Whitepaper

Robustness of Machine Learning Systems

An Overview of Defences against Adversarial AI Attacks

Authors
Niels Brink, Jip van Stijn, Piotr Stachyra,
Olivier Spinnler, Yori Kamphuis

TNO innovation for life

Contents

Introduction

Alan Turing once said: “A computer would deserve to be called intelligent if it could deceive a human into believing that it was human”. Currently, we cannot confirm that any system successfully deceived a human into believing that it is a human. However, there are plenty of cases of computers deceiving other computers, for example by fooling it into thinking a picture of a hamster is actually a burrito (Anley, 2022). The ability to deceive Artificial Intelligence (AI) models has sparked discussion among researchers about their robustness and safety. In order to counter the risks that come with Adversarial AI attacks, a novel study branch has emerged that deals with defence methods against such attacks.

As the world is rapidly becoming aware of the increasing capabilities of AI, promising new applications are expected and being implemented in a plethora of domains. The cyber domain is no exception: cybersecurity applications, including in cyber-physical systems such as factories, power plants, and oil and gas facilities, are being deployed with AI components (Alotaibi & Rassam, 2023). However, the introduction of AI and specifically Machine Learning (ML) technologies could create new attack vectors that can be exploited through Adversarial Machine Learning (AML) techniques (Brink, et al., 2023).

In 2023, the authors of this whitepaper provided an overview of the academic literature on AML and identified five main methods through which ML models may be attacked (Brink, et al., 2023).[1] This overview showed that research on attack methods is developing quickly, and yet we might only be seeing the tip of the iceberg, as some studies may not be published due to confidentiality. Adding to that, there is scant public knowledge about which attacks are being carried out, in the wild, making it difficult to create an accurate depiction of the actual threat landscape. At the same time, the number of studies on defensive methods against AML attacks is increasing rapidly, as the 24 papers published in 2014[2] are dwarfed by the 3848 and 5415 papers published in 2022[3] and 2023[4] respectively. However, this increase is proceeding in a less structured manner: scholars often present breakthrough techniques which in reality are existing methods with relatively small alterations. This results in a diverse range of terminology in this field of research, making it difficult to discern general trends and promising results. Providing a structured overview of AML defence methods is a crucial step towards enabling developers to identify the most popular and promising ways to defend AI-based systems in the emerging threat landscape, improving the robust and secure use of AI (Brink, et al., 2023).

This whitepaper takes up this challenge by structuring existing AML defence mechanisms in the cyber domain, answering the following two research questions:

1. Which defence mechanisms are being discussed in academic literature, and how can they be structured?
2. What are the general trends in AML literature?

This whitepaper first gives a brief overview of the different AML attacks mentioned earlier.

Then, the results of the literature review into the defences against AML attacks are presented in the AML Defence Framework, presented below. This framework was used to uncover overarching trends in the literature, which are described after the AML Defence Framework. The paper finalises with a section of conclusions regarding the research field of defences against AML attacks.

[1] See https://www.tno.nl/en/newsroom/2023/02/first-overview-cyberattack-techniques-ai/. This framework was based on predictive ML models. Recent research suggests that Generative AI may be vulnerable to different, newer types of attacks, such as abuse attacks (Vassilev, Oprea, Fordyce, & Anderson, 2024). The focus of this whitepaper is on more ‘traditional’, predictive ML models.

[2] (Brink, et al., 2023)

[3] Scopus search term: (ALL (adversarial AND machine AND learning)) AND (cyber) AND PUBYEAR = 2022 AND (LIMIT-TO (DOCTYPE , “ar”) OR LIMIT-TO (DOCTYPE , “cp”) OR LIMIT-TO (DOCTYPE , “re”)).

[4] Scopus search term: (ALL (adversarial AND machine AND learning)) AND (cyber) AND PUBYEAR = 2023 AND (LIMIT-TO (DOCTYPE , “ar”) OR LIMIT-TO (DOCTYPE , “cp”) OR LIMIT-TO (DOCTYPE , “re”)).

Background Knowledge: Adversarial ML Attacks

According to Brink et al. (2023), AI-based systems in the cyber domain could be attacked using AML attacks. They furthermore divided those AML attacks into five categories:

1. **Poisoning**: manipulating the training data.
2. **Backdoor**: adding code to the model that ensures normal operation until a specific input is given by the attacker.
3. **Evasion**: manipulating the input to mislead the model.
4. **Membership inference**: using access to the model to learn characteristics about the training data.
5. **Model stealing**: creating a copy of the original model by exploiting access to it.

The different attacks were plotted on the European Telecommunications and Standardisation Institute's (ETSI) ML lifecycle model, which outlines the six stages of a ML model's development, operation, and updates. In this whitepaper, we map the defensive techniques on the ETSI ML lifecycle model in a similar manner. This categorisation aims to create a understanding of the various possibilities for mitigating existing threats to machine learning models, bringing to light general patterns in the defences. When plotting possible defences onto this model, one must keep in mind that defences in a specific phase do not solely counter attacks that target that specific phase. They may also mitigate different attacks. For example, by using multiple models to collectively vote on the main model's output, one could prevent model stealing attacks, but potentially also evasion and poisoning attacks. Thus, this defence in the operational phase could mitigate attacks in the output, input, and preparation phases.

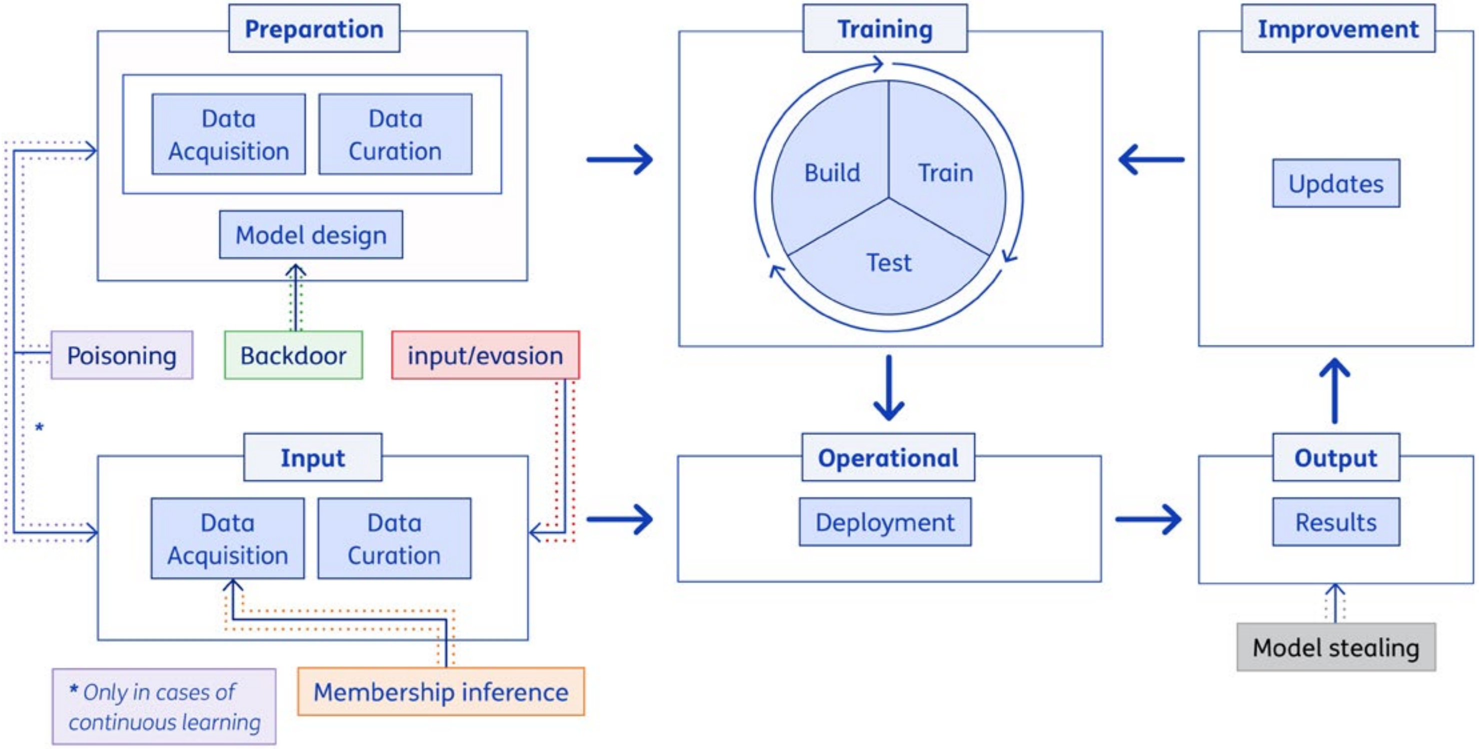

Figure 1: Possibilities to attack the ML life cycle (Brink, et al., 2023), supplement to (ETSI, 2020, p. 11).

AML Defence Framework

As presented above, AML attacks against machine learning models can be mapped to the phases in the ETSI ML life cycle (ETSI, 2020). We propose an extension of ETSI's framework by plotting the defence mechanisms found in the literature onto their ML life cycle. To allow for the analysis of the most relevant papers in this quickly developing field, we collected a large sample of papers through a Scopus database. Next, we assessed them using ASReview, a ML tool for conducting systematic literature reviews (ASReview Lab, 2022). Figure 2 shows the result of this analysis: a set of AML defence categories (or 'families') for each of the ETSI ML life cycle phases. Each of the categories is colour-coded according to the attack type that it is described to protect against.

The following sections explain the ML life cycle phases, the categories of defences within them, and builds on those by adding the specific defences within those categories. The most relevant defences will be explained in more detail in the text, but for those that are not, the papers are also included in the bibliography. Note that the current research has not empirically verified the effectiveness of these defences in practical applications. However, advances in research since the introduction of these attacks allowed us to draw certain conclusions about their effectiveness.

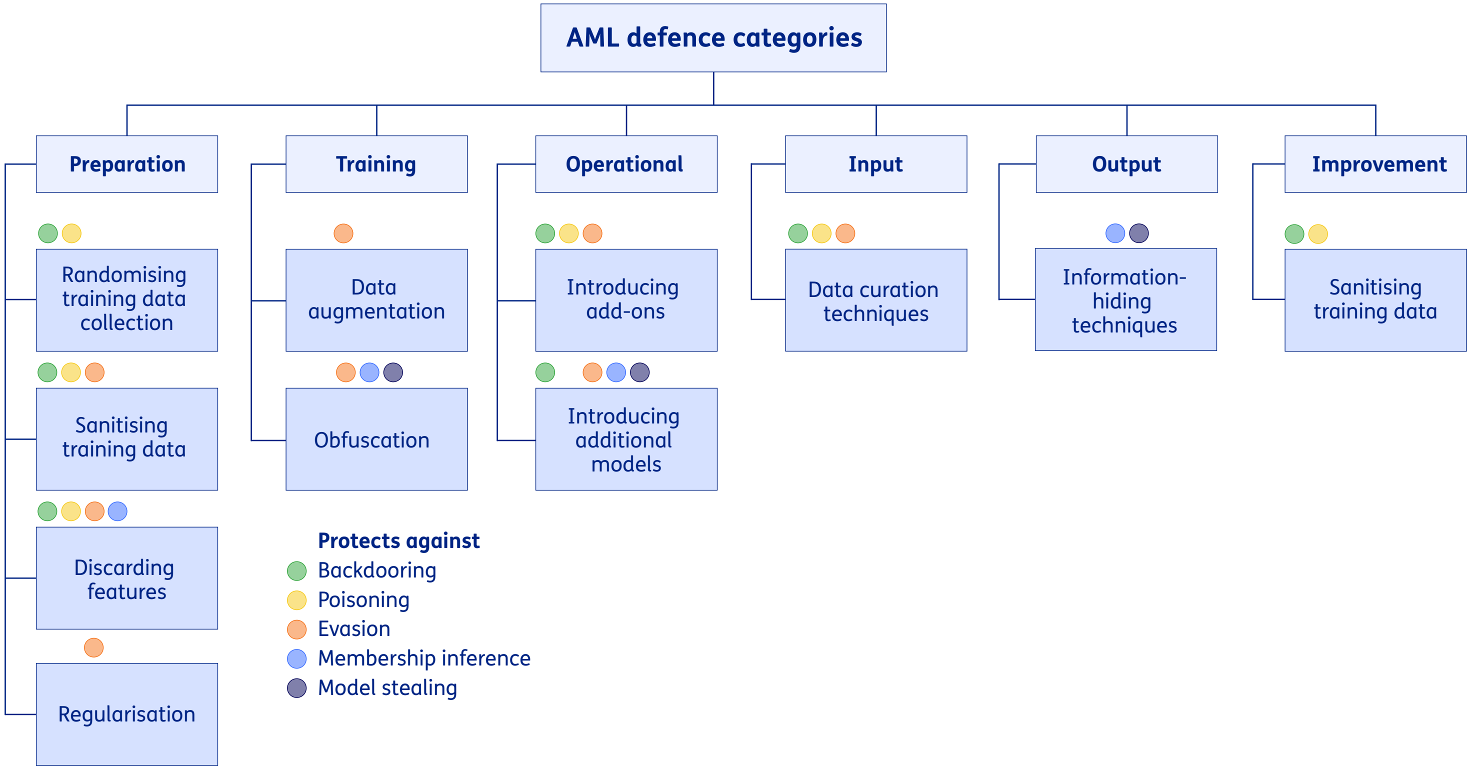

Figure 2: AML defence categories within the ML life cycle.

Preparation phase

The preparation phase consists of data acquisition, data curation, and model design. The end-objective is obtaining a dataset of sufficient quality and a correct format of inputs for the model (ETSI, 2020). In this context, sufficient quality refers to the state in which the dataset represents the phenomenon to a sufficient extent, and which allows the model to produce meaningful and accurate results.

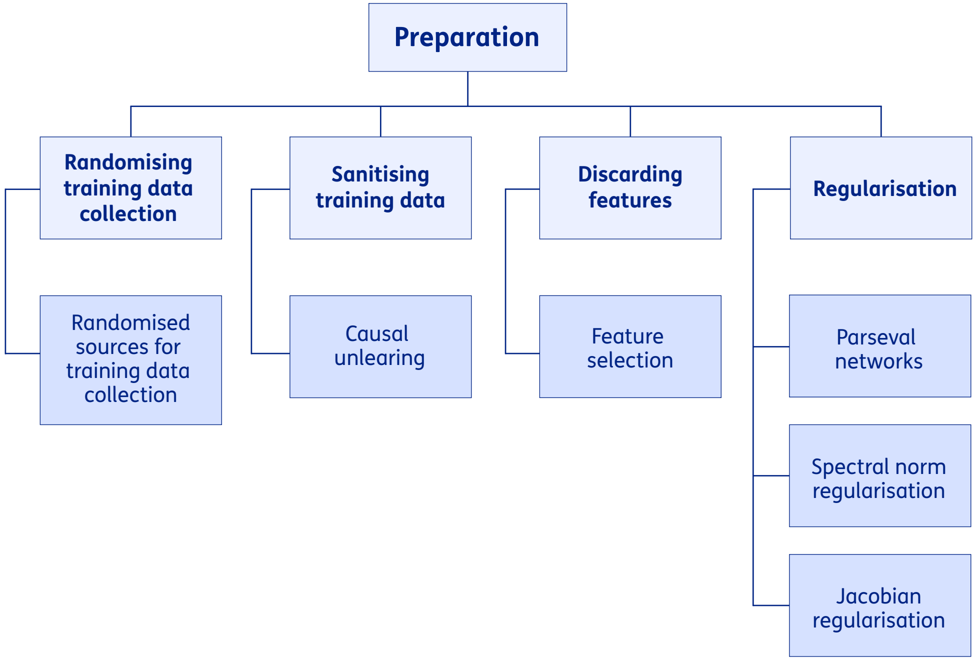

Figure 3: AML defences in the preparation phase.

The literature describes four main streams of defence in the preparation phase. First is **randomisation** of the way that training data is collected. By gathering training data from different sources at different times, it would be more difficult for adversaries to poison a significant amount of the training data (Biggio & Roli, 2018). This is especially challenging for the recently developed models that use datasets of considerable sizes (millions of records). Examples include computer vision datasets such as LAION-400M or COYO-700M, that contain 400 million and 700 million images respectively. These are hosted as an image's URL with their corresponding labels. Regardless of the file referred to by the URL, the URL is immutable, whereas the served content can be altered. When the domains of the hosted images expire, the attacker can purchase them and include the poisoned samples in the training set for any model which uses these datasets (Carlini, et al., 2023). Randomisation of the data retrieval processes can help to mitigate the problem of poisoned data by invalidating the attacker's assumption about the records that will be accessed and about the time that they will be collected. The latter is important in scenarios where the dataset maintainer snapshots a data collection from a specific source. In this case, the attacker can poison the original data source just before the snapshotting takes place. Wikipedia's articles and database are examples of this, as these are periodically updated (Carlini, et al., 2023).

Second, **collected training data could be sanitised**, for example, using techniques for analysing the input data distribution. One such technique, causal unlearning, an anti-poisoning method, aims to detect polluted data in the dataset. By removing different sets of samples from the dataset and observing if a misclassification still occurs, the cause of it can be detected (Cao, et al., 2018). Third, the attack surface of the model could be limited by **discarding some of the features**, which would constrain the space for malicious perturbations. Finally, one can attempt to create a more robust model by introducing certain **regularisation methods.** Apart from preventing overfitting and granting better generalisability, they can lead to better robustness of the models facing adversarial examples (Yoshida & Miyato, 2017; Cisse, Bojanowski, Grave, Dauphin, & Usunier, 2017; Hoffman, Roberts, & Yaida, 2019). One of the most promising defences that utilises such a method is termed a Parseval network, in which the global Lipschitz constant is constrained during training.

While this method is described as promising, it is computationally expensive, which increases the difficulty of implementation (Cisse, Bojanowski, Grave, Dauphin, & Usunier, 2017). Still, when comparing the necessary computational load, the novelty, and efficacy, this method appears to be the most promising out of the three defences in the regularisation section.

Table 1: Defence method sources for the preparation phase.

Defence method	Reference
Randomisation	(Joseph, Laskov, Roli, Tygar, & Nelson, 2013)
Causal unlearning	(Cao, et al., 2018)
Feature selection	(Zhang, Chan, Biggio, Yeung, & Roli, 2015)
Parseval networks	(Cisse, Bojanowski, Grave, Dauphin, & Usunier, 2017)
Spectral norm regularisation	(Yoshida & Miyato, 2017)
Jacobian regularisation	(Hoffman, Roberts, & Yaida, 2019)

Training phase

The training phase involves three activities: the model is built, trained, and evaluated upon test time. It can be seen as an iterative sequence of these actions. After evaluating the model's performance, one might want to make changes to the model's definition – a part of source code related to the model – and effectively repeat the process until the expectations for its performance are met. In a nutshell, training is an optimisation process of finding model parameter values that allows for solving a given task with a desired efficacy. The model can find the best fit to the training data by controlling the change to the objective function, which allows for measuring how much error a model is producing for its results. Ideally, during each iteration of training, this value is minimised, and the model improves its performance (Fan, 2023). This process is of great importance for defenders, since an attacker might attempt to manipulate or deceive it, specifically in the case of poisoning and evasion attacks.

Defences deployed in the training phase can generally be divided into two categories: **data augmentation** techniques and **obfuscation** techniques. The most researched variation of the first category is adversarial training, in which adversarial examples are included in the training dataset so the model can learn to differentiate them from benign examples. In an ideal situation, where all (future) adversarial samples are known, this is a promising technique – we simply train the model to recognise which inputs are malicious. However, in reality, we cannot know the complete set of adversarial examples that the model currently is and will be subjected to. Therefore, the main challenge when using adversarial training is selecting the adversarial samples that are representative of the range of the adversarial samples that the model could encounter. Ideally, this process should be repeated periodically, including previously unknown adversarial samples. Another proposed method is Gaussian data augmentation, which adds noise to the inputs of a model. That noise consists of values which are drawn at random from a Gaussian distribution (Rochac, Liang, Zhang, & Oladunni, 2019). By assumption, such noisy inputs enhance the capabilities of a model to be more robust to AML attacks. The benefit of data augmentation techniques is that, in addition to showing promising results, they also allow for relatively easy security improvements to the running model, since they can be applied during model retraining.

The defences present in the second category rely on gradient obfuscation (Athalye, Carlini, & Wagner, 2018). Since white-box attacks rely on the gradient of the model, the efficacy of these attacks normally decreases when obfuscation methods are used. However, novel attack techniques have been developed that render this defence mechanism ineffective (Athalye, Carlini, & Wagner, 2018).

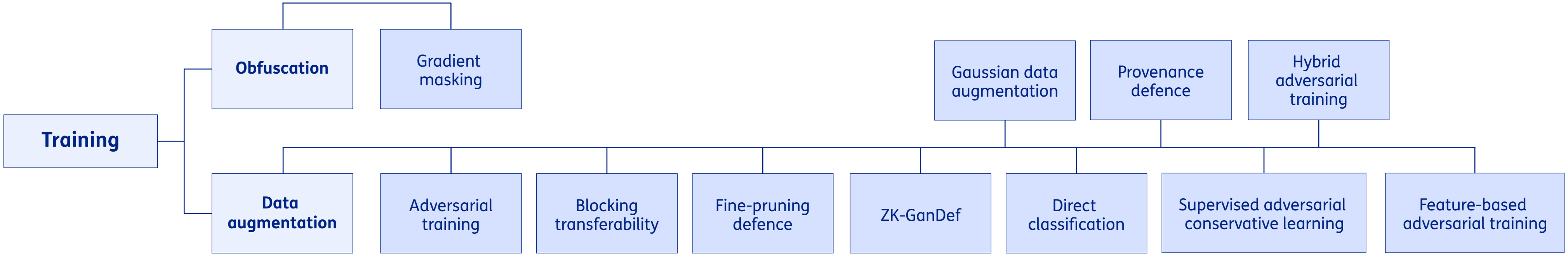

Figure 4: AML defences in the training phase.

Hence, when the attack is indeed white-box, gradient obfuscation is not deemed useful, as attackers can circumvent this defence method. Additionally, this defence method could also be circumvented if the attacker is able to develop a black-box substitute model to craft adversarial examples (Papernot, et al., 2017). Still, this defence method provides protection against certain attacks, more if defenders can prevent attackers from obtaining a substitute model, and the underlying techniques could provide for novel defence mechanisms.

Table 2: Defence method sources for the training phase.

Defence method	Reference
Adversarial training	(Goodfellow, Shlens, & Szegedy, 2015)
Blocking transferability	(Hosseini, Chen, Kannan, Zhang, & Poovendran, 2017)
Fine-pruning defence	(Liu, Dolan-Gavitt, & Garg, 2018)
ZK-GanDef	(Liu, Khalil, & Khreishah, 2019)
Gaussian data augmentation	(Rochac, Liang, Zhang, & Oladunni, 2019)
Direct classification	(Grosse, Manoharan, Papernot, Backes, & McDaniel, 2017)
Provenance defence	(Baracaldo, Chen, Ludwig, & Safavi, 2017)
Supervised adversarial contrastive learning	(Li, et al., 2023)
Hybrid adversarial training	(Ryu & Choi, 2022)
Feature-based adversarial training	(Ryu & Choi, 2022)
Gradient masking	(Tramèr, et al., 2018)

Operational phase

Once the model is developed and evaluated successfully, it can be released to the production environment. Regarding the model's deployment, one should consider the model itself, as well as its embeddings. This includes the way the model will be interacted with and how it will operate.

Defences in the operational phase generally consist of **introducing an add-on** to the target model or **introducing additional models** to detect or block AML attacks. An add-on to the model could mitigate attacks by monitoring the behaviour of the model, for example, by focusing on monitoring changes in the value of the loss function. This defence method, called loss-based defence, monitors deviations from the expected values, which are marked as suspicious, triggering the termination of the model's operations (Chen, Zou, Su, & Zhang, 2020). Additionally, one may add additional models to counter attacks, creating a sort of layered defence strategy. For example, to prevent model stealing attacks, ensemble defence methods can be implemented, in which the output is determined through a voting scheme between different models. Within this scheme, the output that the models collectively determine to be the best is returned as final output by the model (Chen, Zou, Su, & Zhang, 2020). For an adversary to succeed in attacking this model, they would have to devise an attack that can account for all the different models in the ensemble.

It is important to consider the impact of defensive techniques on the runtime of the system however. Therefore, these methods must be efficient enough not to overburden the existing infrastructure.

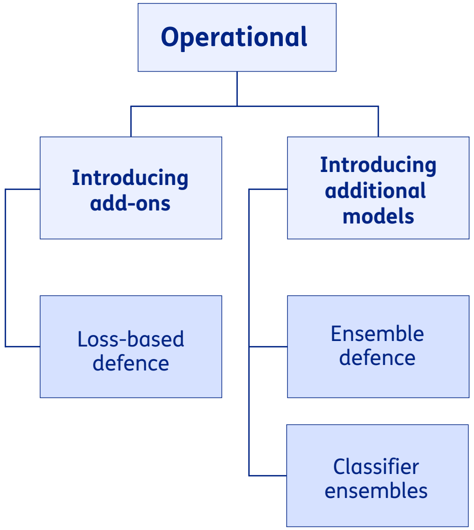

Figure 5: AML defences in the operational phase.

Table 3: Defence method sources for the operational phase.

Defence method	Reference
Loss-based defence	(Yang, Wu, Li, & Chen, 2017)
Ensemble defence	(Hitaj & Mancini, 2018)
Classifier ensembles	(Biggio, Fumera, & Roli, 2010)

Input phase

This phase includes mechanisms for submitting and sanitising the input from the user. In the case of a server-side application, this can include an upload functionality with a pre-processing pipeline implemented in the back end.

Defences in the input phase are largely similar to those in the preparation phase, as this phase mainly consists of processes related to both data acquisition and data curation.

A promising technique implemented in computer vision tasks is termed feature squeezing. This technique allows for reducing the colour depth and smoothing out differences between pixels. This effectively shrinks the size of the space in which the attacker might introduce malicious perturbations. To increase the chances of an effective attack, the adversary needs to elevate the intensity of perturbations, generating more visible malicious noise in the altered image. Overall, the transformations proposed in this technique account for the fact that the model is not robust. Some of the tested attacks failed or their effect was reduced significantly (Xu, Evans, & Qi, 2017). Besides the computer vision domain, this technique can be introduced to other tasks such as automatic speech recognition systems, which may utilise spectrograms for the audio data representation (OpenAI, 2023).

The input phase is also where adversarial examples may be detected. One manner of achieving this is by comparing the prediction of a deep neural network based on the original input with the one based on the squeezed input (Xu, Evans, & Qi, 2017).

Another technique, that both detects and mitigates adversarial examples, relies on an image's compression levels. The fact that these compression levels are randomly applied to different regions of an image allows one to rectify the perturbed input (Liu, et al., 2018).

Clearly, the attempts to suppress the malicious effect of perturbations at the input phase rely on certain forms of transformations, which leads to a reduction of the impact of malicious noise. Spatial and magnitude alterations, as well as compression techniques, seem promising. However, at the same time, they re-form the original inputs, which raises questions on preserving some of the significant input attributes. In this context, the defensive methods applied in the input phase must be validated in terms of information loss, in addition to their efficacy. Some of the operations might result in a decrease of data quality, its meaningfulness, or intelligibility. Moreover, any form of decomposition can result in a less interpretable process of classification. From an explainable AI perspective of research, this might be problematic, as formulating conclusions on how the model operates and how it produces a specific output might become impossible given such an obfuscated form of input data.

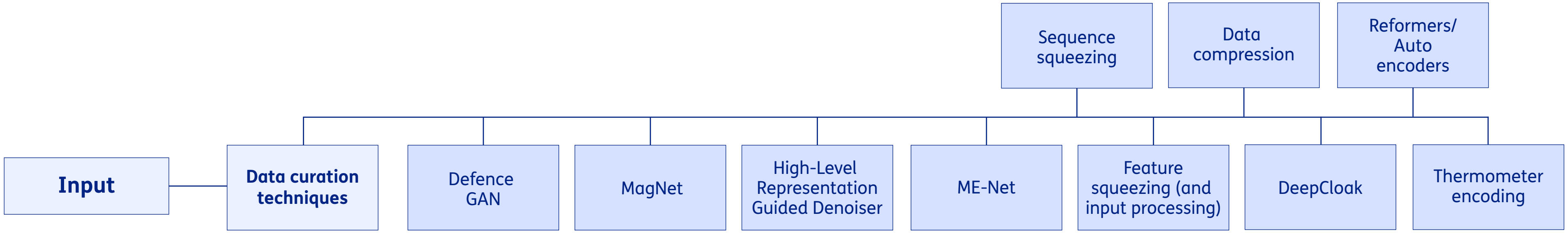

Figure 6: AML defences in the input phase.

Table 4: Defence method sources for the input phase.

Defence method	Reference
Defence GAN	(Samangouei, Kabkab, & Rama, 2018)
MagNet	(Meng & Chen, 2017)
High-Level Representation Guided Denoiser	(Liao, et al., 2018)
ME-Net	(Yang, Zhang, Katabi, & Xu, 2019)
Sequence squeezing	(Rosenberg I. , Shabtai, Elovici, & Rokach, 2019)
Feature squeezing (and input processing)	(Xu, Evans, & Qi, 2018)
Data compression	(Dziugaite, Ghahramani, & Roy, 2016)
DeepCloak	(Gao, Wang, Lin, Xu, & Qi, 2017)
Reformers/Autoencoders	(Liu, Xie, & Srivastava, 2017)
Thermometer encoding	(Buckman, Roy, Raffel, & Goodfellow, 2018)

Output phase

This phase includes means for presenting the results of the model's task. Based on the quality of the output, the developer might decide that additional actions must be performed to improve the performance of the model, a specific module, or the entire application.

Defences in this phase generally aim to ensure that only necessary information is relayed to users, **omitting non-essential data** that might be useful to attackers. Thus, the two threats in the output phase are model stealing attacks, as constructing a replica model relies on the data produced in this very phase, and membership inference attacks, as attacks extract the data in this stage. To mitigate such attacks, we can attempt to hide certain information from the potential attacker by reducing or eliminating any feedback given by the ML model or providing less meaningful outputs (Clark Jr & Doran, 2018). This can, for example, be done by providing labels instead of classes' probabilities. Since users generally do not need this information, these strategies could limit an attacker's access to important data without significantly compromising the model's usability for other users.

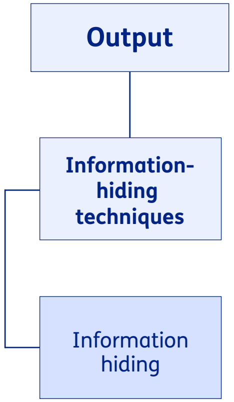

Figure 7: AML defences in the output phase.

Table 5: Defence method sources for the output phase.

Defence method	Reference
Information hiding	(Barreno, Nelson, Sears, Joseph, & Tygar, 2006)

Improvement phase

The improvement phase aims to adapt the model to allow it to handle previously unseen features. ML models require updates to their parameters, which are achieved by fine-tuning the model with newly collected data. The focus can be on fine-tuning the last layers of the model, which is also a way to achieve transfer learning (Fan, 2023). The improvement might also focus on the application's performance by optimising the workload of the modules (e.g., batching), the scalability of operators, finding a more suitable framework, or by changing the architecture or hyperparameters (Kogan, 2023).

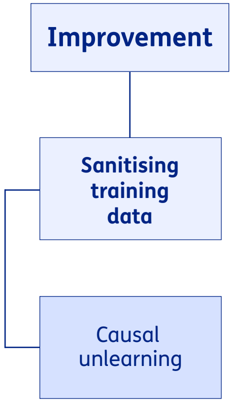

Figure 8: AML defences in the improvement phase.

As this phase focuses on improving the model with new training data, defences used in the preparation phase to **sanitise the training data** could also be applied here. One could again check for poisoned data, for example using causal unlearning (Cao, et al., 2018).

Regardless of the specifics of the end-product, it is important to consider that the software accompanying the model may be vulnerable to some server-side or client-side attacks. This can leave the users or the application vulnerable to further exploitation by the attackers. Despite the focus on adversarial machine learning, we emphasise that the security of the end-product must be evaluated both from the perspective of the security of the model, as well as the security of all the incorporated components.

Table 6: Defence method sources for the improvement phase.

Defence method	Reference
Causal unlearning	(Cao, et al., 2018)

Overarching trends

Utilising the extensive insights gained through our research into the defences, we can distinguish five overarching trends. These trends highlight the connections between the defences in the different categories and provide general lessons for implementing defences.

Emphasis on evasion attacks

Firstly, while there are five types of AML attacks in the domain of predictive AI, most of the defences are geared towards evasion attacks. This may be because these attacks have received the most attention in research, potentially leading to evasion attacks being the most likely attack type one could encounter. This may also stem from the fact that AML attacks, in the literature, are often confined solely to evasion attacks (Li, Fung, & Charland, 2022; AL-Essa, Andresini, Appice, & Malerba, 2022; Liu, Khalil, & Khreishah, 2019). As most literature focuses on evasion attacks, papers developing defence methods for novel or significant attacks primarily focus on evasion attacks as well.

Prominence of certain techniques

Secondly, even though there is a sizeable list of defences, some of them gained more attention than others. One of them is adversarial training, a method that involves adding attack samples to the training data. As the model is then trained to recognise these malicious samples, it can better recognise and deal with them upon deployment. This method's popularity stems from its ability to drastically improve a model's ability to withstand AML attacks and from it being relatively straightforward to implement even after the model is deployed. This has spurred other researchers to explore ways to improve adversarial training. Examples of this include eliminating the need for real malicious samples, and devising ways to determine which malicious samples should be included for optimal performance. (Liu, Khalil, & Khreishah, 2019). However, the effectiveness of this defence technique is limited by the fact that selecting an optimal set of samples remains a challenge.

Other techniques are also featured prominently in multiple defence methods, such as the usage of autoencoders. These neural networks take an input, encode a compressed version of the input, and output the reconstructed input from the code (Dartat, 2017). They achieve this by first learning the manifold of benign data, so when an evasion attack targets the boundary of a benign example, the autoencoder reforms the input and pushes it to the correct benign sample (Meng & Chen, 2017). This demonstrates that certain promising techniques could be implemented in various ways.

Incorporation of detection methods

Thirdly, a sizeable number of defences incorporate methods for detecting attacks into the defence method, resulting in a mechanism that can both detect and mitigate an attack. An example of such a defence method is blocking transferability (Hosseini, Chen, Kannan, Zhang, & Poovendran, 2017). This method proposes a solution to the problem of the transferability of attacks, meaning that if an attacker develops an attack using one model, those attacks are likely to work against another model, even if they differ significantly (Hosseini, Chen, Kannan, Zhang, & Poovendran, 2017). The blocking transferability method builds on this concept, using benign input to train the model to learn how benign data is distributed, similar to how it knows how training data is distributed to create its classes (Hosseini, Chen, Kannan, Zhang, & Poovendran, 2017). Then, the model is trained to discard all inputs whose distribution differs from benign inputs, discarding the adversarial inputs.

Focus on security-enhancing systems in the cyber domain

Fourthly, there is an important lesson regarding defences from the cyber domain that demonstrates the importance of considering AML defences. Namely, the ML applications that were aimed at improving security-enhancing systems in the cyber domain introduced new vulnerabilities, as their potential susceptibility to AML attacks could lead to a compromise of the entire system. This is specifically the case for security-enhancing systems that focus on anomaly detection, such as Network Intrusion Detection Systems (NIDS) (Mbow, Sakurai, & Koide, 2022). While these systems functioned sufficiently, the advances in ML have made them even more accurate in the face of rapidly evolving and increasing amounts of attacks (Jmila & Ibn Khedher, 2022). However, ML models' vulnerability to AML attacks poses new risks which could render the NIDS as a whole vulnerable. Using maliciously perturbed samples, attacks could bypass the NIDS and, for example, gain access to the system the NIDS was supposed to protect (Alotaibi & Rassam, 2023). Given that ML models are increasingly being adopted in systems such as autonomous vehicles and chemical plants (Gu & Easwaran, 2019), this new attack vector could cause significant damage if exploited. Thus, simply adding ML to an existing system may not mitigate the current security problems and could even create new attack vectors.

Difficulties of evaluating defences

Lastly, while defences are being developed against all AML attack categories, using (combinations of) novel techniques, some defences may also seem more promising on paper than they are in the real world. This may stem from the fact that defences are often evaluated incorrectly, leading to incorrect assumptions about their efficacy (Carlini, et al., 2019). For example, a promising technique by Papernot, McDaniel, Wu, Jha, and Swami (2016), termed defensive distillation, was proven to be not as robust as the authors claimed (Carlini & Wagner, 2017). In this case, the constrained set of tests led to an overly optimistic evaluation of the network. Thus, one should inspect whether defences can achieve their claimed efficacy before implementing them, and periodically review their efficacy in a changing threat landscape. Defences may not always mitigate all (sub)types of attacks. However, as the highest level of security is not always necessary, this should not be a reason to automatically disqualify a particular defence mechanism. Since perfect defences are impossible, the goal should be to increase the cost of attacking to the level necessary to deter the adversary. This requires comprehensive modelling of the expected threat. Generally, scholars model the threat according to three axes: goals (what outcome does the adversary seek), capabilities (what constraints do they face), and knowledge (what do they know about the model for example) (Biggio & Roli, 2018; Duddu, 2018). Additionally, there is a fourth axis which is often only mentioned implicitly: strategy. This axis explores whether the attacker will observe and gather information (passive attacks) or actively target the model, disrupting its functioning (active attacks) (Dasgupta & Collins, 2019; Dai, Sthapit, Epiphaniou, & Maple, 2021; Rosenberg, Shabtai, Elovici, & Rokach, 2021). To reduce implicit assumptions clouding the threat modelling and to conduce the debate about the threat, all four axes should be considered when modelling the adversary. Based on the outcome of this threat modelling, the defenders select the defences that could increase the cost of attacking to the level necessary to deter the adversary.

Conclusions

AML research into the cyber domain continues to progress rapidly, as evidenced by the rapid increase of the number of papers published, including a 40% increase from 2022 to 2023. While these developments demonstrate the increasing importance of this research field and the new findings continue to enable new methods, digesting all this information has become a task which cannot be performed solely by humans anymore.

Additionally, this increase may also indicate that AML attacks are becoming more realistic. At the same time, however, the development of defences continues to significantly lag behind the developments occurring on the attack side (Carlini, et al., 2019). And even if feasible defences exist against expected attacks, selecting the appropriate one(s) remains difficult. Therefore, one should first seek to analyse the threats the model is expected to encounter, which will help inform the selection of applicable defences.

Despite being less developed than the attack methods literature, research on defence methods continues to progress rapidly. As this whitepaper has shown, the academic community presents defences that are stated to increase the robustness of the model in all stages of the ML life cycle. By inventorying the existing defence methods and categorising them within the ML life cycle, this whitepaper has provided a clear and structured overview of methods that can be used in an attempt to improve ML models' robustness against AML attacks. Depending on the stage of development (including when the model is finished), the presented framework outlines the defences that could be implemented.

As Brink et al. (2023) concluded, research stemming from other domains is largely generalisable to the cyber domain. This held true for the defences, as demonstrated by various authors who applied existing defence mechanisms to the cyber domain (Apruzzese, Andreolini, Colajanni, & Marchetti, 2020; Mbow, Sakurai, & Koide, 2022; AL-Essa, Andresini, Appice, & Malerba, 2022). Thus, future research should also explore the developments occurring in the other domains to inform research in the cyber domain.

Another challenge pertains to the testing of defence mechanisms. As this whitepaper has argued, there are various defences available, but they do not always perform equally well, depending on the testing method. While certain defences, such as defensive distillation, demonstrated promising results in the authors’ testing environment, the same defence mechanism failed when it was put up against attacks in a scenario that was closer to the real world (Carlini & Wagner, 2017). A valuable contribution in future research would be the development of an operationally relevant evaluation of the various defence methods against AML. This effort could also investigate whether defences are transferable, just as attacks are.

By providing a first structured overview of AML defences, this whitepaper contributes the structure and clarity necessary to grasp the developments occurring within this rapidly developing research field. The next steps in this path towards clarity should be verifying which defence methods live up to their claimed performance in the real world. This would bring us one step closer to securely deploying ML models.

Bibliography

Abhishek, R., Anshuman, C., Charles, A. K., & Prasant, M. (2019). A Moving Target Defense against Adversarial Machine Learning. Proceedings of Second ACM/IEEE Workshop on Security and Privacy in Edge Computing (EdgeS&P '19) (pp. 383-388). New York: ACM.

AL-Essa, M., Andresini, G., Appice, A., & Malerba, D. (2022). An XAI-based adversarial training approach for cyber-threat detection. 2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech) (pp. 1-8). Falerna: IEEE.

Alotaibi, A., & Rassam, M. A. (2023). Adversarial machine learning attacks against intrusion detection systems: A survey on strategies and defense. Future Internet 15(2), 1-34.

Anley, C. (2022, July 6). Whitepaper – Practical Attacks on Machine Learning Systems. Retrieved from NCC Group: https://research.nccgroup.com/2022/07/06/whitepaper-practical-attacks-on-machine-learning-systems/

Apruzzese, G., Andreolini, M., Colajanni, M., & Marchetti, M. (2020). Hardening random forest cyber detectors against adversarial attacks. IEEE Transactions on Emerging Topics in Computational Intelligence, 4(4), 427-439.

ASReview Lab. (2022). ASReview Lab, v0..19.

Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples.

Baracaldo, N., Chen, B., Ludwig, H., & Safavi, J. A. (2017). Mitigating poisoning attacks on machine learning models: A data provenance based approach. Proceedings of the 10th ACM workshop on artificial intelligence and security (pp. 103-11-). New York: Association for Computing Machinery.

Barreno, M., Nelson, B., Sears, R., Joseph, A. D., & Tygar, J. D. (2006). Can machine learning be secure? Proceedings of the 2006 ACM Symposium on Information, computer and communications security (pp. 16-25). Taipei: Association for Computing Machinery.

Biggio, B., & Roli, F. (2018, December). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84, pp. 317-331.

Biggio, B., Fumera, G., & Roli, F. (2010). Multiple classifier systems for robust classifier design in adversarial environments. International Journal of Machine Learning and Cybernetics, 1, 27-41.

Brink, N., Kamphuis, Y., Maas, Y., Gwen, J.-F., van Stijn, J., Poppink, B., ... Chiscop, I. (2023, February). Adversarial AI in the cyber domain. Retrieved from TNO: https://www.tno.nl/en/newsroom/2023/02/first-overview-cyberattack-techniques-ai/

Buckman, J., Roy, A., Raffel, C., & Goodfellow, I. (2018). Thermometer encoding: One hot way to resist adversarial examples. 6th International Conference on Learning Representations (ICLR 2018) (pp. 1-22). Vancouver: International Conference on Learning Representations.

Cao, Y., Fangxiao Yu, A., Aday, A., Stahl, E., Merwine, J., & Yang, J. (2018). Efficient Repair of Polluted Machine Learning Systems via Causal Unlearning., (pp. 735-747).

Carlini, N., & Wagner, D. (2017). Towards Evaluating the Robustness of Neural Networks. Retrieved from Arxiv: https://arxiv.org/abs/1608.04644

Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., ... Kurakin, A. (2019, February 18). On Evaluating Adversarial Robustness. Retrieved from Arxiv: https://arxiv.org/pdf/1902.06705.pdf

Carlini, N., Jagielski, M., Choquette-Choo, C. A., Paleka, D., Pearce, W., Anderson, H., ... Tramèr, F. (2023). Poisoning Web-Scale Training Datasets is Practical.

Chen, J., Zou, J., Su, M., & Zhang, L. (2020). A review of poisoning attack and defense of deep learning models. Journal of Cyber Security, 14-29.

Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., & Usunier, N. (2017). Parseval Networks: Improving Robustness to Adversarial Examples. Proceedings of the 34 th International Conference on Machine Learning, (pp. 1-10). Sydney.

Clark Jr, G. W., & Doran, M. V. (2018). Machine Learning Security Vulnerabilities in Cyber-Physical Systems. Proceedings of The 9th International Multi-Conference on Complexity, Informatics and Cybernetics (IMCIC 2018) (pp. 41-46). Orlando: International Institute of Informatics and Cybernetics.

Dai, G., Sthapit, S., Epiphaniou, G., & Maple, C. (2021). Artificial Intelligence Technologies in Building Resilient Machine Learning. Competitive Advantage in the Digital Economy (pp. 50-55). IET.

Dartat, A. (2017, October 3). Applied Deep Learning - Part 3: Autoencoders. Retrieved from Towards Data Science: https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798

Dasgupta, P., & Collins, J. (2019). A survey of game theoretic approaches for adversarial machine learning in cybersecurity tasks. AI Magazine, 31-43.

Duddu. (2018). A survey of adversarial machine learning in cyber warfare. Defence Science Journal, 68(4), 356.

Dziugaite, G. K., Ghahramani, Z., & Roy, D. M. (2016). A study of the effect of JPG compression on adversarial images. International Society for Bayesian Analysis (ISBA 2016) World Meeting (pp. 1-8). Sardinia: International Society for Bayesian Analysis.

ETSI. (2020). Securing Artificial Intelligence (SAI): Problem Statement. ETSI.

Fan, M. T. (2023). Machine Learning with Confidential Computing: A Systematization of Knowledge.

Gao, J., Wang, B., Lin, Z., Xu, W., & Qi, Y. (2017). Deepcloak: Masking deep neural network models for robustness against adversarial samples. 5th International Conference on Learning Representations (pp. 1-8). Toulon: International Conference on Learning Representations.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations (pp. 1-11). San Diego: International Conference on Learning Representations.

Grosse, K., Manoharan, P., Papernot, N., Backes, M., & McDaniel, P. (2017). On the (statistical) detection of adversarial examples. ArXiv, 1-13.

Gu, X., & Easwaran, A. (2019). Towards safe machine learning for cps: infer uncertainty from training data. Proceedings of the 10th ACM/IEEE International Conference on Cyber-Physical Systems, 249-258.

Hitaj, D., & Mancini, L. V. (2018). Have you stolen my model? evasion attacks against deep neural network watermarking techniques. ArXiv, 1-7.

Hoffman, J., Roberts, D. A., & Yaida, S. (2019). Robust Learning with Jacobian Regularization.

Hosseini, H., Chen, Y., Kannan, S., Zhang, B., & Poovendran, R. (2017). Blocking transferability of adversarial examples in black-box learning systems. Retrieved from arXiv: https://arxiv.org/pdf/1703.04318.pdf

Jmila, H., & Ibn Khedher, M. (2022). Adversarial machine learning for network intrusion detection: A comparative study. Computer Networks 214, 1-14.

Joseph, A. D., Laskov, P., Roli, F., Tygar, J. D., & Nelson, B. (2013). Machine learning methods for computer security (Dagstuhl Perspectives Workshop 12371). Dagstuhl Reports (pp. 109-130). Wadern: Dagstuhl Research Online Publication Server.

Kogan, A. (2023). Improving Inference Performance of Machine Learning with the Divide-and-Conquer Principle.

Li, M. Q., Fung, B. C., & Charland, P. (2022). DyAdvDefender: An Instance-based Online Machine Learning Model for Perturbation-trial-based Black-box Adversarial Defense. Information Sciences, 357-373.

Li, W., Zhao, B., An, Y., Shangguan, C., Ji, M., & Yuan, A. (2023). Supervised contrastive learning for robust text adversarial training. Neural Computing and Applications, 35(10), 7357-7368.

Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., & Zhu, J. (2018). Defense against adversarial attacks using high-level representation guided denoiser. Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1778-1787). Sal Lake City: IEEE.

Liu, G., Khalil, I., & Khreishah, A. (2019). ZK-GanDef: A GAN Based Zero Knowledge Adversarial Training Defense for Neural Networks. 2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN) (pp. 64-75). Portland: IEEEE.

Liu, K., Dolan-Gavitt, B., & Garg, S. (2018). Fine-pruning: Defending against backdooring attacks on deep neural networks. International symposium on research in attacks, intrusions, and defenses (pp. 273-294). Heraklion: Springer Cham.

Liu, Y., Xie, Y., & Srivastava, A. (2017). Neural Trojans. 2017 IEEE International Conference on Computer Design (ICCD) (pp. 45-48). Boston: IEEE.

Liu, Z., Liu, Q., Liu, T., Xu, N., Lin, X., Wang, Y., & Wen, W. (2018). Feature Distillation: DNN-Oriented JPEG Compression Against Adversarial Examples.

Mbow, M., Sakurai, K., & Koide, H. (2022). Advances in Adversarial Attacks and Defenses in Intrusion Detection System: A Survey. International Conference on Science of Cyber Security, 196-212.

Meng, D., & Chen, H. (2017). MagNet: A Two-Pronged Defense against Adversarial Examples. CCS '17: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 135-147). Dallas: Association for Computing Machinery.

OpenAI. (2023, November 16). Introducing Whisper. Retrieved from https://openai.com/research/whisper

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). Practical Black-Box Attacks against Machine Learning. Proceedings of the 2017 ACM on Asia conference on computer and communications security, 506-519.

Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016). Distillation as a defense to adversarial perturbations against deep neural networks. Security and Privacy (SP) IEEE Symposium, 582-597.

Rochac, J. F., Liang, L., Zhang, N., & Oladunni, T. (2019). A Gaussian Data Augmentation Technique on Highly Dimensional, Limited Labeled Data for Multiclass Classification Using Deep Learning. 2019 Tenth International Conference on Intelligent Control and Information Processing (ICICIP), (pp. 145-151). Marrakesh.

Rosenberg, I., Shabtai, A., Elovici, Y., & Rokach, L. (2019). Defense methods against adversarial examples for recurrent neural networks. ArXiv, 1-20.

Rosenberg, I., Shabtai, A., Elovici, Y., & Rokach, L. (2021). Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain. ACM Computing Surveys (CSUR), 54(5), 1-36. doi:https://doi.org/10.1145/3453158

Ryu, G., & Choi, D. (2022). A hybrid adversarial training for deep learning model and denoising network resistant to adversarial examples. Applied Intelligence, 53(8), 9174-9187.

Ryu, G., & Choi, D. (2022). Feature-based adversarial training for deep learning models resistant to transferable adversarial examples. IEICE TRANSACTIONS on Information and Systems, 105(5), 1039-1049.

Samangouei, P., Kabkab, M., & Rama, C. (2018). Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models. International Conference on Learning Representations 2018 (pp. 1-17). Vancouver: International Conference on Learning Representations.

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I. J., Boneh, D., & McDaniel, P. (2018). Ensemble adversarial training: Attacks and defenses. The Sixth International Conference on Learning Representations (pp. 1-22). Vancouver: International Conference on Learning Representations.

Vassilev, A., Oprea, A., Fordyce, A., & Anderson, H. (2024). Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. NIST.

Xu, W., Evans, D., & Qi, Y. (2017). Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. Network and Distributed Systems Security Symposium (NDSS). San Diego.

Xu, W., Evans, D., & Qi, Y. (2018). Feature squeezing: Detecting adversarial examples in deep neural networks. Network and Distributed Systems Security Symposium (NDSS) (pp. 1-15). San Diego: Network and Distributed Systems Security .

Yang, C., Wu, Q., Li, H., & Chen, Y. (2017). Generative poisoning attack method against neural networks. ArXiv, 1-8.

Yang, Y., Zhang, G., Katabi, D., & Xu, Z. (2019). Me-net: Towards effective adversarial robustness with matrix estimation. Proceedings of the 36th International Conference on Machine Learning (pp. 7025-7034). Long Beach: Proceedings of Machine Learning Research.

Yoshida, Y., & Miyato, T. (2017). Spectral Norm Regularization for Improving the Generalizability of Deep Learning. ArXiv, 1-12.

Zhang, F., Chan, P. P., Biggio, B., Yeung, D. S., & Roli, F. (2015). Adversarial feature selection against evasion attacks. IEEE transactions on cybernetics, 46(3), 766-777.

Authors
Niels Brink, Jip van Stijn, Piotr Stachyra,
Olivier Spinnler, Yori Kamphuis

Contact
Yori Kamphuis

Lead Counter AI
TNO: Resilience & Security – Unit Defence,
Safety and Security

☎ +31 611856745

https://www.linkedin.com/in/yorikamphuis

TNO is The Netherlands' leading not-for-profit institute for applied scientific research. Its mission is to create impactful innovations for the sustainable wellbeing and prosperity of society. The unit Defence, Safety and Security aims to protect what we hold dear and ensure that people can live together in freedom and security. And to that end, we develop strategic knowledge, technology, and capabilities.

TNO innovation for life

tno.nl