Towards Efficient Features Dimensionality Reduction for Network Intrusion Detection on Highly Imbalanced Traffic

Razan Abdulhammed, Hassan Musafer, Miad Faezipour, and Abdelshakour Abuzneid

Department of Computer Science and Engineering

University of Bridgeport, Bridgeport, CT

Abstract

The performance of an IDS is significantly improved when the features are more discriminative and representative. This research effort is able to reduce the CICIDS2017 dataset's feature dimensions from 81 to 10, while maintaining a high accuracy of 99.6% in multi-class and binary classification. Furthermore, we propose a Multi-Class Combined performance metric $Combined_{Mc}$ with respect to class distribution to compare various multi-class and binary classification systems through incorporating FAR, DR, Accuracy, and class distribution parameters. In addition, we developed a uniform distribution based balancing approach to handle the imbalanced distribution of the minority class instances in the CICIDS 2017 network intrusion dataset.

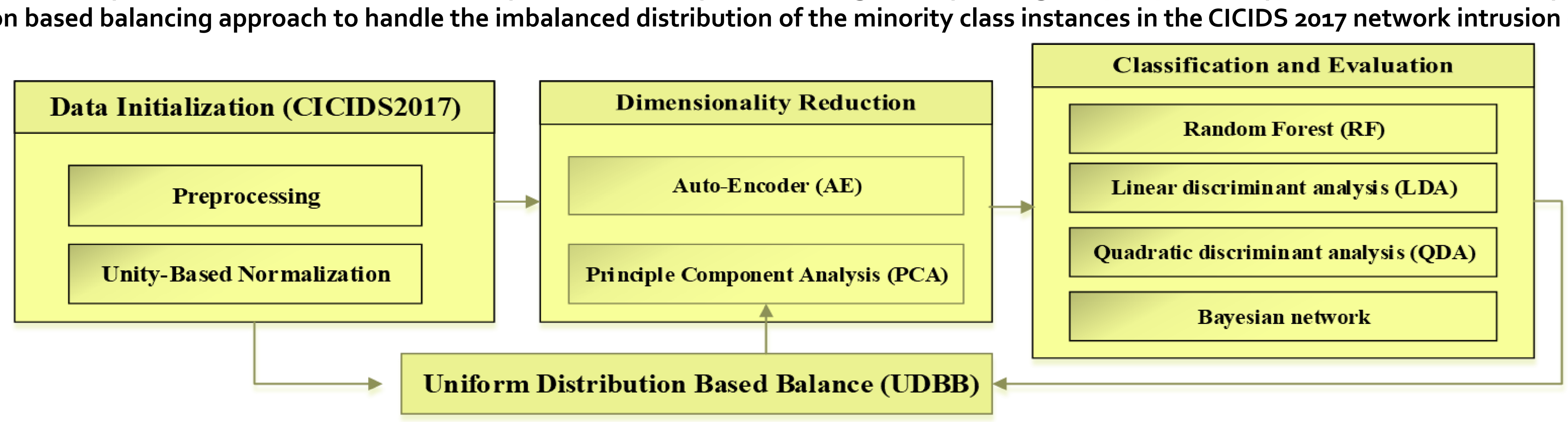

Figure 1. Proposed Framework

Features Dimensionality Reduction Framework

The procedure of our proposed framework, as presented in Figure 1, mainly includes Preprocessing, Unity-Based Normalization, Dimensionality reduction, Classification and Evaluation and finally, combating imbalanced class distributions using the uniform distribution based balance approach.

Highly Imbalanced CICIDS2017 Dataset

The CICIDS2017 [1] covers various attack scenarios that represent common attack families. The attacks include Brute Force Attack, HeartBleed Attack, Botnet, DoS Attack, Distributed DoS (DDoS) Attack, Web Attack, and Infiltration Attack. CICIDS2017 was collected based on real traces of benign and malicious activities of the network traffic. The total number of records in the dataset is 2,830,108. The benign traffic encompasses 2,358,036 records (83.3% of the data), while the malicious records are 471,454 (16.7% of the data). CICIDS2017 is one of the unique datasets that includes up-to-date 14 types of attacks. Furthermore, the features are exclusive and matchless in comparison with other datasets such as AWID[2,3], and CIDD-001 [4]. For this reason, CICIDS2017 was selected as the most comprehensive IDS benchmark to test and validate the proposed ideas.

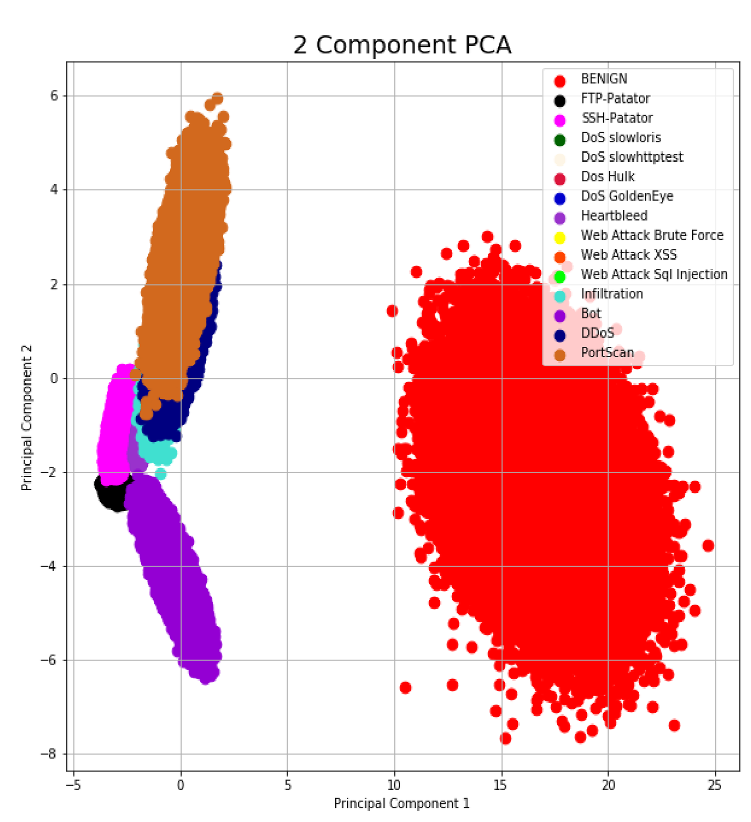

Figure 2. 2 D Visualization After Applying UBBD

Calculate $Combined_{Mc}$ with respect to Class Distribution

Feed Confusion Matrix CM
For i =1 to C
Calculate the total number of FP for Ci as the sum of values in the ith column excluding TP
Calculate the total number of FN for Ci as the sum of values in the ith row excluding TP
Calculate the total number of TN for Ci as the sum of all columns and rows excluding the ith row and column
Calculate the total number of TP for Ci as the diagonal of the ith cell of CM
Calculate the total number of instances for Ci as the sum of the ith row
Calculate the total number of instances in the dataset as the sum of all rows
Calculate Acc , DR . FAR for each class
Ci Calculate the distribution of each Ci using Eq. 1
i ++
Calculate $Combined_{MC}$ using Eq 2

Proposed $Combined_{Mc}$ Pseudo Code Calculation

$$CombinedMc = \sum_{i=1}^{c} \lambda_i \left(\frac{Acc_i + DRi}{2} - FAR_i\right) \quad \ldots 2$$

$$dist = \lambda_i = \frac{Number\ of\ instances\ in\ Class\ i}{Number\ of\ instances\ in\ the\ dataset} \quad \ldots 1$$

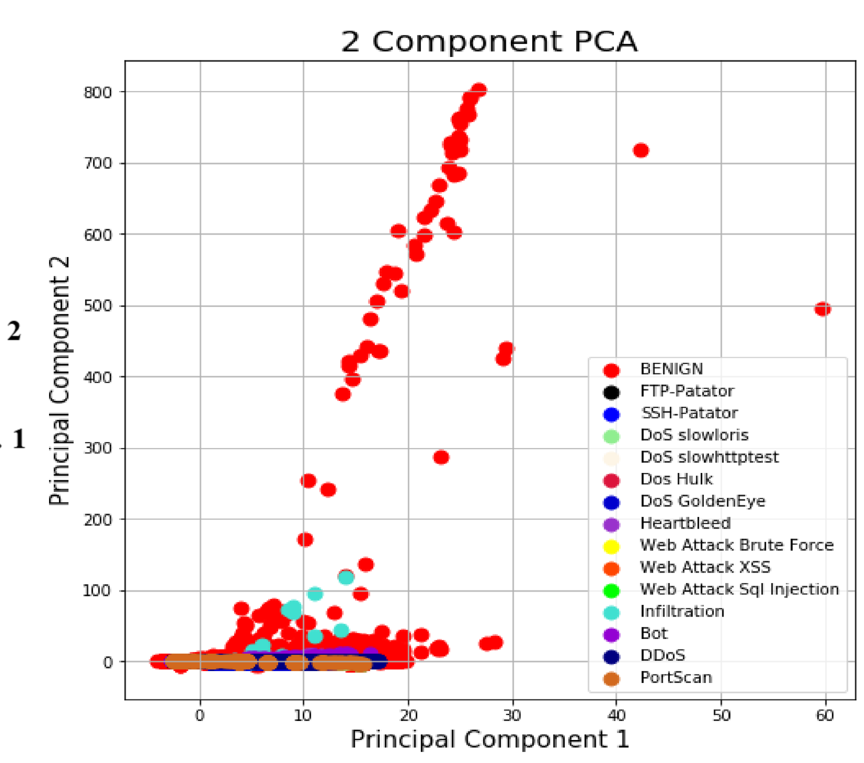

Figure 3. 2 D Visualization Before Applying UBBD

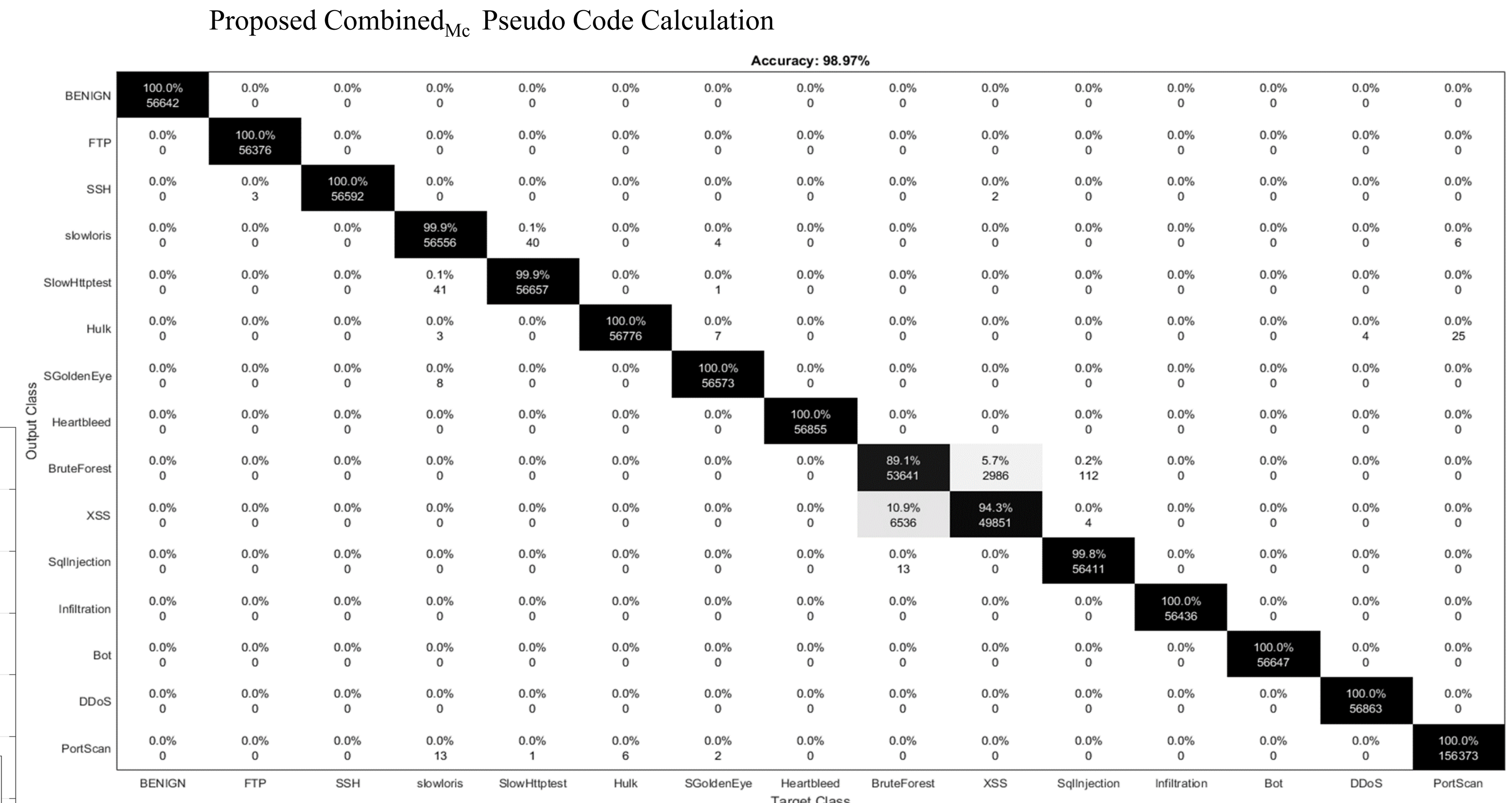

Figure 4. Confusion Matrix for (PC A− RF) Mc−10 After Applying UDBB Approach

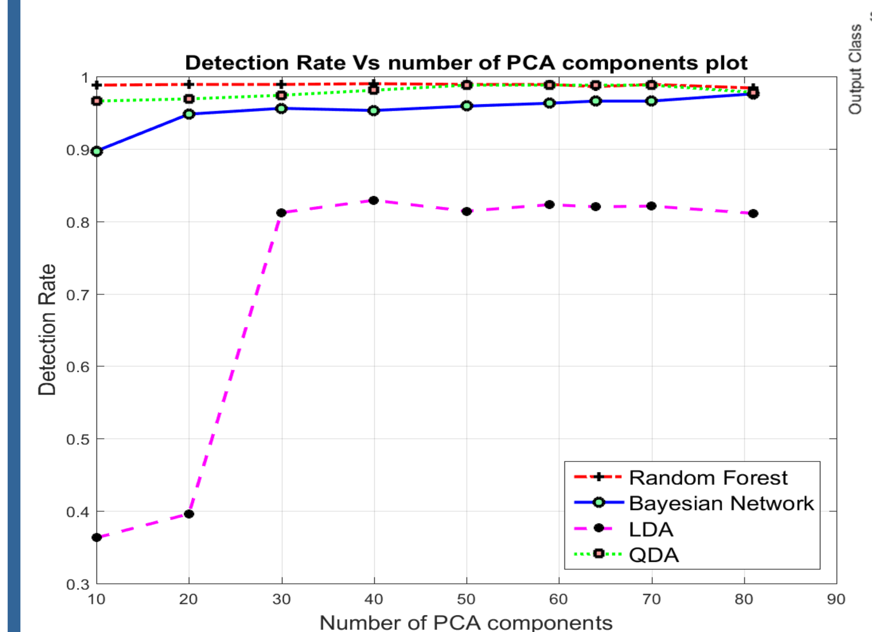

Figure 5. Binary Class Classification: Detection Rate in terms of number of components using PCA

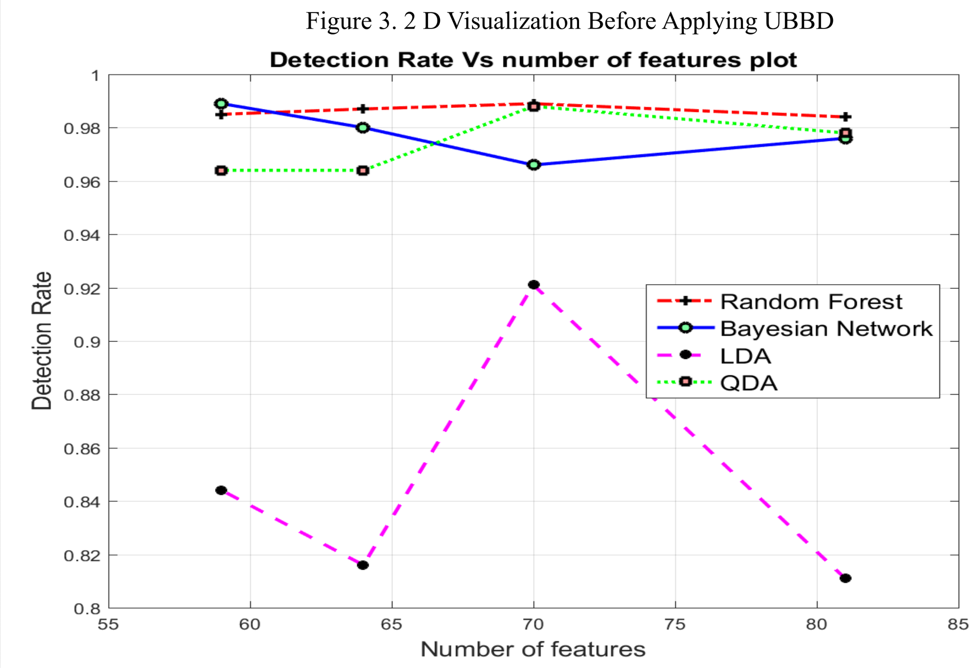

Figure 6. Binary Class Classification: Detection Rate in terms of number of features using Auto-Encoder

Table 2 Comparison with previous work

Reference	Classifier name	F-measure	Feature selection/extraction (Features Count)
[1]	MLP	0.948	Payload related features
[6]	SVM	0.921	DBN
[7]	KNN	0.997	Fisher Scoring (30)
[8]	XGBoost for DoS Attacks	0.995	(80)
[9]	Deep Learning for Port Scan Attacks	Accuracy 97.80	(80)
	SVM for Port Scan Attacks	Accuracy 69.79	(80)
[10]	XGBoost	Accuracy 98.93	DDR Features Selections (36)
[11]	Deep Multi Layer Perceptron (DMLP) for DDoS Attacks	Accuracy 91.00	Recursive feature elimination with Random Forest
Proposed Framework	Random Forest	0.995	Auto-encoder (59)
Proposed Framework	Random Forest	0.996	PCA with Original Distribution (10)
Proposed Framework	Random Forest	0.988	PCA With UDBB(10)

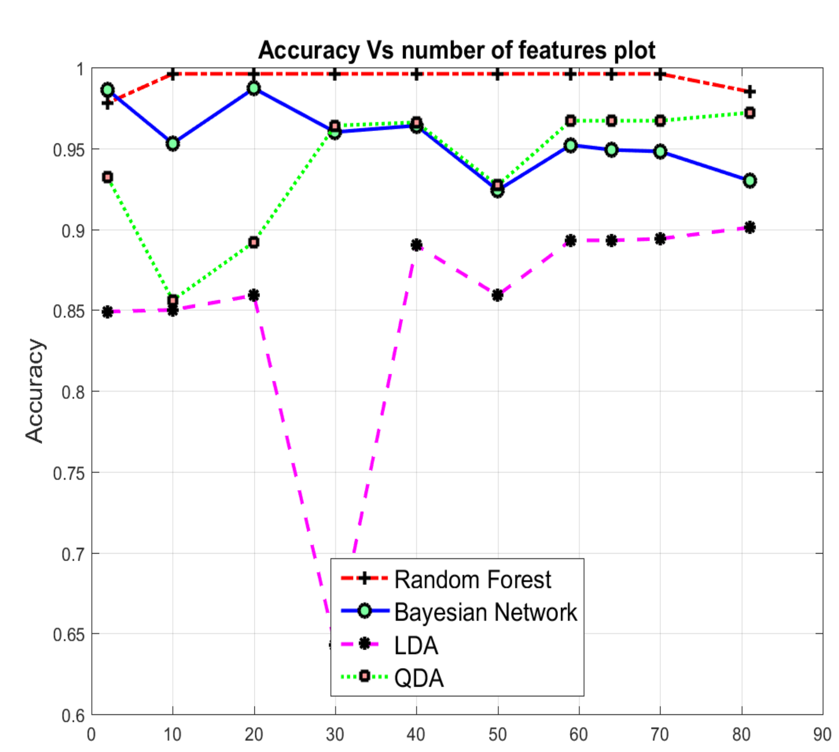

Figure 7. Multi Class Classification: Accuracy in terms of number of components using PCA

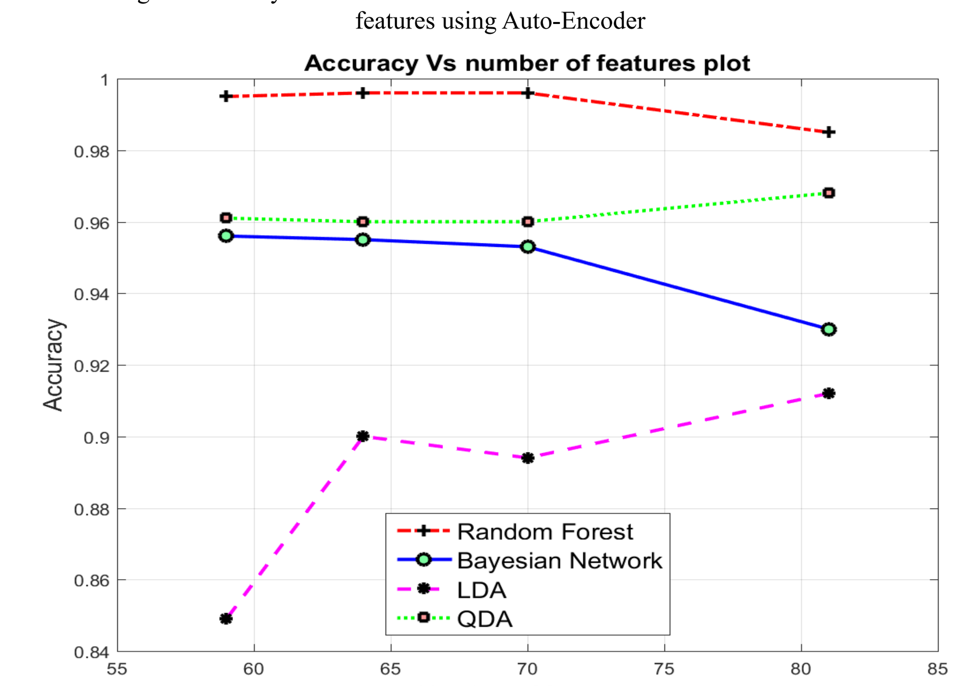

Figure 8 Multi Class Classification: Accuracy in terms of number of features using Auto-Encoder

Table 3 Time to build and test the models

Classifier	Time to Build the Model (Sec.)	Time to Test the Model (Sec.)
	Binary-class Classification	
LDA	12.16	5.56
QDA	12.84	6.57
RF	752.67	21.52
BN	199.17	11.07
	Multi-class Classification	
LDA	17.5	2.96
QDA	15.35	3.16
RF	502.81	41.66
BN	175.17	10.07

Table 1 UDBB Approach Pseudo Code

Input Training Set: D_{Train}
Set Distribution to Uniform
C : Number of Classes
F_T : Total number of features in D_{Train} Training Set
I_{old} : Total number of Instances in D_{Train}
Calculate the required number of Instances in each class: $I_{Resample}$

Training Set $D_{Trainnew} = \emptyset$
For each class C_i Do
While $i \neq I_{Resample}$
For each feature $F_1, \ldots, F_T$
Generate new sample using uniform distribution
Assign Class label
Return $DTrain_{new}$

Conclusions and Future work

As exemplified from the obtained results, the PCA approach is able to preserve important information in CICIDS2017, while efficiently reducing the features dimensions in the used dataset, as well as presenting a reasonable visualization model of the data. These findings provide insights for extended future research work including: fault tolerance, model resilience, quality of Experience , and Adaption to non stationary

References

[1] Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization,"in Proceedings of fourth international conference on information systems security and privacy, ICISSP , 2018.
[2] Abdulhammed, Razan, et al. "Deep and Machine Learning Approaches for Anomaly-Based Intrusion Detection of Imbalanced Network Traffic." IEEE sensors letters 3.1 (2019): 1-4.
[3] M. R. Abdulhammed, M. Faezipour, A. Abuzneid, and A. Alessa, "Effective features selection and machine learning classifiers for improved wireless intrusion detection," in 2018 International Symposium on Networks, Computers and Communications (ISNCC) IEEE, 2018, pp. 1–6.
[4] R. Abdulhammed, M. Faezipour, and K. Elleithy,Intrusion Detection in Self organizing Network:A Survey NewYork: CRC Press Taylor Francis Group, 2017, ch. 13, pp. 393–449.
[5] Watson, G. A Comparison of Header and Deep Packet Features when Detecting Network Intrusions. Technical report, 2018.
[6] Marir, N.; Wang, H.; Feng, G.; Li, B.; Jia, M. Distributed Abnormal Behavior Detection Approach based on Deep Belief Network and Ensemble SVM using Spark. IEEE Access 2018
[7] Aksu, D.; Üstebay, S.; Aydin, M.A.; Atmaca, T. Intrusion Detection with Comparative Analysis of Supervised Learning Techniques and Fisher Score Feature Selection Algorithm. International Symposium on Computer and Information Sciences. Springer, 2018, pp. 141–149.
[8] Bansal, A.; Kaur, S. Extreme Gradient Boosting Based Tuning for Classification in Intrusion Detection Systems. International Conference on Advances in Computing and Data Sciences. Springer, 2018, pp. 372-380.
[9] Aksu, D.; Aydin, M.A. Detecting Port Scan Attempts with Comparative Analysis of Deep Learning and Support Vector Machine Algorithms, IEEE IBIGDELFT, 2018, pp. 77–80.
[10] Bansal, A. DDR Scheme and LSTM RNN Algorithm for Building an Efficient IDS. Master's thesis, 2018.
[11] Ustebay, S.; Turgut, Z.; Aydin, M.A. Intrusion Detection System with Recursive Feature Elimination by Using Random Forest and Deep Learning Classifier. IEEE IBIGDELFT, 2018, pp. 71–76.

ECT
Emerging Communications Technologies Research Center

D-BEST Lab
Digital / Biomedical Embedded Systems & Technology Lab