Enhancing Cyber-threat detection coupling Deep Neural Ensemble Learning with XAI

Malik Al-Essa[1,*,†], Giuseppina Andresini[1,2,*,†], Annalisa Appice[1,2,†] and Donato Malerba[1,2,†]

[1]University of Bari Aldo Moro, Bari, Italy
[2]Consorzio Interuniversitario Nazionale per l'Informatica - CINI, Bari, Italy

Abstract

In the digital age, the use of deep learning is one of the most powerful machine learning paradigms for cybersecurity. Despite the amazing results recently achieved with deep learning methods in securing the digital infrastructures of modern organizations, the security of neural models can easily be jeopardized by adversarial attacks. This article describes a recently published cyber-threat detection method, named PANACEA, that combines Adversarial Training and eXplainable Artificial Intelligence (XAI) to increase the diversity of multiple neural models fused together through a neural ensemble system. Experiments carried out on several benchmark cybersecurity datasets show the beneficial effects of the proposed combination of Adversarial Training, Ensemble Learning and XAI on the accuracy of multi-class classifications of cyber-data achieved by the neural method.

Keywords

Ensemble Learning, Adversarial Training, eXplainable Artificial Intelligence, Cyber-threat Detection

1. Introduction

During the last decade, the cybersecurity literature has conferred a high-level role in deep learning as a powerful learning paradigm to detect ever-evolving cyber-threats in modern security systems. In particular, recent cybersecurity studies have shown that deep learning performance can be further strengthened with ensemble learning systems [1] that are able to obtain better generalization by reducing the dispersion of predictions of single models and gaining model accuracy. However, selecting the ensemble member models based on the local model accuracy may lead to the issue of excessive ensemble because the performance of the ensemble system may not be significantly improved by some of the selected models. Therefore, several scholars encourage the diversity among individual models of deep ensembles, in addition to the accuracy of individual models, to learn diverse aspects of training data [2].

In [3, 4], we have recently proposed a new XAI-based method, named PANACEA, that is mainly founded on the idea that different sub-areas of the input feature space can be equally relevant to achieve a correct decision for multiple samples produced in the same situation. Hence, an accurate ensemble system may be produced through the fusion of base models that perform decisions which give more importance to different sub-areas of the input feature space. For this purpose, we use the XAI DALEX framework [5] to explain the global feature importance in neural models. Specifically, we adopt a combination of XAI and clustering to select ensemble base models that achieve high explanation diversity. Finally, we use a multi-headed neural network architecture that fine-tunes simultaneously base neural models selected through DALEX-based clustering, by taking advantage of a back-propagation strategy to share knowledge among multiple base models incorporated as sub-network blocks in the ensemble system.

Motivations for adopting this neural ensemble method in cybersecurity problems can be mainly founded in the peculiarities of the network intrusion detection problems, where samples of different attack families commonly have signatures involving different features. For example, as illustrated by [6], "the time between the SYN ACK and the ACK response" is relevant for detecting shellcode intrusions, while it becomes less important when detecting other types of attacks. Shellcode, in fact, is an exploiting attack in which the attacker penetrates a piece of code from a shell to control a target machine using the standard TCP/IP socket connections.

Based upon these considerations, our point of view is that being able to fuse deep neural models that give relevance to different network traffic feature signatures (and, consequently, input feature sub-spaces) may help in improving the accuracy of a multi-class deep neural ensemble trained to recognize different cyber-attack patterns such as various categories of network traffic intru-

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy
*Corresponding author.
†These authors contributed equally.
malik.alessa@uniba.it (M. Al-Essa); giuseppina.andresini@uniba.it (G. Andresini); annalisa.appice@uniba.it (A. Appice); donato.malerba@uniba.it (D. Malerba)
0000-0002-0892-975X (M. Al-Essa); 0000-0002-5272-644X (G. Andresini); 0000-0001-9840-844X (A. Appice); 0000-0001-8432-4608 (D. Malerba)

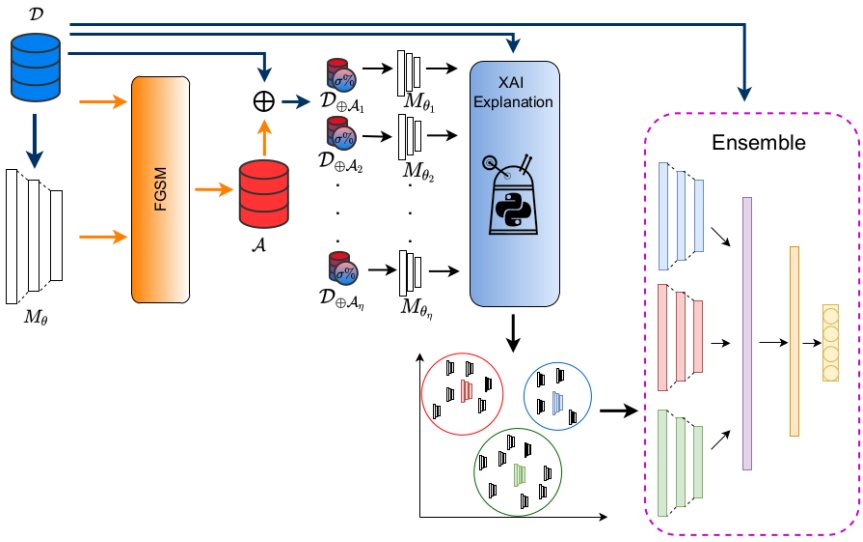


Figure 1: Schema of PANACEA

sions. Our argument is mainly supported by experiments performed with three benchmark network intrusion detection datasets, namely NSL-KDD, UNSW- NB15 and CICIDS17, that comprise multiple real categories of network traffic intrusions (comprising rare attacks). In addition, to explore the adaptability of the proposed method to other cyber-threat detection problems, we also evaluated the effectiveness proposed method in a benchmark malware detection problem, namely CICMalDroid20, since we expect that, similarly to network traffic intrusions, different malware categories may have diverse feature signatures.

This paper summarises some of the main results published in [3, 4]. The PANACEA method is presented in Section 2. Section 3 illustrates the main results achieved in the evaluation of the proposed method. Finally, Section 4 draws conclusions and sketches future research directions.

2. PANACEA method

Let us consider a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ of N training samples, where $\mathbf{x} \in \mathbb{R}^d$ is a d-dimensional vector of input features that describe cyber-data samples, and $y \in \{1, \ldots, K\}$ is the label variable with K classes (*benign* class and several categories of *cyber-threats*), according to labels of samples historically collected.

The PANACEA method, illustrated in Figure 1, is based on the following steps:

- The training of an initial neural model $M_\theta : \mathbb{R}^d \mapsto Y$ with parameter θ learned from $\mathcal{D}$.
- The generation of an adversarial set $\mathcal{A}$ produced by $\mathcal{D}$ with data perturbation threshold ϵ by using M_θ. The adversarial samples are produced using the FGSM algorithm.
- The training of η neural model candidates learned from $\mathcal{D}$, augmented with subsets of σ adversarial samples randomly selected from $\mathcal{A}$.
- The use of a post-hoc global XAI technique, namely DALEX, to explain the decisions of neural model candidates and generate a feature-vector explanation of each neural model candidate.
- A clustering stage (k-medoids method) to group neural model candidates with similar feature explanation vectors in the same clusters, and neural model candidates with dissimilar feature explanation vectors in separate clusters. Since each cluster medoid is a neural model candidate that acts as the cluster's prototype, k medoids (chosen using the Elbow method) are selected as the base neural models for the ensemble fusion.
- A multi-headed neural network that fuses together base neural models selected through clustering.

Notice that the performance of PANACEA may depend on the input parameters:(1) ϵ that represents the amount of data perturbation considered to generate adversarial samples; (2) σ that defines the number of adversarial samples randomly selected for learning each neural model candidate with the adversarial training strategy; (3) η that is the number of distinct neural model candidates learned with the adversarial training strategy. In general,

the perturbation ϵ is selected as a small value in the range between 0 and 0.1 [7], to scale the noise and ensure that perturbations are small enough to remain undetected to the human eye, but large enough to fool the attacked neural model. In PANACEA the value of ϵ is automatically selected based on the characteristics of adversarial samples. This is based on the idea that the value at which a lower ϵ stops perturbing training samples, by diminishing the number of misclassified adversarial training samples, may correspond to an adequate value of ϵ for gaining accuracy with the adversarial training strategy. Based on this idea, for each ϵ in the range [0, 0.1], the adversarial set $\mathcal{A}_\epsilon$, produced from the original training set with initial neural model M_θ as target model, is considered. The Overall Accuracy (OA) of M_θ is computed on each $\mathcal{A}_\epsilon$ and the Elbow method is used to pick the knee of the OA($\mathcal{A}_\epsilon$) curve as the value of ϵ. Notably, this procedure for the automatic selection of ϵ is independent of both σ and η that remain user-defined parameters

3. Evaluation study

Four benchmark multi-class datasets, i.e., NSL-KDD, UNSW-NB15, CICIDS17 (network security datasets) and CICMalDroid20 (malware security dataset) were considered to evaluate the performance of PANACEA. Experiments were conducted by dividing each dataset into training set and testing set. The detailed description of the experimental set-up is reported in [4].

The most of experiments were conducted with $\sigma = 5\%$ and 10% of the training set size, considering the values of elbow ϵ automatically selected with the Elbow method and fixing $\eta = 100$ for all datasets. However, further experiments exploring the sensitivity of the performance of PANACEA to the number of models η are illustrated in [4].

Table 1 reports the number of neural models (k) that the clustering step of PANACEA selected for the ensemble fusion, as well as WeightedF1, MacroF1 and OA of PANACEA in the considered experimental setting. All the accuracy metrics were measured on the testing set of each dataset. As BASELINE, we considered the deep neural network that was trained in the first step of PANACEA as the initial neural model for the adversarial sample production. We recall that the number of clusters k was automatically identified during the clustering step of PANACEA. The results show that PANACEA outperforms BASELINE, independently of the number σ of adversarial samples processed in UNSW-NB15, CICIDS17 and CICMalDroid20. In these three datasets, the gain in accuracy is commonly observed equally along WeightedF1, MacroF1 and OA. The only exception is the MacroF1 of PANACEA with $\sigma = 5\%$ in UNSW-NB15. However, both WeightedF1 and OA of PANACEA outperform WeightedF1 and OA of BASELINE also in this configuration. In addition, there is at least one tested configuration of PANACEA that outperforms BASELINE in NSL-KDD. Finally, also in NSL-KDD the gain in accuracy is observed along WeightedF1 and OA, but not along MacroF1. This is due to the presence of minority classes in both NSL-KDD and UNSW-NB15. In fact, in both datasets, the ensemble strategy allows us to gain accuracy by better classifying samples of majority classes, while we may lose accuracy by classifying samples of minority classes. This intuition is confirmed by the analysis of detailed F1 per class, reported [4]. Notably [4] also reports an extensive analysis of the accuracy performance of PANACEA compared to several, recent state-of-the-art competitors, as well as the analysis of the accuracy performance achieved by PANACEA by using PGD, DeepFool and LowProFool in place of FGSM.

To examine in-depth diversity, Figure 2 depicts the top-15 relevant features on the global decisions of the base neural models selected in NSL-KDD. Feature ranking maps show how diverse input features play prominent roles in explaining the decisions of the base neural models selected for the ensemble fusion in PANACEA. For example, the input feature “serror_rate", that is ranked in third place for the neural model medoids of clusters 2, 3 and 7 of NSL-KDD, is not even in the top-15 for the medoid of cluster 6. Notably, humans may inspect this explanation result to confirm the selection of neural model candidates automatically selected by PANACEA or perform a manual update of the automatic selection (with model deletions or additions) according to background knowledge.

We complete this article by illustrating an example that shows how the ensemble model of PANACEA gains accuracy in a cyber-threat detection task compared to the single model of BASELINE. For this purpose, we consider an R2L sample of the test set of NSL-KDD that was wrongly classified by BASELINE in the class Normal, while it was correctly recognised in the class R2L by PANACEA. We analyse this sample by using SHAP that is a local algorithm to measure the effect of an input feature on the assignment of a sample to a class with a neural model. Figure 3 shows the five most important input features identified by SHAP to see the sample in the class R2L with the models learned by both BASELINE and PANACEA. Let us consider that only PANACEA predicted this sample in the class R2L.

Both BASELINE and PANACEA share the same top-3 features, i.e., *service_http*, *service_ftp_data* and *dst_host_srv_count*. Notably, these three features are recognised as important to detect R2L attacks also in [8]. The input feature in the fourth place of the feature ranking of PANACEA is *protocol_type_tcp* that does not appear in the feature ranking of BASELINE. The authors of [9] report that the simultaneous use of the TCP protocol and the FTP service is to be considered a symptom of

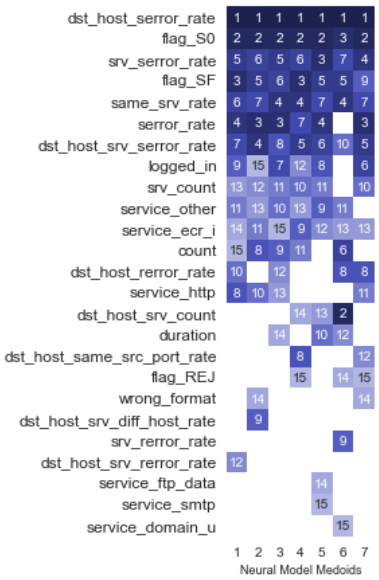


Figure 2: Top-15 feature ranking map of the base neural models selected through the clustering step of PANACEA in NSL-KDD

a possible Warez Master attack in network traffic. Warez Master is a subcategory of R2L attacks, where attackers exploit a system bug associated with FTP to send packets of illegal software to a target host [9]. We note that FTP is a service based on the TCP protocol. Therefore, this example shows how the ensemble model of PANACEA manages to bring out the existence of feature patterns useful for the recognition of attack classes that are often ignored by the single model of BASELINE. These conclusions are also supported by the study of [8], that identifies both *service_ftp_data* and *protocol_type_tcp* features as the most important features to detect R2L attacks. In addition, BASELINE, differently from PANACEA, identifies *serror_rate* as one of the most relevant features for recognizing the sample as an R2L attack. However, neither [8] nor [9] identify this feature as one of the most prominent features for this type of attack.

In short, the emergence of *protocol_type_tcp* can be considered as an important input feature instead of *serror_rate* motivates the ability of PANACEA in correctly recognising the considered R2L sample and, in general, the ability of outperforming BASELINE in the recognition of R2L attacks (that passes from F1(R2L)=0.55 for BASELINE to F1(R2L)=0.64 for PANACEA.

4. Conclusion

In this paper, we have summarized the main results of our newest research illustrated in [4], where we have described a deep learning method for multi-class classification of cyber-data.[1] The proposed method trains an ensemble of base neural models, whose weights are initialised with an adversarial training strategy. We use an XAI-based approach to increase the diversity of the neural models selected to be fused together through the ensemble system.

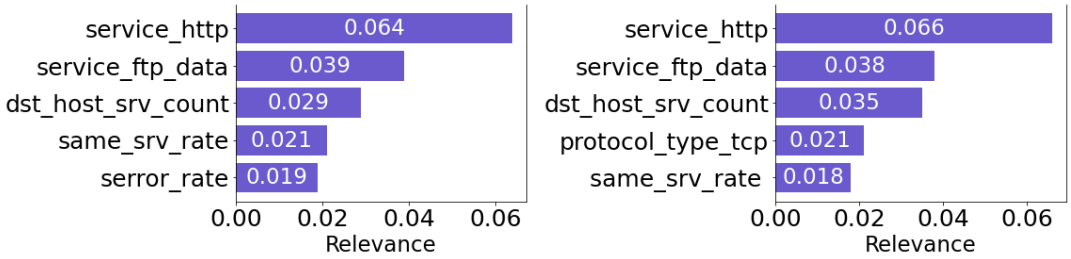
Notably, this article delves into one of the current research directions carried out by Laboratory KDDE (https://kdde.di.uniba.it/) at the University of Bari "Aldo Moro", which aims at exploring a Symbiotic AI approach to Cybersecurity. The team has recently published several papers in this field (e.g., [10, 11, 6, 12, 13]). In particular, the newest studies [3, 4] stay under the umbrella of Symbiotic AI, as they explore how Explainability of AI systems can be leveraged as a valuable means to allow deep neural models to gain accuracy under critical conditions commonly occurring in cybersecurity problems, e.g., class imbalance, attack signature diversity. They provide a mechanism that can explain to humans how the candidate models are selected for ensemble systems. On the other side, these studies stay under the umbrella of Cybersecurity, as XAI is used to improve the performance of a cyber-threat detection ensemble model on multiple attack categories by allowing us to identify and use the multiple input sub-space that can help in detecting attacks with diverse signature. In addition, the use of XAI tools allows us to perform a step forward to gain the trust of stakeholders in AI decisions. In fact, it allows us to disclose cyber-data patterns that are hidden in how the AI models achieve a decision and explain why a black box model can actually achieve higher performance than another one in cyber-threat detection.

By continuing along this research direction, the team is working on the use of XAI to examine and explain the evasion ability of state-of-the-art attack methods formulated for Windows PE malware detection problems. In addition, the team is investigating emerging learning frameworks (such as distillation) to leverage explanations disclosed through attention layers to improve the performance of deep neural models trained for cyber-threat detection.

5. Acknowledgments

Malik AL-Essa is supported by PON RI 2014-2020 - Machine Learning per l'Investigazione di Cyber-minacce e la Cyber-difesa - CUP H98B20000970007. Giuseppina Andresini is supported by the project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI, under the NRRP MUR program funded by the NextGenerationEU.

[1]The original research illustrated in [4] was published under Creative Commons License Attribution 4.0 (CC BY 4.0) https://creativecommons.org/licenses/by/4.0/



(a) BASELINE

(b) PANACEA

Figure 3: Top-5 input features considered by both BASELINE (3a) and PANACEA (3b) to recognize an R2L attack in the class R2L

Table 1
WeightedF1, MacroF1 and OA of PANACEA with $\sigma = 5\%$ and 10% of the training set size and BASELINE. k denotes the number of distinct neural models automatically selected in the clustering step of PANACEA over $\eta = 100$ neural model candidates. The best results are in bold.

dataset	method	k	WeightedF1	MacroF1	OA
NSL-KDD	BASELINE	-	0.80	**0.64**	0.80
	PANACEA ($\sigma = 5\%$)	8	0.79	0.60	0.80
	PANACEA ($\sigma = 10\%$)	7	**0.83**	**0.64**	**0.84**
UNSW-NB15	BASELINE	-	0.74	0.42	0.74
	PANACEA ($\sigma = 5\%$)	12	0.77	0.41	0.77
	PANACEA ($\sigma = 10\%$)	11	**0.78**	**0.44**	**0.77**
CICIDS17	BASELINE	-	0.92	0.64	0.91
	PANACEA ($\sigma = 5\%$)	9	0.98	0.73	0.97
	PANACEA ($\sigma = 10\%$)	9	**0.99**	**0.94**	**0.99**
CICMalDroid20	BASELINE	-	0.83	0.80	0.83
	PANACEA ($\sigma = 5\%$)	13	**0.90**	**0.88**	**0.89**
	PANACEA ($\sigma = 10\%$)	18	0.85	0.83	0.86

Annalisa Appice and Donato Malerba are partially supported by project SERICS (PE00000014) under the NRRP MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

References

[1] B. A. Tama, S. Lim, Ensemble learning for intrusion detection systems: A systematic mapping study and cross-benchmark evaluation, Computer Science Review 39 (2021) 1–27. doi:10.1016/j.cosrev.2020.100357.

[2] X. Dong, Z. Yu, W. Cao, Y. Shi, Q. Ma, A survey on ensemble learning, Frontiers of Computer Science 14 (2020) 241–258. doi:10.1007/s11704-019-8208-z.

[3] M. Al-Essa, G. Andresini, A. Appice, D. Malerba, Panacea: A neural model ensemble for cyber-threat detection, 2023. doi:10.1109/DSAA60987.2023.10302518.

[4] M. AL-Essa, G. Andresini, A. Appice, D. Malerba, Panacea: a neural model ensemble for cyber-threat detection, Machine Learning (2024). doi:10.1007/s10994-023-06470-2.

[5] P. Biecek, DALEX: Explainers for complex predictive models in R, Journal of Machine Learning Research 19 (2018) 1–5.

[6] G. Andresini, A. Appice, F. P. Caforio, D. Malerba, G. Vessio, ROULETTE: A neural attention multi-output model for explainable network intrusion detection, Expert Systems with Applications (2022) 117144. doi:10.1016/j.eswa.2022.117144.

[7] T. Bai, J. Luo, J. Zhao, B. Wen, Q. Wang, Recent advances in adversarial training for adversarial robustness, in: 30th International Joint Conference on Artificial Intelligence, IJCAI 2021, IJCAI.ORG, 2021, pp. 4312–4321. doi:10.24963/ijcai.2021/591.

[8] M. Sabhnani, G. Serpen, KDD feature set complaint heuristic rules for R2L attack detection, in: International Conference on Security and Management, SAM 2003, CSREA Press, 2003, pp. 310–316.

[9] M. Wang, K. Zheng, Y. Yang, X. Wang, An explainable machine learning framework for intrusion detection systems, IEEE Access 8 (2020) 73127–73141. doi:10.1109/ACCESS.2020.2988359.

[10] F. P. Caforio, G. Andresini, G. Vessio, A. Appice, D. Malerba, Leveraging grad-cam to improve the accuracy of network intrusion detection systems, in: 24th Conference on Discovery Science , DS 2021, volume 12986 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 385–400.
[11] G. Andresini, F. Pendlebury, F. Pierazzi, C. Loglisci, A. Appice, L. Cavallaro, INSOMNIA: towards concept-drift robustness in network intrusion detection, in: 14th ACM Workshop on Artificial Intelligence and Security, ACM, 2021, pp. 111–122.
[12] M. AL-Essa, G. Andresini, A. Appice, D. Malerba, XAI to explore robustness of features in adversarial training for cybersecurity, in: Foundations of Intelligent Systems, Springer International Publishing, 2022, pp. 117–126. doi:10.1007/978-3-031-16564-1_12.
[13] M. Al-Essa, G. Andresini, A. Appice, D. Malerba, An XAI-based adversarial training approach for cyber-threat detection, in: 2022 IEEE International Conference on Cyber Science and Technology Congress, CyberSciTech 2023, IEEE, 2022, pp. 1–8. doi:10.1109/DASC/PiCom/CBDCom/Cy55231.2022.9927842.

A. Online Resources

The source code of PANACEA implementation is available online at https://github.com/malikalessa/PANACEA.