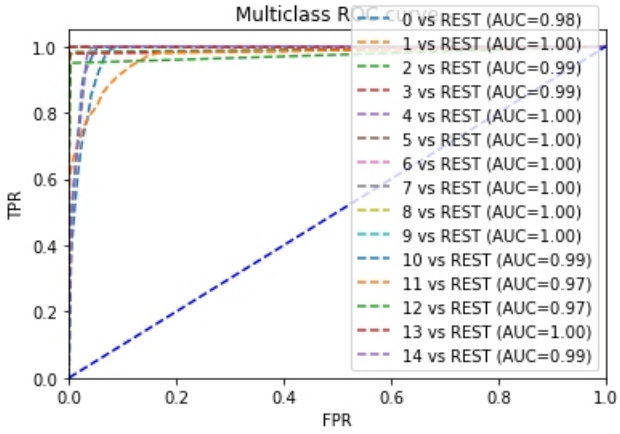

Figure 6b: ROC Encoder and random forest.

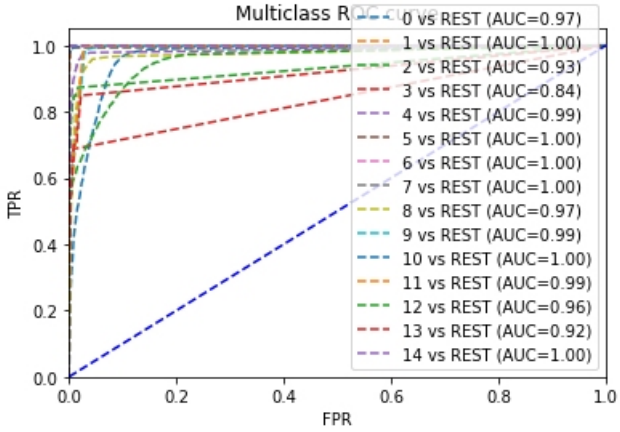

Figure 6c: PCA and combine classifier.

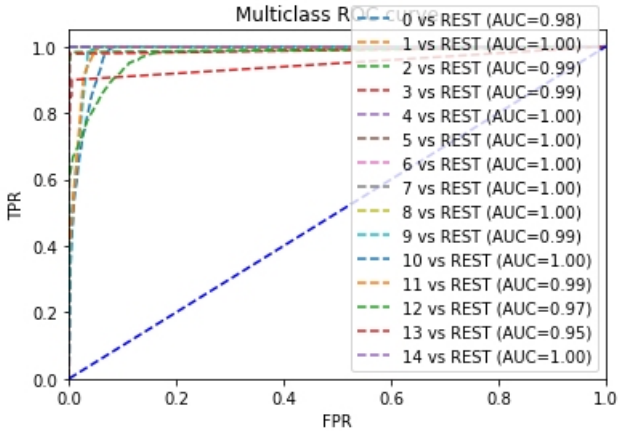

Figure 6d: Encoder and combine classifier.

The analysis of the results on the testing data shows us that the Random Forest classification algorithm applied with one of the reduction techniques performs better than the combination of several classification algorithms on the same dataset. The best performance is a 91%.

CONCLUSION

Security threats to computer systems are multiple and attacks are daily and increasingly sophisticated and complex. The compromise of a system remains a possibility regardless of the preventive means put in place. So in this paper we have pre-processed the data, reduced the data, applied learning supervised and unsupervised algorithms. The imbalance observed in public data was managed by using the ADASYM technique. Research carried out in this paper can be further investigated in order to improve our approach and the learning methods used. Thus, we plan to use convolutional neural networks to explore other datasets.

REFERENCES

Ansam Khraisat, Iqbal Gondal, Peter Vamplew and Joarder Kamruzzaman. ≪ Survey of intrusion detection systems: techniques, datasets and challenges ≫. Cybersecurity, (2019) 2:20.

Biswas, S. K. (2018). "Intrusion detection using machine learning: A comparison study". International Journal of Pure and Applied Mathematics, 118 (19), 101–114.

Canadian institute of cybersecurity, university of new brun-swick https://www.unb.ca/cic/datasets/ids-2018.html

David Pierrot, Nouria Harbi, Jérôme Darmont. ≪ Détection des intrusions et aide à la décision ≫. 2018, page 4–5.

Dr. S. Vijayarani 1 and Ms. Maria Sylviaa. S, ≪ Intrusion Detection System – A Study ≫. International Journal of Security, Privacy and Trust Management (IJSPTM) Vol 4, No 1, Février 2015.

Haq, N. F., Onik, A. R., Hridoy, M. A. K., Rafni, M., Shah, F. M., & Farid, D. M. (2015). Application of machine learning approaches in intrusion detection system: a survey. IJARAI-International Journal of Advanced Research in Artificial Intelligence, 4(3), 9–18. https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python

He H., Bai Y., Garcia E. A. & Li S. (2008). Adasyn : Adaptive synthetic sampling approach for imbalanced learning. In Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on, pp. 1322–1328.

Iqbal H. Sarker. ≪ Machine Learning: Algorithms, Real-World Applications and Research Directions ≫. SN Computer Science (2021) 2:160.

J. T. Hounsou, P. Ciza Niyomukiza, T. Nsabimana, G. Vlavonou, F. Frati, and E. Damiani "Learning Vector Quantization and Radial Basis Function Performance Comparison Based Intrusion Detection System", IHSI 2021: Intelligent Human Systems Integration 2021 pp. 561–572.

J. T. Hounsou, T. Nsabimana, & J. Degila (2019). "Implementation of network intrusion detection system using soft computing algorithms (self organizing feature map and genetic algorithm)". Journal of Information Security 10:1–24.

Mitchell, T. M., Utgoff, P. E., Banerji, R. "Learning by experimentation: Acquiring and refining problem-solving heuristics". Machine learning. Springer, Berlin, Heidelberg. pp. 163–190, 1983.

Mrutyunjaya Panda, Abraham Ajith & Patra Ranjan. A Hybrid Intelligent Approach for Network Intrusion Detection. International Conference on Communication Technology and System Design, 2011.

Rajeev Singh, ≪ Introduction to Intrusion Detection System ≫. International Journal of Electrical and Electronics Research (IJEER). Vol. 2, Issue 1, pp: (1-6), Month: January-March 2014.

Samuel, A. L. "Some Studies in Machine Learning Using the Game of Checkers". IBM journal of Research and Development 3, vol. 44, no 1.2, pp. 206–226, 1959.

Vahid Farrahi & Marzieh Ahmadzadeh ≪ KCMC: A Hybrid Learning Approach for Network Intrusion Detection using K-means Clustering and Multiple Classifiers ≫. International Journal of Computer Applications (0975 – 8887), Volume 124 – No.9, August 2015.

Varun Gupta, Monika Mittal. ≪ KNN and PCA classifier with Autoregressive modelling during different ECG signal interpretation ≫. 6th International Conference on Smart Computing and Communications, ICSCC 2017, 7–8 December 2017, Kurukshetra, India.

Weimer, M. "Machine Teaching--A Machine Learning Approach to Technology Enhanced Learning". PhD Thesis, Technische Universität, Brunswick, 2010.

Y. E. Kurniawati, A. E. Permanasari and S. Fauziati, "Adaptive Synthetic-Nominal (ADASYN-N) and Adaptive Synthetic-KNN (ADASYN-KNN) for Multiclass Imbalance Learning on Laboratory Test Data," *2018 4th International Conference on Science and Technology (ICST)*, Yogyakarta, Indonesia, 2018, pp. 1–6.

Yasi, Wang, Hongxun Yao, Sicheng Zhao. ≪ Auto-encoder based dimensionality reduction ≫. Neurocomputing, Volume 184, 5 April 2016, Pages 232–242.

Z. Muda, W. Yassinn M. Sulaiman, I. Udzir. ≪ Intrusion detection based on K-Means clustering and Naïve Bayes classification ≫. 7th International Conference on Information Technology in Asia, 2011.