doi:10.20944/preprints202210.0157.v1

ADVERSARIAL ARTIFICIAL INTELLIGENCE IN INSURANCE: FROM AN EXAMPLE TO SOME POTENTIAL REMEDIES

Behnaz Amerirad, McGill University, Canada

Matteo Cattaneo, Innovation Lab, Reale Mutua Group[1]

Ron S. Kenett, KPA Group, Samuel Neaman Institute, Technion, Israel and University of Torino

Elisa Luciano, University of Torino, Collegio Carlo Alberto[2]

Abstract

Artificial intelligence (AI) is a tool that financial intermediaries and insurance companies use in most cases or are willing to use it in almost all their activities. AI can have a positive impact on almost all aspects of the insurance value chain.: pricing, underwriting, marketing, claims management, after-sales services. While it is very important and useful, AI is not free of risks, including its robustness against cyber-attacks and so-called adversarial attacks. Adversarial attacks are conducted by external entities to misguide and defraud the AI algorithms. The paper is designed to provide a review of adversarial AI and discuss its implications for the insurance sector.

The study starts with a taxonomy of adversarial attacks and presents a fully-fledged example of claims falsification in health insurance. Some remedies, consistent with the current regulatory framework, are presented.

1. Introduction

AI is a set of techniques and technologies that the financial industry considers as a strategic priority and potentially as a source of relevant competitive advantages and in which it invests a significant efforts and resources. This paper focuses on insurers who apply AI in their processes, exploiting the large databases they already have or the data they can collect from customers through, for example, web-based interaction

[1] *The opinions expressed here are those of the Author and do not represent his institution view.*

[2] *Corresponding Author: elisa.luciano@unito.it*

and wearable devices. Indeed, AI is more and more used in product design and development, pricing and underwriting, marketing and distribution, customer service and relationship with the clients, claims management, from claims filing to settlement.

While it is very important and useful, AI is not free of risks. EIOPA itself highlights the relevance of AI "robustness". AI systems should be robust not only in the sense of being fit-for-purpose, regularly maintained and subject to tests, but also in being deployed in infrastructures protected from cyberattacks. Within the notion of cyberattacks one can confidently consider adversarial AI attacks aimed at defrauding an AI system, a way that the result is often detected by the human eye, but not by the system.

To prevent and fix adversarial attacks, an underwriter must first categorize such attacks according to several dimensions, i.e., have a taxonomy. After providing a taxonomy in Section 1, the paper provides an example built on public data from health insurance. In Section 2, it illustrates the subtlety and powerfulness of those attacks. We conclude with practical remedies against adversarial attacks, in Section 3.

2. Adversarial Attacks and Their Taxonomy

An adversarial example is a sample of input data that has been very slightly altered in a way that is intended to mislead a ML system (Kurakin, Goodfellow and Bengio, 2017a). The result is that the AI application makes incorrect predictions.

Although AI applications on images, videos, text, or voice, are becoming increasingly sophisticated, they still are vulnerable to adversarial attacks based on specific perturbations of their input data. Sometimes these perturbations can be small, imperceptible to human detection. Under this context, not only Machine Learning (ML) systems are fooled for their detection, but also higher level of these perturbations can increase the success rate of the attack by lowering the accuracy of the system. A famous adversarial example is the image of a panda, provided by (Goodfellow, Shlens and Szegedy, 2015). In this example, the author explained that how small, invisible perturbations on the input pixels of a panda image result in its misclassification as a gibbon. The Appendix clarifies how adversarial attacks are generated based on different algorithms.

Adversarial attacks can be categorized based on its goal, properties, or capabilities as shown in Figure 1.

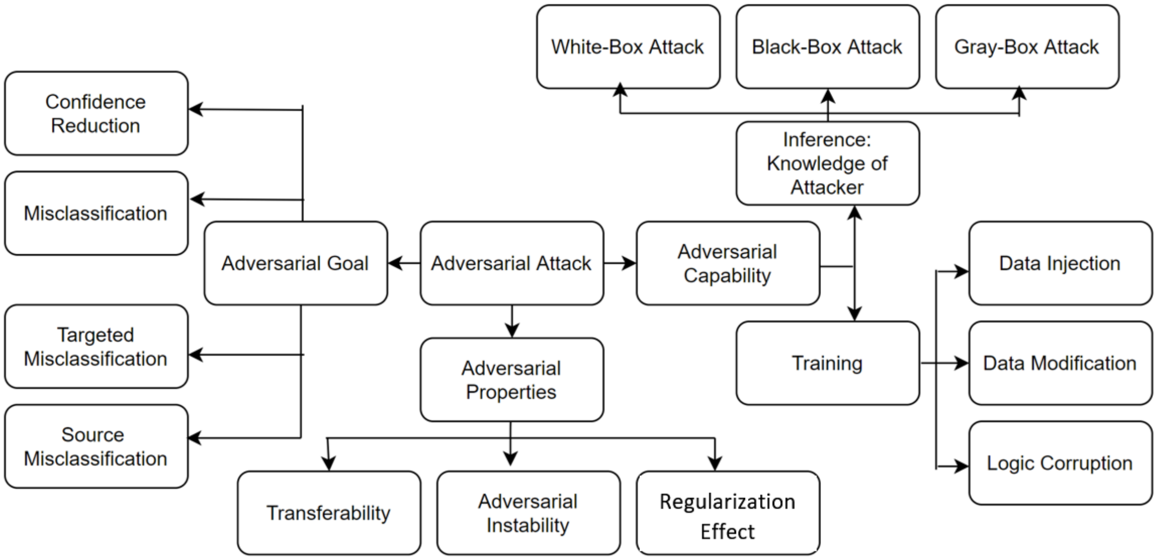

Figure 1 Adversarial Attack Taxonomy

2.1 Adversarial goal

Based on their goal, adversarial attacks can be divided into four categories: i) those that lead to confidence reduction, ii) to an untargeted misclassification, iii) to a target misclassification and iv) those that target a specific source for misclassification (see (Qiu, Liu, Zhou and Wu, 2019).

- When the aim is confidence reduction, the attackers attempt to reduce the accuracy of the target model prediction, i.e., the attack results in the model having very low accuracy.[3]
- When the aim is to obtain an untargeted misclassification, the attacker tries to change the original class of the input to any class that differs from the original one.
- When the aim is to obtain a targeted misclassification, the adversaries attempt to change the output to a specific target class.
- Finally, in source/target misclassification, the adversaries try to change the output classification of a particular input.

2.2 Adversarial Capability

While attacks on the training phase seek to learn, influence, or corrupt the model itself, attacks in the inference phase do not tamper with the targeted model but rather either produce adversary selected outputs or gather evidence about the model characteristics. (Ren, Zheng, Qin and Liud, 2020).

[3] Accuracy in AI is defined as the ratio of true positives and true negatives to all positive and negative outcomes. It measures how frequently the model gives a correct prediction out of the total predictions it made.

Attacks in the training phase: The attack strategies used in the training phase can be divided into three categories:

- **Data Injection:** The attacker has no access to training data or learning algorithms, but can add new data to the training data set in order to falsify the target model.
- **Data Modification:** Without access to the learning algorithm, but to all training data, the attacker can poison the target model by manipulating the training data.
- **Logic Corruption:** The attacker has access to interfere with the target model's learning algorithms.

Attacks in the inference phase: There are three common threat models in the inference phase for adversarial attacks: the white-box, gray-box, and black-box models (Ren et al., 2020). The effectiveness of such attacks is largely determined by the information available to the attacker about the model and its use in the target environment.

- **White-Box Attack:** In white-box attacks, the attackers know the details of the target model, including the model architecture, model parameters and training data. The attackers use the available information to identify the most vulnerable areas of the target model, and then use adversarial pattern generation methods to create inputs that exploit these vulnerabilities. (Qiu et al., 2019).
- **Black-Box Attack:** In the black box model, attackers do not know the structure of target networks and parameters but exploit system vulnerabilities using information about the environment or past inputs. Black box attacks can always compromise a naturally trained non-defensive system.
- **Gray-Box Attack:** In the Gray box model, an attacker is assumed to know the architecture of the target model but does not have access to the model parameters. In this threat model, it is assumed that the attacker creates adversarial examples at a surrogate classifier of the same architecture. Due to the additional structural information, a gray-box attacker always shows better attack performance compared to a black-box attacker.

We give, in Appendix 1, more details on the further split of White and Black-box attacks models, together their intuition.

2.3 Adversarial Properties

Adversarial attacks have three basic properties,: i) transferability, ii) adversarial instability and iii) the possibility of reaching regularization effect (Zhang and Li, 2018).

- **Transferability:** Transferable adversarial examples are not limited to attacking specific model architectures but can be generated by one model and tend to deceive other models with the same probability.
- **Adversarial Instability:** After physical transformations of the data, such as translation, rotation, and illumination of images for image-based attacks, the ability of the latter may be lost. In such a case, the AI model correctly classifies the data, and the adversarial attack is said to be unstable.
- **Regularization Effect:** Consists in training the AI so as to reveal its defects – especially neural network systems – and consequently to improve its resilience. Adversarial training (Goodfellow et al., 2015), that we discuss in the last Section, is an example of regularization method.

3. A Health Insurance Example and the Assessment of Damages

The Adversarial attack example we provide below consists of falsifying the health status of potential customers, by corrupting the breast images of female patients, who are not affected by malign cancer but will end up being classified as such by the AI system. Regardless of the high personal costs and implications, an occurrence of this type may be extremely costly for an insurance company in countries like the United States, where the health insurance system is almost completely private, and household massively rely on insurance coverage. The US spent approximately 18% of GDP on healthcare already in 2016, well before the COVID pandemic. This amount runs the risk of being inflated by fraud. Although there are no recent estimates of fraud, we know that it was already estimated this number to be $272 billion in 2011, as (Finlayson, Chung, Kohane and Beam, 2019). report. Unfortunately, fraud can be committed by a diverse set of individuals, including professionals in healthcare together with their patients.

Fraud in health insurance may occur both in underwriting phase and in the claim filing one. In the first case it may occur when the applicants make false, misleading, or at least incomplete information about their medical history or current health in order to deceive the system and gain more benefits. As part of the underwriting process, insurers determine the price of coverage by assessing the risks based (also) on the applicant's medical history. For this evaluation, insurers are allowed to ask questions about applicants' pre-existing conditions and then decide whom to offer coverage to, whom to deny coverage to, and whether to charge additional fees for individually purchased coverage.

Even in the claim filing phase, adversarial attacks occur when an attacker with access to medical imaging material can alter the content to make a misdiagnosis. Specifically, attackers can add or remove evidence of certain medical conditions from 3D medical scans, including: copying content from one image to another (image splicing), duplicating content within the same image to cover or add something (copy-move), and enhancing an image to give it a different appearance (image retouching), as in (Singh, Kumar, Singh and

Mohan, 2017) (Sadeghi, Dadkhah, Jalab, Mazzola and Uliyan, 2018). For example, the attacks may consist in injecting and removing pixels on CT scans of patients' lung cancer (Mirsky, Mahler, Shelef and Elovici, 2019).

Finlayson et al. discuss how pervasive is fraud in healthcare and the need for intelligent algorithms to diagnose the condition of insurance claimants for reimbursement (Finlayson et al., 2019). Their model was first developed for the classification of diabetic retinal disease using fundoscopic images, pneumothorax using chest X-rays, and melanoma using dermatoscopic images. Subsequently, the robustness of the model has been tested using both PGD and universal attacks that are imperceptible by humans (see the definition in the Appendix).

Another study (Wetstein, Gonz´alez-Gonzalo, Bortsova, Liefers, Dubost, Katramados, Veta 2020) evaluated several unexplored factors, including the degree of perturbation and the transmissibility of the adversarial attack, affecting the susceptibility of DL, in Medical Image Analysis systems (MedIA) mainly focused on diabetic retinopathy detection, ChestX-Ray for thoracic diseases, and histopathological images of lymph node sections.

(Hirano, Minagi and Takemoto, 2021) used universal attack in clinical diagnosis for classification of skin cancer, diabetic retinopathy and pneumonia. Their results confirmed that DNNs are susceptible universal attacks, resulting in an input being assigned to an incorrect class, and cause the DNN to classify an input into a specific class. Our example is built on public data and proceeds as follows.

3.1 Dataset

We explore the possibility of adversarial attacks on insurance claimants' information on breast abnormalities of mammograms. Our data source (Suckling, 1996) is the mini database MIAS, which consists of 323 mammogram images, each with a size of 1024x1024 pixels. In the database MIAS, the mammogram images are divided into three classes: glandular dense, fatty, and fatty glandular. Each class is subdivided into images of normal, benign, and malignant tissue.

Each abnormal image, either benign or malignant, has a type such as calcification, mass, and asymmetry. A total of 207 normal images and 116 abnormal (64 benign and 52 malignant) images were obtained. In this study, only the abnormal images from the dataset are used to classify the "benign" and "malignant" classes.

Our construction of the attack has been conducted using Python and consisted of three main stages: preprocessing, training, and adversarial attack on the abnormal images of breast cancer.

First stage

A common step in computer-aided diagnosis systems is preprocessing, which improves the characteristics of the image by applying a series of transformations to improve performance (Li, Ge, Zhao, Guan and Yan, 2018). An applied approach in this research is data augmentation, often used in the context of Deep Learning (DL), which refers to the process of generating new samples from existing data, used to improve data sparsity and prevent overfitting, as in (Kooi, Litjens, Ginneken, Gubern-Mérida, Sánchez, Mann, Karssemeijer 2017), who studied large scale deep learning for computer aided detection of mammographic lesions. Here, all breast cancer images are rotated to artificially expand the size of a training dataset by creating modified versions of the same images This allows us to improve the performance and generalization ability of the model.

Second stage

In the training stage part of this study, we use a novel Convolutional Neural Network (CNN) model that has been previously proposed to classify benign or malignant tumors. Our methodology has achieved the highest accuracy rate (which is the ratio of the sum of the true positive and true negative predictions out of all the predictions), 99% and 97.0% in the train and test data sets, respectively. The high accuracy as well as other excellent evaluation indicators show that the CNN has a high performance. This is key in our mammography diagnosis.

Let us remind that the the loss is the error one tries to minimize in the AI process. Figures 2a and 2b provide an overview of the training process, by depicting the loss and the accuracy of the training (indicated in green) and validation datasets (indicated in blue) as a function of the epoch. As shown in Figure 2., in some cases, the loss function and accuracy of the validation set are better than the training set's counterparts. To clarify the reason, it should be noted that loss and accuracy are measured after each period of training. As the model improves in the learning process, malignancy status of cancer is more accurately detected in the validation data set compared to the training dataset.

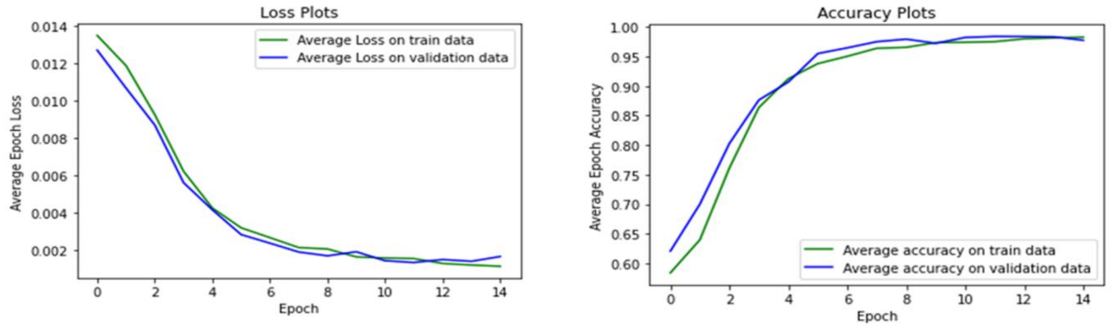

Figure 2 a. Loss Plots for Train and Test Dataset

Figure 2 b. Accuracy Plots for Train and Test Dataset

Figure 2. Loss and Accuracy Plots for the Training Phase

Figure 3 presents the confusion matrix, which presents the true negative and positives on the main diagonal, the false negatives on the top right cell and the false positives in the bottom left one. The sum of the main diagonal cells therefore indicates that, as anticipated above, the accuracy for the train set is 99% (Figure 3 a), and for the test set it is 97% (Figure 3 b). An implication of this result is that the pre-attack model will detect 97% of all patients with the correct type of cancer.

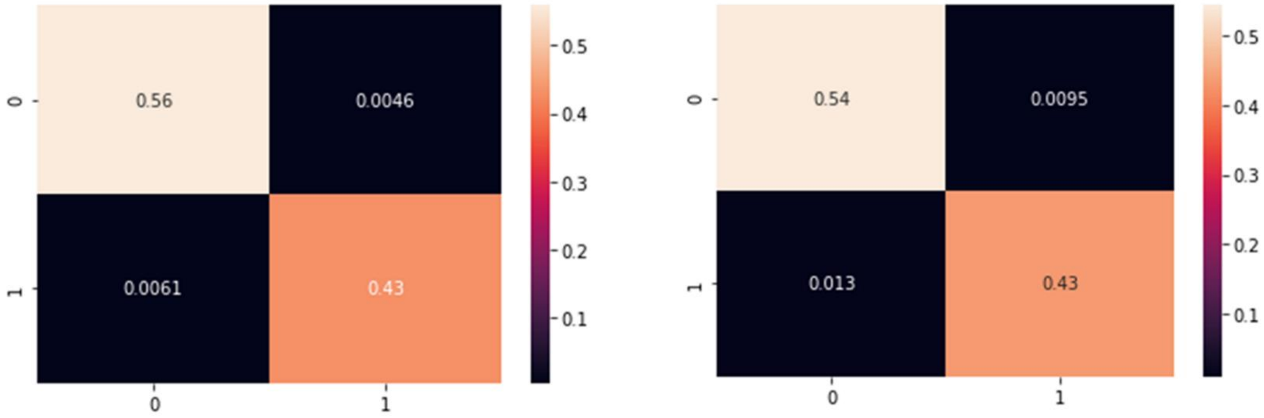

Figure 3 a. Confusion Matrix for Train Dataset

Figure 3 b. Confusion Matrix for Test Dataset

Figure 3. Confusion Matrices for Train and Test Dataset

Moreover, the sensitivity, which gives the model's probability for predicting malignancy when the patient has the malignant cancer, being the ratio of true positives over false negatives and true positives, is 91% for the training and 97.7% for the test dataset. Similarly, specificity, which indicates the probability of predicting a benign model when a patient has benign cancer, being the ratio of true negatives over false positives and true negatives, is very high: 99% and 99.76% in the train and test sets respectively.

Third stage

To evaluate the model robustness of the diagnosis system, we simulated an adversarial attack with two widely used methods: PGD and Universal Patch.

Figure 4 shows the results of a PGD attack on mammography images, where the first column shows clean images, the second and third columns show perturbations and the results of the misclassification of the attack.

Table 1 shows the results of our experiments with different degrees of perturbation in the PGD attack. By its very nature, higher degrees of perturbation lead to much lower performance of the target models. Although this results in a sure misdiagnosis of the systems, it also increases the probability of the insurer noticing when the systems are attacked with. Therefore, conspicuous perturbations that could be easily detected during the insurer's assessment can be weeded out without much effort. There is a trade-off between the perceptibility and the success rate in PGD Attack. A higher perturbation can make an attack appear as a certain deception of the classification system, but the attacked image may in the end appear to be falsified to a trained human eye.

Table 1 The Perturbation Impact on the Accuracy Level of the Target Model

Perturbation ϵ	Model Accuracy	Perturbation ϵ	Accuracy
0	0.959	0.006	0.626
0.001	0.927	0.007	0.610
0.002	0.878	0.008	0.569
0.003	0.821	0.010	0.512
0.004	0.756	0.015	0.431
0.005	0.691	0.20	0.390

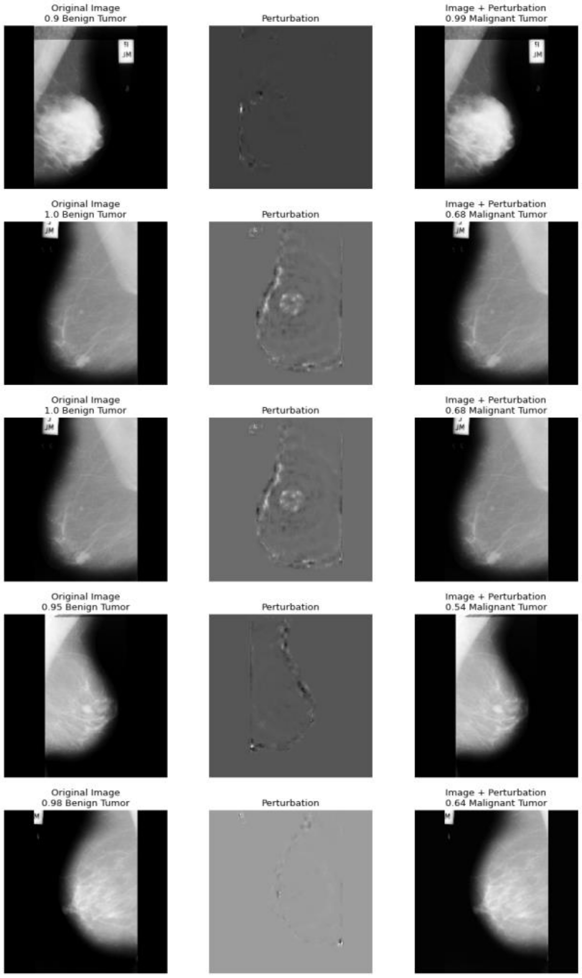

Figure 4. A Sample of Perturbation Caused by PGD and Misclassification Result

We also report another attack on medical images, the so-called universal attack, which comes directly from research of (Moosavi-Dezfooli, Fawzi, Fawzi, & Frossardy, 2017). Figure 5 shows the results of the universal attack on breast cancer image. Not only is benign tumor detected as malignant cancer, like PGD attacks, but in some cases, adversarial images are even diagnosed with higher accuracy than the original image.

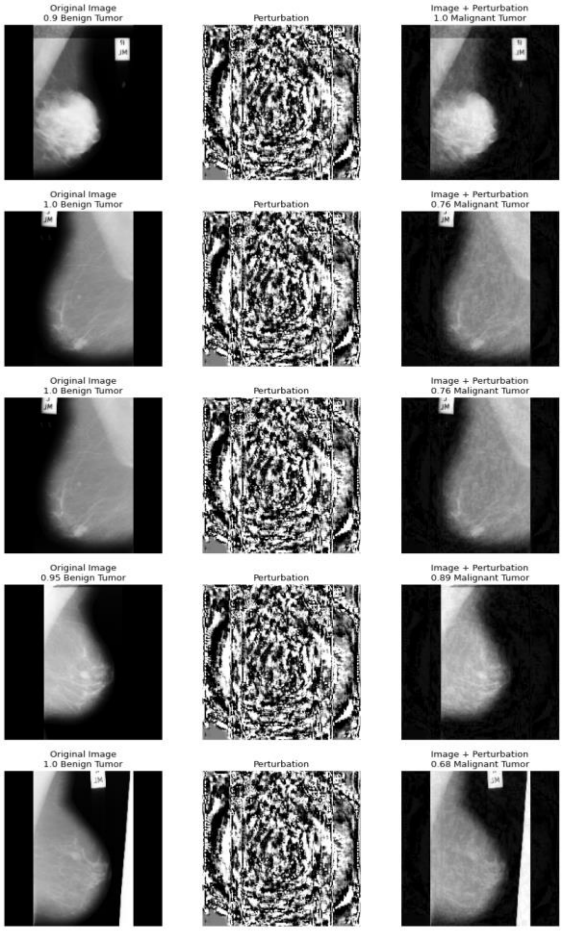

Figure 5. A Sample of Perturbation Caused by Universal attack and Misclassification Result

In the universal attack, the probability of a given classification after the attack can also be higher than before (see the first line of figure 5).

Overall, he results of the experiments show the model can be fooled for detecting the type of patients' cancer. Based on the above case studies, there are two possible scenarios:

1. Healthy individuals submit fake claims.
2. Patients can conceal their pre-existing (PED) conditions with malignant cancer.

For the first possible scenario, assume a person that has been diagnosed with a benign tumor based on their mammogram with 90% accuracy. If the system was attacked by either PGD or by a universal attack, this time the tumor will be detected as malignant even with higher accuracy than the pre-attacked image. On the other hand, in the second scenario, we consider an insurance applicant who seeks for insurance coverage and wants to have low premiums. Also, in this case the individual has an incentive to manipulate the images through medical doctors or image providers by concealing his risk factors. As it is clear in both cases, an

doi:10.20944/preprints202210.0157.v1

attacker - who could be an individual applying directly for insurance or anyone on his behalf - has a financial motivation to attack an automated AI system in order to gain a higher benefit.

4. Preparing for Adversarial Attacks

A number of papers and books suggest how to make AI systems more robust to adversarial attacks. Defenses are divided into heuristic defenses, whose effectiveness is based only on experimental evidence, and has no general validity, and certified or proved defenses, which exploit theoretical properties and are therefore principle-validated. The most well known category of heuristic defenses is adversarial training. Training works exactly as the word suggests, very much in the spirit of training of the whole aI approach. So-called certified defenses do not provide only a training, but also a certification of the accuracy of the AI application with and without specific adversarial attacks.

For a comprehensive taxonomy one can see Ren, Zheng, Qin and Liud (2020), who include among the heuristic methods FGSM adversarial training, PGD adversarial training, ensemble adversarial training, adversarial logit pairing, generative adversarial training, randomization, random input transformation, random noising, random feature pruning, denoising, conventional input rectification, GAN-based input cleaning, auto encoder-based input denoising, feature denoising, Among the provable defenses they distinguish semidefinite programming-based certificated defense, distributional robustness certification, weight-sparse DNNs, dual approach-based provable defense, KNN-based defenses, Bayesian model and consistency-based defenses. As for DNNs, which we have applied above, t he book by Warr (2019) explains how to make .applications of DNNs to image processing more resilient. (Xu, Ma, Liu, Deb, Liu, Tang, and Jain, 2020) extend the analysis to DNNs applied to graphs and text.

It is obvious from the rich taxonomy above that the artirelly at disposal in order to prevent adversarial attacks is rich. Sometimes defense methods are also quite powerful. Kurakin, Goodfellow and Bengio (2017b) for instance show, using ImageNet, that adversarially trained models perform better on adversarial examples than on non-attacked ones, as it happened in some of our examples. That is the case because when constructing the adversarial attack one uses the true characteristics of the example, and the model learns, as any AI model.

Most of the current literature however focuses on specific attacks and on how to strenghten the corresponding AI applications, because there is neither a universal patch nor a consensus on the best defense for a specific attack.

In that sense Ren et al. (2020) state that certified attacks are the state of the art, although, until now, they present the problem of being seldom scalable. There is no defense which succeeds in being efficient and effective against adversarial attacks, since the most effective defense, which according to them is an heuristic adversarial training, is too computationally intensive to be efficient. Other defenses, which are computationally less costly, are quite vulnerable and do not guarantee enough robustness in industries like finance and insurance.

While the research on defenses continues, it is our opinion that to make an AI model in insurance more robust against potential adversarial attacks requires a holistic view. It is not just about defending it through technical solutions, but about understanding the broader impact of such attacks on an organization, and detect where attacks can hurt more, so as to prioritize the search for resiliency.

The first countermeasure one can take adopting this holistic view is similar to the approach suggested by the European Commission and the Joint Research Centre when evaluating model risk and validating models for policy purposes, namely, to conduct sensitivity analysis on the input data and so-called "sensitivity audits". Sensitivity audits ascertain how model results used in impact assessments and elsewhere depend upon the information fed into them, their structure and underlying assumptions. It extends the impact assessment of model assumptions to different sets of input data. For examples of sensitivity audit applications see the EU Science Hub.[4]

Given the wide experience actuaries, risk managers, tariff producers have of data and its order of magnitude, sensitivity analysis is likely to be conducted effectively in insurance, at least when adversarial attacks are not as subtle as the health insurance one we provided, or when, even in that case, the image is complemented by more medical data about the patient.

This does not mean to withhold innovation. In this sense, the World Economic Forum recommends to empower employees so that they look responsibly at AI and raise concerns about it, with the aim of making innovation more helpful, not of lowering its pace. Also in preparing against adversarial attacks, what

[4] https://ec.europa.eu/search/?QueryText=sensitivity+audit&op=Search&swlang=en&form_build_id=form-WZ65edbU064IlfvfZtaOeEFLhj5IOLUbYfEFLNJ707Q&form_id=nexteuropa_europa_search_search_form

matters according to the World Economic Forum is to find an equilibrium between the autonomy of AI and human oversight.[5] Sensitivity is just one step in that direction.

5. Conclusions

Summing up, insurance fraud can be typically qualified as soft insurance fraud and hard insurance fraud. Soft fraud is usually unplanned and arises when the opportunity presents itself. It is the more prevalent form of fraud. An example of this type of fraud would be getting into a car accident and claiming your injuries are worse than they really are, getting you a bigger settlement than you would get if you were telling the truth about your injuries. Hard fraud takes planning. An example of hard fraud would be falsifying documentation of an accident on purpose so that you can claim the insurance money. Adversarial attacks qualify as hard insurance fraud. The ability to monitor and pre-empt adversarial attacks requires insurance companies to upskill their abilities, beyond, for example, recourse to private investigators[6]. This recourse has been quite pervasive in the US, because there, the total cost of insurance fraud (excluding health insurance) is estimated to exceed $40 billion per year, which means an increase in premiums $400 and $700 per year and per household. In the era of AI, defenses against adversarial AI could save a lot of this money.7 How to do this is still an open issue, with an holistic, sensitivity-based approach as a first, universal defense, together with a balance between autonomy and human oversight in AI applications.

Appendix

If an attacker has access to the architecture and parameters of the model, these models are called white-box attacks. If not, these methods are called black-box attacks.

1. White-Box Attacks

To theoretically explain the adversarial attack of group “a”, let the input domain $X \in R^d$, the class domain be $Y \in \{0,1\}^C$, and let $H(x): X \to Y$ be a functional mapping the d −dimensional input domain X to a C − dimensional discrete class domain. Denote the loss function of a network by $J(\theta, x, y)$, where θ are the parameters of the network, x is the input image and y is the class label associated with x. Given a test image x with class y, the goal of an attack procedure is to generate a new image x_{adv} such that $H(x_{adv}) \neq y$ and the amount of perturbation is minimized:

[5] http://ow.ly/o1Uh50L2n5E

[6] See for instance https://www.pinow.com/articles/305/insurers-on-the-alert-for-false-claims-turn-to-private-investigators

[7] See https://www.fbi.gov/stats-services/publications/insurance-fraud

$$minimize \ \|x_{adv} - x\|_p \ s.t. H(x_{adv}) \neq y \tag{1.1}$$

where $\|.\|_p$is the norm that measures the extent of perturbation. Some commonly used $L_p$norms are L_0, L_2, or L_∞. This, as mentioned earlier, applies to an untargeted attack, which means that the attacker only needs to perturb input x to any class that is incorrect. The attack can also be "targeted", in which case the input x is perturbed into a specific incorrect class $y_{target} \neq y$. Accordingly, the problem of the targeted adversarial attack generation is defined as:

$$minimize \ \|x_{adv} - x\|_p \ s.t. H(x_{adv}) = y_{target} \neq y \tag{1.2}$$

In general, targeted adversarial examples are more difficult to generate than untargeted adversarial examples. Different ways to solve both (1.1) and (1.2) lead to different attack methods that have been proposed to generate adversarial examples to attack DNN. Note that the generation of adversarial examples is a post-processing method for an already trained network. Therefore, adversarial generation updates the input x instead of the model parameters, which contrasts with network training where the parameters θ are updated. Moreover, adversarial generation aims to maximize the loss function to fool the network to make errors, while in the training phase the network aims to minimize the loss function. The following is an overview of the most widespread adversarial attacks.

- Fast Gradient Sign Method
- Projected Gradient Descent
- DeepFool
- Carlini and Wagner

1.1 Fast Gradient Sign Method: The Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015) can be targeted or untargeted. FGSM falls into the group that maximizes the attack success rate given a limited budget, which perturbs each feature of an input x by a small amount towards maximizing the prediction loss $J(\theta, x, y)$. FGSM performs a single gradient descent step in the case of a targeted attack (t is the target label instead of true label y)

$$x_{adv} = x - \epsilon.sign\ (\nabla_x J(\theta, x, t)) \tag{1.3}$$

and a single gradient ascent step in the case of untargeted attack

$$x_{adv} = x + \epsilon.sign\ (\nabla_x \ J(\theta, x, y)) \tag{1.4}$$

doi:10.20944/preprints202210.0157.v1

FGM is a fast method that only perturbs the input once, with ϵ – the direction of the steepest ascent - that is fixed. Therefore, it is not guaranteed to successfully perturb the input to an adversarial class (i.e., $H(x_{adv}) \neq y$). The success rate can be improved by increasing the perturbation magnitude ϵ, although this may result in large perturbations that are perceptible to human observers.

1.2 Projected Gradient Descent: As a simple extension of FGSM, Projected Gradient Descent (PGD) (Kurakin et al., 2017a) applies FGSM iteratively with a small step size and projects the intermediate results around the original image x. In (1.5), $clip_{x,\epsilon}(.)$ is an element-wise clipping to ensure that this condition is satisfied. In general, the projection onto an $\epsilon - l^p$ –ball is a difficult problem and closed form solutions are only known for a few values of p. Formally, it is

$$x_{adv}^0 = x, x_{adv}^i = clip_{x,\epsilon}(x_{adv}^{i-1} + \epsilon sign(\nabla_{x_{adv}^{i-1}} \; J(\theta, x_{adv}^{i-1}, y))) \tag{1.5}$$

The perturbation process can stop in two cases: first, when the misclassification $H(x_{adv}) \neq y$ is reached, or second, when a fixed number of iterations has been performed.

Another white-box attack method is called Iterative FGSM (I-FGSM). It was introduced in (Kurakin et al., 2017b) and it iteratively performs the FGSM attack. This is an improved white box attack in which the FGSM attack is updated iteratively at a smaller step size and clips the signals of the intermediate results to ensure its proximity to the original signal. Essentially, I-FGSM is the same as PGD, the only difference being that the PGD attack initializes the perturbation with a random noise, while I-FGSM initializes the perturbation with only zero values (Zhang , Benz, Lin, Karjauv, Wu and Kweon, 2021). This random initialization can help improve the success rate of the attack, especially when the number of iterations is limited to a relatively small value.

1.3 DeepFool: The DeepFool algorithm (Moosavi-Dezfooli, Fawzi, & Frossard, 2016) was developed with the goal of providing an efficient yet accurate method for computing minimal perturbations with respect to the l^p –norm. Since DeepFool iteratively produces the perturbations by updating the gradient with respect to the decision boundaries of the model, it falls into the attack category that attempts to minimize the size of the perturbation. The authors propose DeepFool as an untargeted attack, but the algorithm can in principle be easily modified for the targeted setting.

By considering DNNs, Dezfooli et al. argue that the minimum perturbation of the adversary can be constructed as an orthogonal projection onto the nearest decision boundary hypersurface. To account for the fact that DNNs are not truly linear, the authors propose an iterative procedure in which the orthogonal projection onto the first-order approximation of these decision boundaries is computed at each step. The search ends with finding a true adversarial example (Qiu et al., 2019).

1.4 Carlini & Wagner Attack (C&W): C&W's attack (Carlini and Wagner, 2017) attempts at finding the minimally biased perturbation problem - similarly to the DeepFool algorithm - as follows:

$$min\|x - x'\|_2^2 + c.H(x', t), \quad s.t\ x' \in [0,1]^m \qquad \text{(II.6)}$$

Carlini and Wagner study several loss functions and find that the loss that maximizes the gap between the target class logit and the highest logit (without the target class logit) leads to superior performance (C. Zhang et al., 2021). Then H is defined as $H(x', t) = (max_{i \neq t} Z(x')_i - Z(x')_t)^+$, where Z is the last layer score in DNNs before the so-called Softmax. Minimizing $J(x', t)$ encourages the algorithm to find an x' that has a larger score for class t than any other label, so the classifier will predict x' to be class t. Next, by applying a line search to the constant c, we can find the one that has the smallest distance from x.

The function $H(x, y)$ can also be considered as a loss function for data as $J(x, y)$. It penalizes the situation where there are some labels i whose values $Z(x)_i$ are larger than $Z(x)_y$. It can also be called a margin loss function.

The authors claim that their attack is one of the strongest attacks that breaks many defense strategies that have proven to be successful. Therefore, their attack method can be used as a benchmark to study the security of DNN classifiers or the quality of other adversarial examples.

2 Black-Box Attacks

While the definition of a "white-box" attack on DNNs is clear and precise, i.e., providing complete knowledge of and full access to a targeted DNN, the definition of a "black-box" attack on DNNs may vary with respect to an attacker's capabilities. From an attacker's perspective, a black-box attack may refer to the most difficult case where only benign images and their class labels are given, but the targeted DNN is completely unknown. Therefore, attacks that mainly focus on backpropagation information which is not available in the black box setting. Here, two common black-box attacks are described:

- Substitute Model
- Gradient Estimation

2.1 Substitute Model: the paper (Papernot, McDaniel, Goodfellow, Jha, Celik, and Swami, 2017) presented the first effective algorithm for a black-box attack on DNN classifiers. An attacker can only input x to obtain the output label y from the classifier. The attacker may have only partial knowledge of 1) the classifier's data domain (e.g., handwritten digits, photographs, human faces) and 2) the classifier's architecture (e.g., CNN, DNN).

The authors in (Zhang et al., 2021) exploit the “transferability” property (defined in Section 2.3 above) of adversarial examples: an example x' can attack H_1, it is also likely to attack H_2, which has similar structure to H_1. Therefore, the authors present a method to train a surrogate model H' to mimic the target-victim classifier H, and then create the adversarial example by attacking surrogate model H'. The main steps are as follows:

1. Synthesize a substitute training dataset: Create a “replica” training set. For example, to attack handwritten digits recognition, create an initial substitute training set X by: a) requiring samples from the test dataset; or b) creating handcrafting samples.
2. Training the surrogate model: Feed the surrogate training dataset X into the victim classifier to obtain their labels Y. Select a surrogate DNN model to train on (X, Y) to obtain H'. Based on the attacker′s knowledge, the chosen DNN should have a similar structure to the victim model.
3. Dataset augmentation: Augment the dataset (X, Y) and iteratively re-train the substitute model H'. This procedure helps to increase the diversity of the replica training set and improve the accuracy of the substitute model H'.
4. Attacking the substitute model: use the previously presented attack methods, such as FGSM, to attack the model H'. The generated adversarial examples are also very likely to mislead the target model H, due to the “transferability” property.

2.2 Gradient Estimation: Another approach for black-box attacks is the gradient estimation method ZOO, proposed by (Chen, Zhang, Sharma, Yi and Hsieh, 2017). They apply zero-order optimization over pixel-wise finite differences to estimate the gradient, and then construct adversarial examples based on the estimated gradient using white-box attack algorithms.

According to their assumption of having access to the prediction confidence from the output of the victim classifier, it is not necessary to build the substitute training set and model. Chen et al. give an algorithm to obtain the gradient information around the victim sample by observing the changes in the prediction confidence $H(x)$ as the pixels of x are changed.

Equation (1.7) shows that for each index i of sample x, we add (or subtract) to an ϵ multiple of another vector $e_{i,}$ to have $x_i = x \pm \epsilon e_i$ by. If ϵ is small enough, we can extract the gradient information for $H(.)$ by

$$\frac{\partial H(x)}{\partial x_i} \approx \frac{H(x + \epsilon e_i) - H(x - \epsilon e_i)}{2\epsilon} \tag{1.7}$$

3. Universal Attack

Adversarial attacks described so far always manipulate a single image to fool a classifier with the specific combination of the image and an adversarial perturbation. In other words, these perturbations are image dependent, i.e., one cannot apply a perturbation designed for image A to another image B and expect the attack to work successfully. In the paper (Moosavi-Dezfooli et al., 2017), an algorithm was presented to create universal or image-independent perturbations. Universal perturbations can pose a greater threat than the previous ones in this Appendix. The goal of a universal perturbation is to make the classifier classify the perturbed image differently from what is should, on at least a percentage $1-\delta$ of cases. Let $H()$ be the classifier, η be the adversarial perturbation, P denote the probability. The universal goal is

$$P\big(H(x+\eta) \neq H(x)\big) \geq 1-\delta$$

This goal must be reached under a constraint, that the distance of the perturbed image from the original is small, to ensure imperceptibility of the perturbation and to fool as many images as possible:

$$\|\eta\|_p \leq \epsilon \tag{1.8}$$

In the constraint the p −norm is required to be smaller than a constant ϵ meat to be small.

References

Carlini, N., & Wagner, D. (2017). Towards Evaluating the Robustness of Neural Networks. *In Proceedings of IEEE Symposium on Security and Privacy*, 39–57. doi:10.1109/SP.2017.49

Chen, P. Y., Zhang, H., Sharma, Y., Yi, J., & Hsieh, C. J. (2017). ZOO: Zeroth Order Optimization based Black-Box Attacks to Deep Neural Networks without Training Substitute Models. *In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 15–26. doi:10.1145/3128572.3140448

Finlayson, S. G., Chung, H. W., Kohane, I. S., & Beam, A. L. (2019). Adversarial Attacks Against Medical Deep Learning Systems. *ArXiv, abs/1804.05296*.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. *Conference paper at ICLR*.

Hirano H., Minagi A., & K., Takemoto. (2021). Universal adversarial attacks on deep neural networks for medical image classification. *BMC Med Imaging*, *21*(1). doi:10.1186/s12880-020-00530-y.

Kooi, T., Litjens, G., Ginneken, B., Gubern-Mérida, A., Sánchez, C., Mann, R., . . . Karssemeijer, N. (2017). Large scale deep learning for computer aided detection of mammographic lesions. *Medical Image Analysis*, *35*, 303-312. doi:10.1016/j.media.2016.07.007

Kurakin, A., Goodfellow, I., & Bengio, S. (2017a). Adversarial Examples in the Physical World. *ICLR*, 14.

Kurakin, A., Goodfellow, I., & Bengio, S. (2017b). Adversarial Machine Learning at Scale. *ICLR*. doi:arXiv:1611.01236v2

Li, B., Ge, Y., Zhao, Y., Guan, E., & Yan, W. (2018). Benign and malignant mammographic image classification based on Convolutional Neural Networks. *Association for Computing Machinery*. doi:https://doi.org/10.1145/3195106.3195163

Mirsky, Y., Mahler, T., Shelef, I., & Elovici, Y. (2019). CT-GAN: Malicious Tampering of 3D Medical Imagery using Deep Learning. *USENIX Security 2019*.

Moosavi-Dezfooli, S. M., Fawzi, A., Fawzi, O., & Frossardy, P. (2017). Universal Adversarial Perturbations. *IEEE Conference on Computer Vision and Pattern Recognition*.

Moosavi-Dezfooli, S. M., Fawzi, A., & Frossard, P. (2016). A Simple and Accurate Method to Fool Deep Neural Networks. *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2574–2582. doi:10.1109/CVPR.2016.282

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). A Practical Black-Box Attacks against Machine Learning. *Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security*, 506–519. doi:10.1145/3052973.3053009

Qiu, S., Liu, Q., Zhou, S., & Wu, C. (2019). Review of Artificial Intelligence Adversarial Attack and Defense Technologies. *Applied Science*, *9*(909). doi:10.3390/app9050909

Ren, K., Zheng, T., Qin, Z., & Liud, X. (2020). Adversarial Attacks and Defenses in Deep Learning. *3*(6), 346-360.

Sadeghi, S., Dadkhah, S., Jalab, H. A., Mazzola, G., & Uliyan, D. (2018). State of the Art in Passive Digital Image Forgery Detection: Copy-Move Image Forgery. *Pattern Analysis and Applications*, *21*(2), 291–306.

Singh, A. K., Kumar, B., Singh, G., & Mohan, A. (2017). Medical Image Watermarking Techniques: A Technical Survey and Potential Challenges. *Springer International Publishing*, *Cham*, 13–41. doi:https://doi.org/10.1007/978-3-319-57699-2_2

Suckling, J. (1996). The mammographic image analysis society digital mammogram database. Retrieved from https://www.kaggle.com/kmader/mias-mammography

Warr, K. (2019). *Strengthening Deep Neural Networks, Making AI Less Susceptible to Adversarial Trickery*: O'Reilly Media, Inc.

Wetstein, S. C., Gonz´alez-Gonzalo, C., Bortsova, G., Liefers, B., Dubost, F., Katramados, I., . . . Veta, M. (2020). Adversarial Attack Vulnerability of Medical Image Analysis Systems: Unexplored Factors. *ArXiv*.

Xu, H., Ma, Y., Liu, H., Deb, D., Liu, H., Tang, J., & Jain, A. K. (2020). Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. *International Journal of Automation and Computing, 17*(2), 151-178.

Xu, K., Zhang, G., Liu, S., Fan, Q., Sun, M., Chen, H., . . . Lin, X. (2020). Adversarial T-shirt! Evading Person Detectors in A Physical World. *Computer Vision and Pattern Recognition*. doi:arXiv:1910.11099v3

Zhang, C., Benz, P., Lin, C., Karjauv, A., Wu, J., & Kweon, I. S. (2021). A Survey On Universal Adversarial Attack. *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. doi:arXiv:2103.01498v1

Zhang, J., & Li, C. (2018). Adversarial Examples: Opportunities and Challenges. *IEEE Transactions on Neural Networks and Learning Systems*, 16. doi:10.1109/TNNLS.2019.2933524