

Differentially-Private Federated Intrusion Detection via Knowledge Distillation in Third-party IoT Systems of Smart Airports

Yang Chen ^a, Saba Al-Rubaye ^a, Antonios Tsourdos ^a, Lawrence Baker ^b, Colin Gillingham ^b

^a School of Aerospace, Transport and Manufacturing, Cranfield University, Milton Keynes, United Kingdom

^b NCC Group, Manchester, United Kingdom

{yang.chen, s.alrubaye, a.tsourdos}@cranfield.ac.uk, {Lawrence.Baker, Colin.Gillingham}@nccgroup.com

Abstract—With the increasing deployment of IoT and Industry 4.0, the federated learning system was presented to preserve the privacy between the third-party IoT systems and the security operation center in smart airports. Nonetheless, the extremely skewed distribution of cyber threats increases the complexity of intrusion detection system (IDS) in smart airports, while privacy preservation limits the utility of IDS in the process of server model update. In this article, we have devised a knowledge distillation (KD)-based Convolutional Neural Network and Gated Recurrent Unit (CNN-GRU) model to improve the accuracy of multiple intrusion detection. In addition, the trade-off between privacy and accuracy is achieved by denoising the adaptive parameter update mechanism to upgrade the optimizer of Differentially-Private (DP) Federated IDS. The results indicate high effectiveness and robustness of DP Federated KD-based IDS for third-party IoT systems of a smart airport.

Index Terms—Federated Learning, KD-based IDS, CNN-GRU, Smart Airport, ICPSSs, Trade-off between Privacy and Accuracy

I. INTRODUCTION

The massive adoption of IoT and Industry 4.0 has led to the appearance of smart airports [1]. Unfortunately, the expanded vulnerability of airport network is introduced due to the susceptibility of industrial cyber physical systems (ICPSs). There are two main problems about intrusion detection in smart airport. On the one hand, the frequency of cyber-attacks in smart airports is too low to provide enough labelled abnormal data required for deep learning. On the other hand, the cybersecurity datasets may contain sensitive information, such as Personally Identifiable Information (PII). Types of services are shown as Fig. 1 in smart airport eco-system. It is crucial to handle the above database under a user-level privacy. Imbalanced learning is a popular learning method for intrusion detection where some of classes occupy most of the examples while the remaining classes are trained from a few supervised samples. Differential-Privacy (DP) is regarded as a superb strategy for privacy guarantee in smart airport ICPSSs. Generally, Gaussian noise mechanism is applied to satisfy (ϵ, δ) -DP with $\delta \geq \frac{5}{4} \exp(-(\sigma\epsilon)^2/2)$ and $\epsilon \in (0, 1)$, so that the unanimous noise type can be understood and corrected easily [21]. DP-Federated Learning (FL) is a new kind of distributed learning framework in which multiple users or devices cooperatively train the server model without sharing the local dataset [2].

In this paper, we design a DP-Federated IDS and deploy them via a knowledge distillation (KD) based CNN-GRU

model to increase the test accuracy of the minority cyber-attack. Moreover, we put forward an adaptive parameter update mechanism to keep balance between privacy and utility. Section II presents related work. Section III and IV explain our system architecture and algorithms. Section V analyses our experimental results, and Section VI concludes.



Fig. 1: Smart Airport Ecosystem

II. RELATED WORK

Imbalanced classification is generally referred as an essential problem of intrusion detection. In the field of cybersecurity, some traditional machine learning research has been conducted by [3]. However, with the emerging 6G communication and IoE, imbalanced classification in cyber-attacks may be more effectively dealt with using a deep learning structure, including data-level, algorithm-level and hybrid method [4]. For data-level, [5] mixed random sampling method with transfer learning for plankton classification. At algorithm-level, MFE loss and focal loss were introduced by [6] and [7] separately. Hybrid-method is a special method that combines data-level and algorithm-level, such as LMLE by [8]. Unfortunately, most of the current imbalanced learning methods are only acceptable to image recognition and natural language processing, which demands high data dimension. Nevertheless, some breakthroughs are still made by semi-supervised learning and knowledge distillation for intrusion detection. [9] proposed a lightweight semi-supervised learning

method based on consistency regularization for intrusion detection, but that method can only identify normal and abnormal rather than detailed types of malicious action.

The concept of differential privacy was proposed to quantify the disclosure about private information of some datasets during running of a federated deep learning architecture. To analyse the privacy cost, [10] develops a differentially private SGD algorithm, which calculates the privacy budget by the momentum accountant. [11] adds Laplacian noise to each client of a federated learning framework for user-level privacy. However, the above privacy preservation always works along with the degradation of learning accuracy. Therefore, [12] update norm clip online to maintain the learning accuracy without the significant deterioration of privacy. Regardless, the current trade-off strategy requires longer to converge.

III. SYSTEM DESIGN

A. DP Federated IDS in Smart Airport

A typical federated IDS consists of three layers, including cloud, fog nodes and IoT devices [13]. Likewise, our system is divided into security operation center, IDS nodes and aviation operation system in Fig. 2.

Security Operation Center: This block orchestrates the cyber-security of all IoT systems of smart airports by updating the model parameter from the third-party IoT system, so this block can be regarded as server in the cross-device federated learning system [14]. As a service provider, the security operation center coordinates FL training processes [2] by the following steps. Firstly, this center trains a series of benchmark dataset to generate the initial model parameter. Then, the intrusion detection nodes download the current model parameter from the security operation center to build up the initial local intrusion detection model. When the local training stops due to convergency, the local updated model parameter will be uploaded to the security operation center, where the uploaded parameters will be aggregated for the construction of a top-stage IDS in smart airport.

IDS Node: They are the bridge between the security operation center and the aviation operation block [15]. With better performance in storage and computation, IDS nodes allocate their privacy budgets to the corresponding airport operation system, such as the airport management system and automated baggage handling system. In return, noise is generated from the relevant aviation operation system. Meanwhile, IDS nodes will add the above noise to local model parameters during each training epoch, while uploading the model parameters to the security operation center for aggregation [13].

Aviation Operation System: These IoT systems represent the clients in the federated learning system, which will insert additive noise to protect user-level privacy. Moreover, in comparison with ensemble learning, there is no need to align the amount of data in each third-party system, because each third-party system trains their deep learning model in their own intrusion detection node. Thus, the deployment of them will improve the robustness of smart airport operation [16].

In our DP Federated IDS, the intrusion detection model is the key element to guarantee the utility of the whole IDS system. To this extent, the details of our KD-based intrusion detection model will be described in the following sub-section.

B. KD-based CNN-GRU Model

The KD-based deep intrusion detection model is devised with a pre-trained teacher model and a lightweight student model as Fig. 3. For one thing, the teacher model is configured by CNN-Module and GRU-Module, where CNN-Module consists of three 1D convolutional layer along with MaxPooling to avoid overfitting issues. Furthermore, two-phase Gated Recurrent Unit (GRU) layers together with Dropout layers are connected with Repeat Vector layer in series within GRU-Module. Each module is responsible for different kinds of feature extraction during the deep intrusion detection, after which all feature parameters will be reshaped by the specific concatenation layer. In the end, the feature parameters will be trained and updated via 4-layer fully connection. Essentially, the skewed data distribution is the root reason why cyber-attack classification is so complicated, so it is necessary to randomly under-sample the input data for higher performance of our deep classifier. For another, student model is just the simple version of teacher model, by learning from which student model can acquire the prior knowledge of teacher model. Therefore, KD-based deep intrusion detection model updates the model parameters by referring to both student loss and distillation loss. For distillation module, we figure out soft targets [17] instead of hard targets to distill more information and configuration from the transfer learning within the teacher model, so it will improve the accuracy of student model efficiently even with a basic structure.

Moreover, a new GRU-based seq2seq framework is described as Fig. 4 in GRU-Module, where encoder and decoder are linked with the bottleneck called repeat vector layer, where we transform the dimension of the output sequence from encoder into the category of target label, which will fit each target separately before flattening, in order to outperform the direct link with two vanilla GRU layers.

IV. DENOISING ADAPTIVE PARAMETER UPDATE MECHANISM

In this section, we will discuss the impact of noise in the security operational center. When the updated parameters are uploaded to the security operation center, all of them have been injected with Gaussian noise for the purpose of privacy preservation. However, these noised parameters will confuse the model update of server port, which is the core of DP federated learning. It is important for intrusion detection to complete within an acceptable period of time, because both the failure and chaos of IoT system will result in much lower operational efficiency and even the disorder or service disruption in smart airport, which is detrimental to the extension of IoT/IoE in the promising 6G networks.

To mitigate the noise from the uploaded parameters, DPFedAvg-M with adaptive clipping was offered to enable the

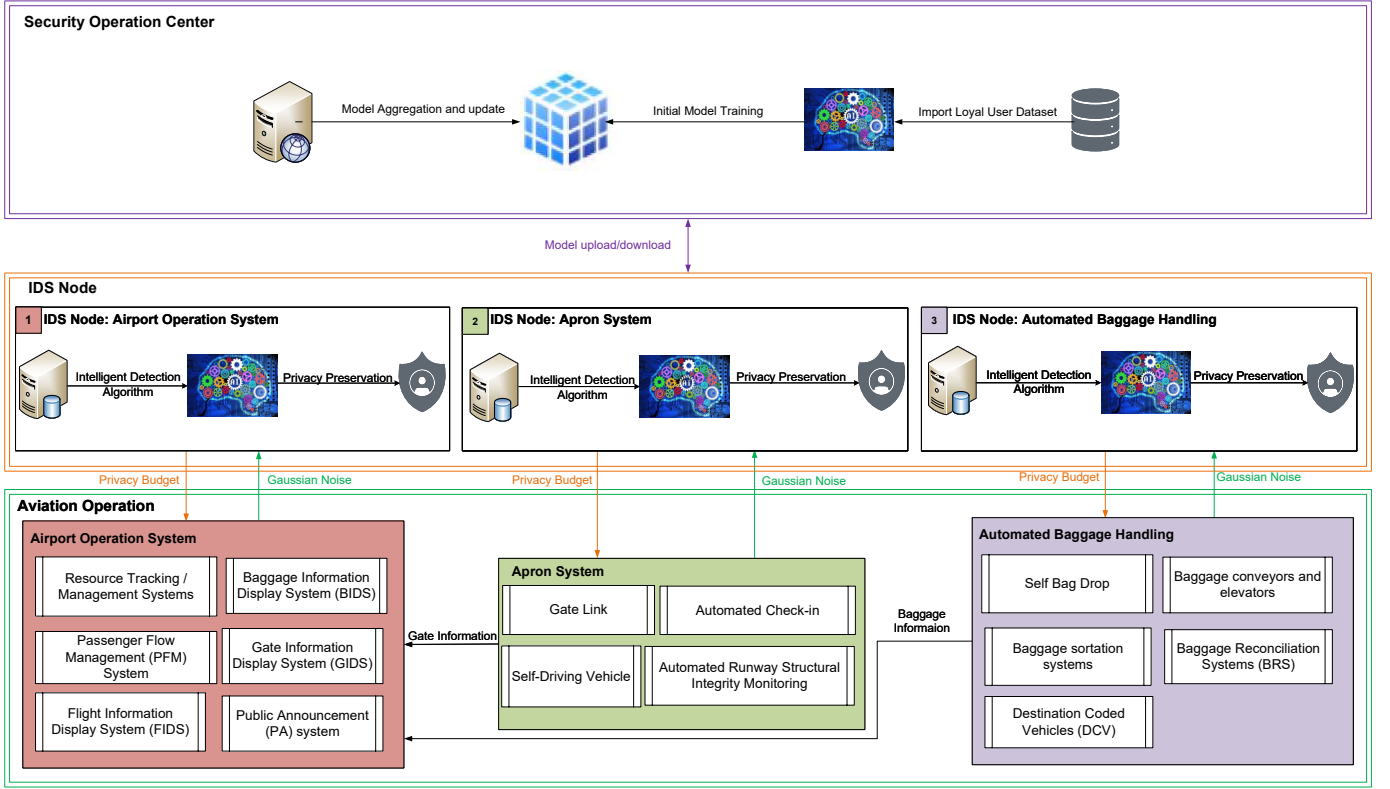


Fig. 2: Overview of DP-Federated IDS in Smart Airport

trade-off between privacy and accuracy of federated learning [12], where clipping bound C was applied to flexibly control the rate of parameter update and the magnitude of noise. Despite that, the obvious performance degradation can be observed in DPFedAvg-M, where our deep intrusion detection algorithms request more time to achieve convergency. In addition, it is notable that the rising noise multiplier will cut down the validation performance identified by the gradient, and convergency value of accuracy and loss function. As a result, we adopt a denoising adaptive parameter update mechanism based on Adam Optimizer [18] for server port, by which the balance between accuracy and privacy will be accomplished throughout the intrusion detection in smart airports.

In traditional Adam Optimizer, all model parameters are updated by the original gradient, while the situation is so different in DPFedAvg-M for the sake of DP noise. In fact, the model parameter of DPFedAvg-M is updated based on both gradient and noise multiplier, where we propose a denoising adaptive parameter update mechanism as Algorithm 1 to avoid the effect of noise multiplier during server update.

At the case of our algorithm, a penalty for noise regularization has been introduced to the model parameters update, which will correct the error resulting from DP noise. According to decoupled weight decay regularization [19], L2 regularization is not equivalent to weight decay in adaptive gradient descent, so we also decouple our noise correction into gradient descent field and parameter update field.

Proposition 1. Gaussian noise regularization factor $\lambda_n \leq$

Algorithm 1 Denoising Adaptive Parameter Update Mechanism

Given: learning rate $\epsilon = 0.001$, exponential decay rates for moment estimates $\rho_1 = 0.9$, $\rho_2 = 0.999$, small constant for numerical stabilization $\delta = 10^{-8}$, DP noise regularization factor $\lambda_n \in \mathbb{R}$, DP noise multiplier $\sigma \in \mathbb{R}$, clipping bound $C \in \mathbb{R}$

Notification: All operators are element wise

- 1: **Initialize:** time step $t = 0$, parameter vector $\theta_{t=0} \in \mathbb{R}^n$, 1st moment variable $s = 0$, 2nd moment variable $r = 0$
- 2: **while** stopping criterion not met **do**
- 3: Sample a minibatch of m examples from the training set with label
- 4: $t \leftarrow t + 1$
- 5: $g_t \leftarrow \frac{1}{m} \sum_i \nabla J_\theta(x_i, y_i) + \lambda_n \cdot C^t \cdot N(0, \sigma^2)$
- 6: $s_t \leftarrow \rho_1 \cdot s_{t-1} + (1 - \rho_1) \cdot g_t$
- 7: $r_t \leftarrow \rho_2 \cdot r_{t-1} + (1 - \rho_2) \cdot g_t^2$
- 8: $\hat{s}_t \leftarrow \frac{s_t}{1 - \rho_1^t}$
- 9: $\hat{r}_t \leftarrow \frac{r_t}{1 - \rho_2^t}$
- 10: $\theta_t = \theta_{t-1} - \epsilon(\frac{\hat{s}_t}{\sqrt{\hat{r}_t} + \delta} + \lambda_n \cdot C^t \cdot N(0, \sigma^2))$
- 11: **end while**

$\frac{\text{sum}^2 + \text{max}^2}{2}$, where sum and max represent the total sum and maximum of input data separately.

Proof. The original parameters update can be denoted as $\theta_t \leftarrow \theta_{t-1} - \epsilon \cdot s_t \cdot \nabla J_\theta(h_\theta(x), y)$ without denoising mechanism,

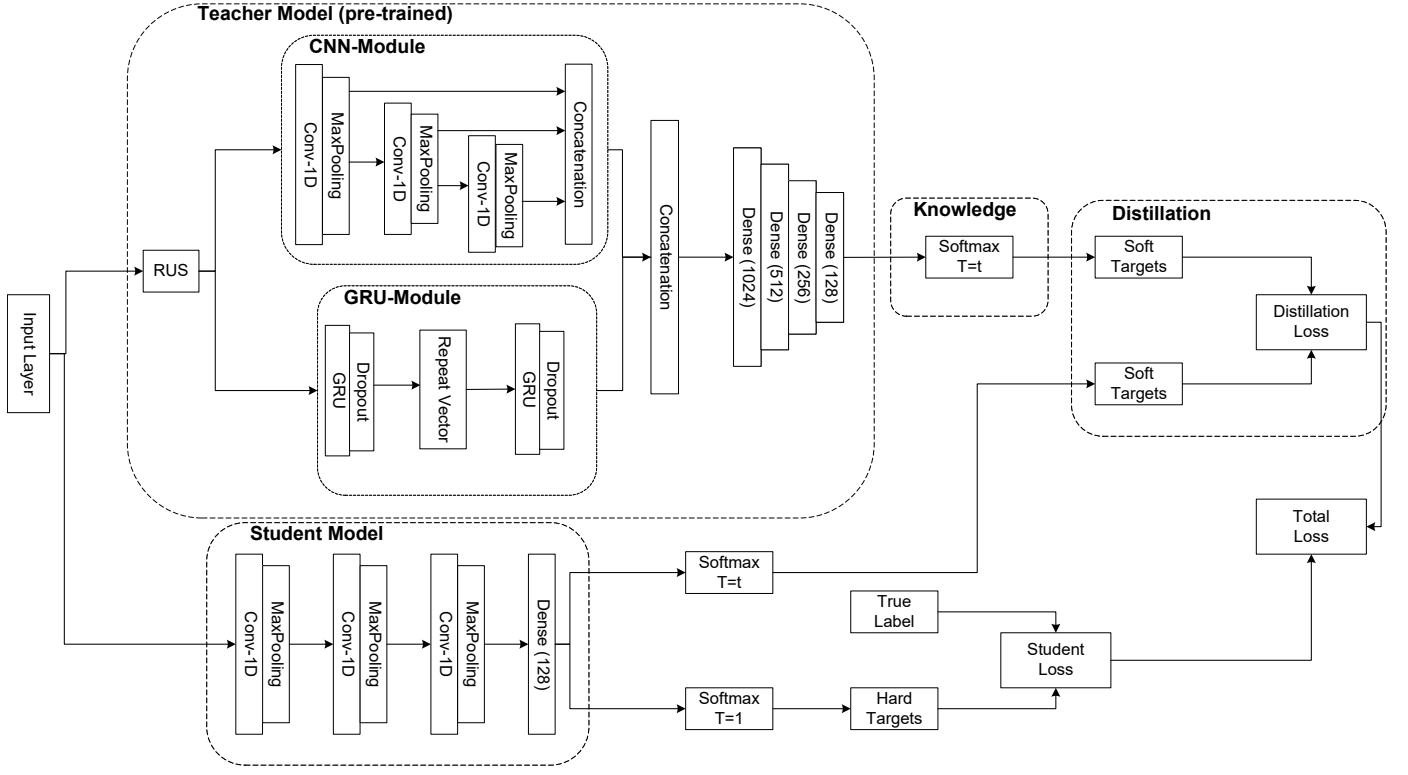


Fig. 3: Architecture of KD-based CNN-GRU Model

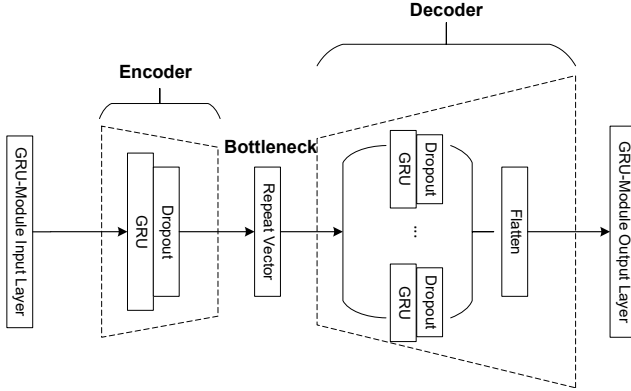


Fig. 4: GRU-based Seq2seq Framework

where s_t represents a complex iteration amount, and $h_\theta(x)$ stands for the prediction of each iteration process. Apart from that, y is defined as the true label, while $\nabla J_\theta(h_\theta(x), y)$ represents the original gradient.

In this proof, we only consider the non-bias prediction function as a matter of convenience. The current prediction function is created by adding Gaussian noise to each parameter. The original and current prediction function have been shown in Equation 1 and 2.

$$h_\theta(x) = \sum_{i=1}^n \theta_i \cdot x_i \quad (1)$$

$$\begin{aligned} h_\theta(x)_{new} &= \sum_{i=1}^n (\theta_i + C^t \cdot N(0, \sigma^2)) x_i \\ &= \sum_{i=1}^n \theta_i \cdot x_i + C^t \cdot N(0, \sigma^2) \cdot \sum_{i=1}^n x_i \end{aligned} \quad (2)$$

Considering Equation 2, the previous θ_t can be refreshed as $\theta_t^j \leftarrow \theta_{t-1}^j - \epsilon \cdot s_t \cdot (\nabla J_\theta(h_\theta(x), y) + C^t \cdot N(0, \sigma^2) \cdot x_j \cdot \sum_{i=1}^n x_i)$, where θ_t^j and θ_{t-1}^j are the j^{th} parameters from step t and step $t-1$. As Algorithm 1, $x_j \cdot \sum_{i=1}^n x_i$ is expressed by λ_n . Referring to Cauchy-Schwarz inequality, the upper bound of λ_n can be set as $\frac{sum^2 + max^2}{2}$.

V. RESULT AND ANALYSIS

In this section, we evaluate the performance of our algorithms with NSL-KDD and a real-world dataset from Smart PiCar testbed [20]. A variety of metrics are exploited to analyze the accuracy of our IDS, as well as the trade-off between accuracy and privacy. The experiments are carried out via Keras API and Tensorflow-Federated framework, executed on Ubuntu 20.04.5 LTS with Intel® Core™ i7-10750H CPU and NVIDIA Quadro T1000 with Max-Q Design (RAM 32GB).

A. Intrusion Detection via Knowledge Distillation

1) *Tuning of Hyper-parameters:* According to [17], it is undeniable that the weighted factor λ and temperature T play the main role for the tuning of hyper-parameter. In terms of the results of multiple experiments, we denote $\lambda = 0.05$ and

$T = 3$ as the hyper-parameters in NSL-KDD, while $\lambda = 0.007$ and $T = 2$ in the real-world data. The Adam optimizer is employed as a reference to assess our denoising Adam in Section V-B. Categorical cross entropy and KL divergence are assigned as the loss function of student/teacher model and distiller respectively.

2) *Evaluation of KD-based Intrusion Detection Algorithm:* Under the above configuration of hyperparameter, the loss function of training set and validation set are shown in Fig. 5, where validation student loss was guided to approach distillation loss by teacher model. Consequently, results of validation student loss will shrink to minima with much less time, which is easier for deployment in smart airport.

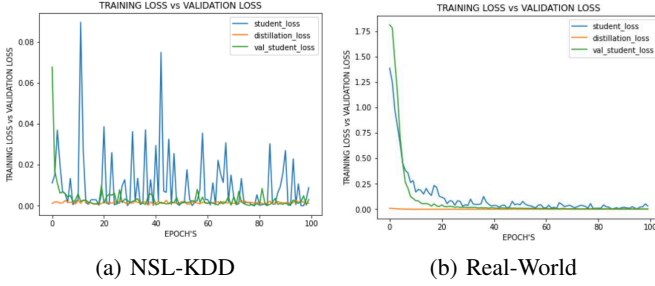


Fig. 5: Loss Function Comparison

To verify the performance of imbalanced classification, the confusion matrix generated by our algorithms is analysed with the current imbalanced learning algorithms in Fig. 6 and Fig. 7, which indicate RUS, Focal Loss and our algorithm can extract more features from the minority class, including R2L, U2R and MITM. In contrast to RUS and Focal Loss, our algorithm holds more properties from the majority classes, the classification accuracy of which won't be reduced too much even with the undersampling data.

B. Trade-off between Privacy and Accuracy

To validate the trade-off performance of denoising adaptive parameter update mechanism, our approaches are compared against the state-of-the-art via different noise multiplier. As an assumption, the local data source will be handled under independent identity distribution (IID) due to the characteristics of the third-party IoT system and the prerequisite of DP-federated learning.

Our experiments are conducted on some pre-configuration, where our datasets are divided into 10 clients and one server, which intend to pursue the private budget up to $(0.1, 10^{-5}) - DP$. In that case, the communication round will be set as 50 to examine whether our mechanism outperforms the conventional Adam optimizer.

The merit of our denoising parameter update mechanism is revealed by Fig. 8 and Fig. 9, which analyse the performance of our mechanism with original adaptive quantile clipping and zero-noise condition. According to Fig. 8 and Fig. 9, our proposed mechanisms improve the convergency efficiency of the original Adam optimizer. Moreover, the adverse impact on converging value is relieved by timely denoising, especially for

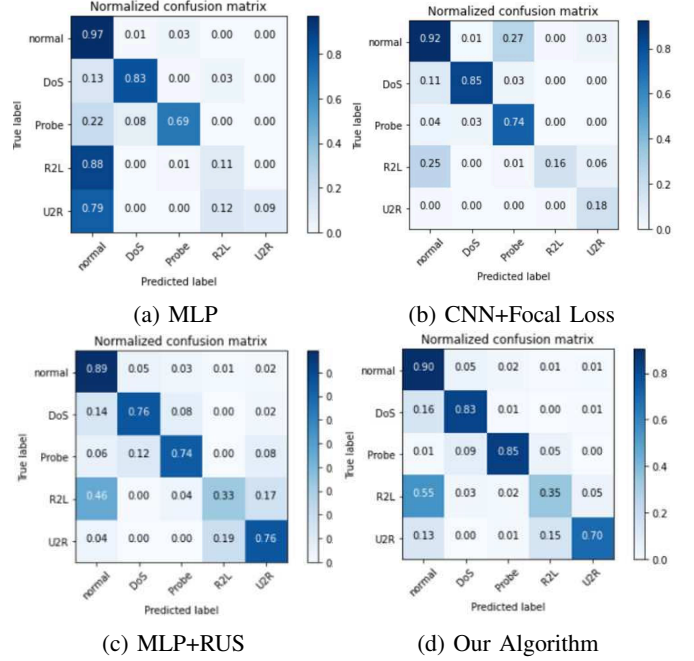


Fig. 6: Confusion Matrix with NSL-KDD

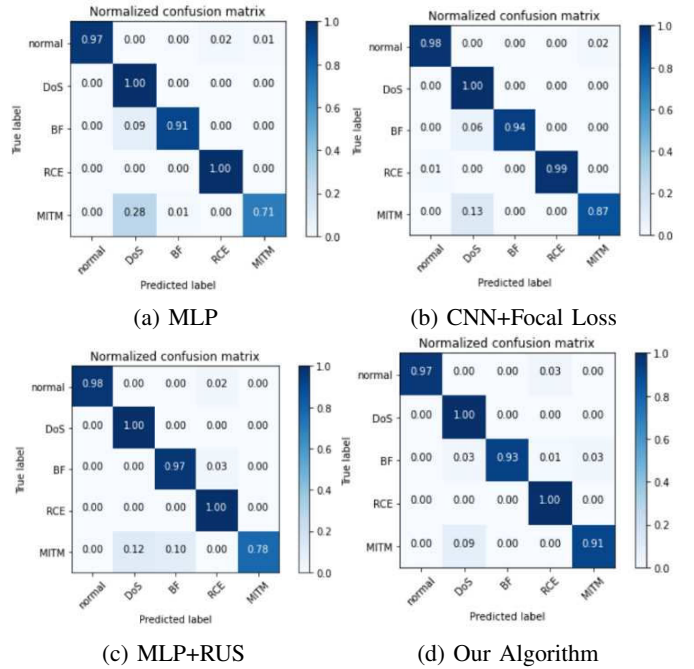


Fig. 7: Confusion Matrix with Real-World Data

high noise multiplier. Therefore, our mechanism is expected to be more appropriate for future target application environments that are time-sensitive, such as smart airports.

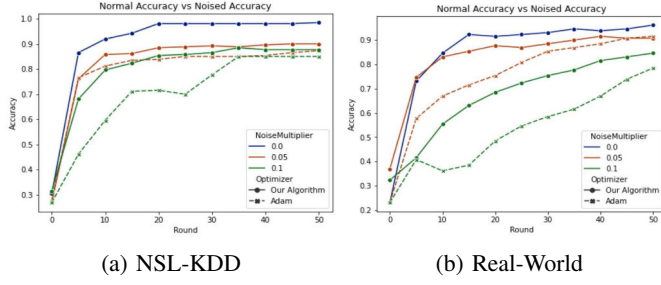


Fig. 8: Training Accuracy with Different Noise Multiplier

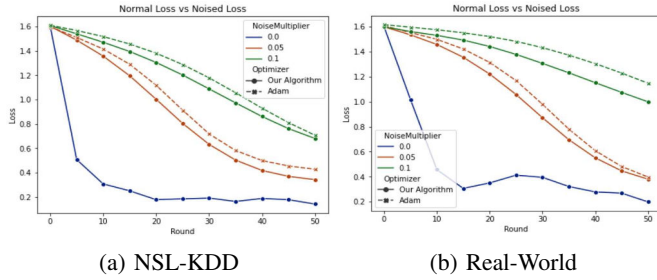


Fig. 9: Loss Function with Different Noise Multiplier

VI. CONCLUSIONS

In this paper, we design a KD-based DP-federated IDS and denoising parameter update mechanism. In contrast with the existing IDS, our system increases the effectiveness to recognize the potential malicious actions. In addition, better trade-off between privacy and utility is achieved by our proposed optimizer for DP-federated learning, which can converge the value close to the non-noise situation even with a high noise multiplier. In our experiments, the accuracy improvement can be achieved up to 30% for detection of the minority malicious actions, while we can ensure the accuracy to converge within 10 rounds even enhancing noise. As a conclusion, our novel IDS is more applicable to detect the cyber-security of smart airport that is stringent to efficiency and accuracy. In the future, we intend to upgrade the trade-off performance of local differential privacy and design a new learning mechanism for knowledge distillation structure.

ACKNOWLEDGMENT

This research was supported by grants received from the UK Department for Transport (DfT) and Home Office, and contributions from NCC Group under the programme of Future Aviation Security Solutions Industrial PhD Partnerships.

REFERENCES

[1] S. Al-Rubaye and A. Tsourdos, 'Airport Connectivity Optimization for 5G Ultra-Dense Networks', *IEEE Trans Cogn Commun Netw*, vol. 6, no. 3, pp. 980–989, Sep. 2020, doi: 10.1109/TCCN.2020.3000509.

[2] H. Brendan McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, 'Communication-Efficient Learning of Deep Networks from Decentralized Data', *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017*, Feb. 2016, doi: 10.48550/arxiv.1602.05629.

[3] N. Peiravian and X. Zhu, 'Machine learning for Android malware detection using permission and API calls', *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, pp. 300–305, 2013, doi: 10.1109/ICTAI.2013.53.

[4] H. Kaur, H. S. Pannu, and A. K. Malhi, 'A Systematic Review on Imbalanced Data Challenges in Machine Learning', *ACM Computing Surveys (CSUR)*, vol. 52, no. 4, Aug. 2019, doi: 10.1145/3343440.

[5] H. Lee, M. Park, and J. Kim, 'Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning', *Proceedings - International Conference on Image Processing, ICIP*, vol. 2016-August, pp. 3713–3717, Aug. 2016, doi: 10.1109/ICIP.2016.7533053.

[6] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, and P. J. Kennedy, 'Training deep neural networks on imbalanced data sets', *Proceedings of the International Joint Conference on Neural Networks*, vol. 2016-October, pp. 4368–4374, Oct. 2016, doi: 10.1109/IJCNN.2016.7727770.

[7] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, 'Focal Loss for Dense Object Detection', *IEEE Trans Pattern Anal Mach Intell*, vol. 42, no. 02, pp. 318–327, Feb. 2020, doi: 10.1109/TPAMI.2018.2858826.

[8] C. Huang, Y. Li, C. C. Loy, and X. Tang, 'Learning deep representation for imbalanced classification', *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 5375–5384, Dec. 2016, doi: 10.1109/CVPR.2016.580.

[9] R. Zhao, T. Tang, G. Gui, and Z. Xue, 'A Lightweight Semi-Supervised Learning Method Based on Consistency Regularization for Intrusion Detection', *ICC 2022 - IEEE International Conference on Communications*, pp. 3124–3129, May 2022, doi: 10.1109/ICC45855.2022.9838567.

[10] M. Abadi et al., 'Deep Learning with Differential Privacy', *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, doi: 10.1145/2976749.

[11] S. U. Stich, 'The Error-Feedback Framework: Better Rates for SGD with Delayed Gradients and Compressed Updates', *Journal of Machine Learning Research*, vol. 21, pp. 1–36, 2020, Accessed: Aug. 20, 2022. [Online]. Available: <http://jmlr.org/papers/v21/19-748.html>.

[12] G. Andrew, O. Thakkar, H. B. McMahan, and S. Ramaswamy, 'Differentially Private Learning with Adaptive Clipping', *Adv Neural Inf Process Syst*, vol. 21, pp. 17455–17466, May 2019, doi: 10.48550/arxiv.1905.03871.

[13] C. Zhou, A. Fu, S. Yu, W. Yang, H. Wang, and Y. Zhang, 'Privacy-Preserving Federated Learning in Fog Computing', *IEEE Internet Things J*, vol. 7, no. 11, pp. 10782–10793, Nov. 2020, doi: 10.1109/IJOT.2020.2987958.

[14] J. Dean et al., 'Large Scale Distributed Deep Networks', *Adv Neural Inf Process Syst*, vol. 25, 2012.

[15] R. Yu, G. Xue, and X. Zhang, 'Application Provisioning in FOG Computing-enabled Internet-of-Things: A Network Perspective', *Proceedings - IEEE INFOCOM*, vol. 2018-April, pp. 783–791, Oct. 2018, doi: 10.1109/INFOCOM.2018.8486269.

[16] N. Koroniotis, N. Moustafa, F. Schiavo, P. Gauravaram, and H. Janicke, 'A Holistic Review of Cybersecurity and Reliability Perspectives in Smart Airports', *IEEE Access*, vol. 8, pp. 209802–209834, 2020, doi: 10.1109/ACCESS.2020.3036728.

[17] G. Hinton, O. Vinyals, and J. Dean, 'Distilling the Knowledge in a Neural Network', Mar. 2015, doi: 10.48550/arxiv.1503.02531.

[18] D. P. Kingma and J. L. Ba, 'Adam: A Method for Stochastic Optimization', *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, Dec. 2014, doi: 10.48550/arxiv.1412.6980.

[19] I. Loshchilov and F. Hutter, 'Decoupled Weight Decay Regularization', *7th International Conference on Learning Representations, ICLR 2019*, Nov. 2017, doi: 10.48550/arxiv.1711.05101.

[20] A. Aloose, S. Al-Rubaye, A. Zolotas, H. He, and C. Shaw, 'A Novel Approach for Detecting Cyberattacks in Embedded Systems Based on Anomalous Patterns of Resource Utilization-Part i', *IEEE Access*, vol. 9, pp. 103204–103229, 2021, doi: 10.1109/ACCESS.2021.3088395.

[21] C. Dwork and A. Roth, 'The Algorithmic Foundations of Differential Privacy', *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–487, Aug. 2014, doi: 10.1561/04000000042.



DSPACE

<https://dspace.org/>

Differentially-private federated intrusion detection via knowledge distillation in third-party IoT systems of smart airports

Chen, Yang; Al-Rubaye, Saba; Tsourdos, Antonios; Baker, Lawrence; Gillingham, Colin

2023-10-23

Institute of Electrical and Electronics Engineers (IEEE)

<https://doi.org/10.1109/ICC45041.2023.10279722>

Chen Y, Al-Rubaye S, Tsourdos A, et al., (2023) Differentially-private federated intrusion detection via knowledge distillation in third-party IoT systems of smart airports. In: ICC 2023 - IEEE International Conference on Communications, 28 May - 1 June 2023, Rome, Italy