risks

MDPI

*Article*

# Adversarial Artificial Intelligence in Insurance: From an Example to Some Potential Remedies

**Behnaz Amerirad [1], Matteo Cattaneo [2], Ron S. Kenett [3,4] and Elisa Luciano [2,4,5,*]**

[1] Desautels Faculty of Management, McGill University, Montréal, QC H3A 1G5, Canada
[2] Reale Group, 10122 Torino, Italy
[3] Samuel Neaman Institute, Technion City, Haifa 32000, Israel
[4] Department of Economics, Social Studies, Applied Mathematics and Statistics, University of Torino, 10134 Torino, Italy
[5] Collegio Carlo Alberto, 10122 Torino, Italy
* Correspondence: elisa.luciano@unito.it

**Abstract:** Artificial intelligence (AI) is a tool that financial intermediaries and insurance companies use or are willing to use 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, and after-sales services. While it is very important and useful, AI is not free of risks, including those related to its robustness against so-called adversarial attacks, which are conducted by external entities to misguide and defraud the AI algorithms. The paper is designed to review adversarial AI and to discuss its implications for the insurance sector. We give a taxonomy of adversarial attacks and present an original, fully fledged example of claims falsification in health insurance, as well as some remedies which are consistent with the current regulatory framework.

**Keywords:** AI in insurance; adversarial AI; insurance fraud; machine learning in insurance (ML)



**Citation:** Amerirad, Behnaz, Matteo Cattaneo, Ron S. Kenett, and Elisa Luciano. 2023. Adversarial Artificial Intelligence in Insurance: From an Example to Some Potential Remedies. *Risks* 11: 20. https://doi.org/10.3390/risks11010020

Academic Editor: Mogens Steffensen

Received: 7 October 2022
Revised: 2 January 2023
Accepted: 5 January 2023
Published: 11 January 2023



## 1. Introduction

AI comprises a set of techniques and technologies that the financial industry considers to be a strategic priority and a potential source of relevant competitive advantages and in which it invests 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 and wearable devices. Indeed, AI is used more and more in product design and development, pricing and underwriting, marketing and distribution, customer service and relationships with clients, and claims management, from claims filing to settlement.

While it is very important and useful, AI is not free of risks. The European Insurance and Occupational Pensions Authority (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 that are protected from cyberattacks. Within the notion of cyberattacks, one can confidently consider adversarial AI attacks aimed at defrauding an AI system in such a way that the result is undetected by humans.

Adversarial attacks against AI are a very modern form of fraud against underwriters; to understand them, in Section 2 we review frauds in the insurance sector and their impact. In Section 3, we associate some typical applications of AI with their corresponding example of adversarial attack/fraud. To prevent and fix adversarial attacks, an underwriter should categorize such attacks according to several dimensions, i.e., they should have a taxonomy. We provide it in Section 4. Section 5 provides our own example built on the public data from health insurance, to illustrate the subtlety and powerfulness of those attacks. We provide practical remedies against adversarial attacks in Section 6, and we conclude in Section 7.

## 2. Fraud in Insurance

Unfortunately, fraud against insurance companies has been perpetrated by customers (as well as by criminal organizations tout court) almost always, but not exclusively, in the claim phase,. Despite legislative interventions and the vast efforts that underwriters put into detecting fraud, the estimated amount is still surprisingly high. According to the European federation of insurance companies, "Insurance Europe", the fraud cost in Europe has been estimated at 10% of the total amount of claims, which is equivalent to the cost of EUR 13 billion per year for European citizens. Healthcare fraud is also quite widespread, with the unfortunate cooperation of several actors, including hospitals, doctors, and nurses. It covers both billing excessive expenses, billing unperformed treatments or services or drugs, inflating the drug type and cost, or even submitting a diagnosis which gives the right to more of a refund than a true one. The National Healthcare Anti-Fraud Association estimates that 3% of all healthcare spending in the United States is lost to healthcare fraud, while the European Healthcare Fraud & Corruption Network (EHFCN) estimates that European Union countries lose about EUR 56 billion to healthcare fraud in Europe each year.[1]

Insurance fraud can be 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 that your injuries were worse than they really were, 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 the documentation of an accident on purpose so that you could claim the insurance money.

Fraud prevention, both soft and hard, is obviously crucial to insurance companies, and academia has worked a lot in order to strengthen the assessment of fraud risk, based on the observed past data.

A traditional approach relies on audits or costly state verifications that assume insurers can indeed verify whether there is fraud by performing a control and paying the corresponding cost. Following an audit, a claim turns out to be classified according to a binary criterion: it is either fraudulent or not. Dionne et al. (2008) study an optimal strategy for audit-based fraud detection and characterize it as a "red flag" one: when some indicator thresholds are overcome, then a special investigation is needed. Those indicators depend on the claim frequency, and the corresponding optimal policy is also verified on the empirical side to act as a fraud deterrent.

Because audits can fail to classify a claim correctly, and frauds can go undetected, the need to have methodologies other than costly state verification, which give a probability of being fraudulent as a result, has arisen (see Caron and Dionne 1999). Artís et al. (1999) develop that idea, and by doing so, they provide a much more nuanced and cautious approach.

Given that the quality and coverage of the data used for that assessment may be diverse, Artís et al. (2002) enhance the methodology, taking the misclassification error into account. They develop a maximum likelihood estimate with that feature and show that it performs very well in the automobile case. This is per se very important, given that in some countries, such as Italy, third-party insurance is compulsory.

Both audit-based and probabilistic assessment approaches can greatly benefit from the support of machine learning and AI, as we explain in the next section.

## 3. AI and Adversarial AI in Insurance

AI has found large applications in the insurance sectors, in almost all phases of the business, given that insurance companies can usually count on rich and always evolving datasets.

First of all, AI can be used to design new products or services so as to tailor them to the needs of customers or potential customers. The information acquired on a particular target market at this stage can be powerfully submitted to a machine learning algorithm to predict—say—the frequency as well as the severity of the damages incurred if we are in

the P&C domain or the life expectancy of the insured if we are in the life and similar-to-life one. Both the quality and the quantity of the granular data are relevant, but so are their integrity: data corrupted by adversarial attacks can lead to products that do not correspond to an insurance company's typical customer profile. An example in that sense is the ability of adversarial AI to corrupt text, as submitted by—say—a potential customer to a chatbot when describing his requests.

A second, related, and important use of AI in insurance consists in the assessment of the riskiness of a potential client. This is key in pricing the products offered to her while respecting non-discrimination. Adversarial attacks can target the risk profile of a customer—in a collective contract, say—in order to provide her with a cheaper contract, to the detriment of the underwriter. In this sense, adversarial attacks are a modern form of adverse selection realization. An example in that sense would again be the text corruption of the collective risk profile or captioning corruption in order to mitigate it without perception by the underwriters. We will explain below that text corruption may have such a property.

A third important application of AI in insurance consists in designing the whole profile of a potential customer, beyond her risk profile, and permitting the customer-centric design of marketing campaigns or product and services bundling. In the era in which customers can remain in contact with their insurance company or get in contact with it through apps, which are often also used for buying instant contracts (health or car accident, for instance), it is important that the info they provide is instantly transformed by the app into a suggestion to the client. This usually happens through the adoption of AI algorithms. The same applies to the post-sale relationship of a client with her insurance company, which often goes through chatbots. Again, AI is the spirit of chatbots. Chatbots and apps are therefore a likely goal of adversarial attacks, which can worsen the quality of the performance of the interaction or step into it and make the customer's experience very bad.

The last very large domain of AI adoption in insurance is fraud prevention. The PRIDIT approach (Principal Component Analysis of RIDITs, which are the indicators of the threshold variables for claim detection) was introduced by Brockett et al. (2002). Ai et al. (2013) improve on the PRIDIT approach by developing PRIDIT-fre, which addresses the bias of unobserved frauds. A number of other articles focus instead on the best machine learning (ML) algorithm to detect fraud. In doing this, they obtain results in line with the corresponding algorithms' predictive power, which we will also encounter in our application for health insurance. See, for instance, the review in Grize et al. (2020).

Obviously, adversarial AI also applies to fraud detection. Our example in Section 5 will be exactly of that type: the adversarial attacker, while submitting the image of a patient or several patients for refund, will compromise the AI algorithm of the insurer so that the latter will consider the illness to be much more serious (and refund-deserving) than it truly is.

## 4. Adversarial Attacks and Their Taxonomy

An adversarial example is a sample of text or image input data that has been very slightly altered in a way that is intended to mislead an ML system (Kurakin et al. 2017a). The result is that the AI application makes incorrect predictions.

AI insurance applications on images, videos, text, or voice, and even captioning of images are becoming increasingly sophisticated. They are vulnerable to adversarial attacks based on specific perturbations of their input data. Sometimes these perturbations can be small and imperceptible to human detection, both for text and images. Not only are ML-systems fooled in their detection but a higher level of these perturbations can also 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 et al. 2015). In this example, the author explained how small, invisible perturbations on the input pixels of a panda image result in its misclassification as a gibbon. Appendix A clarifies how adversarial attacks are generated based on different algorithms.

The taxonomy of adversarial attacks is based on either their goal or their properties or their capabilities, which we list in Table 1 and illustrate in the next sections.

**Table 1.** Adversarial Taxonomy.

| **Adversarial Goal** | **Adversarial Properties** | **Adversarial Capability** | |
|---|---|---|---|
| **Confidence Reduction** | | **Training** | **Inference** |
| Misclassification | Transferability | Data Injection | White-Box Attack |
| Targeted Misclassification | Adversarial Instability | Data Modification | Black-Box Attack |
| Source Misclassification | Regularization Effect | Logic Corruption | Gray-Box Attack |

### *4.1. Adversarial Goal*

Based on their goal, adversarial attacks can be divided into four categories: those that lead to confidence reduction, to an untargeted misclassification, and to a target misclassification and those that target a specific source for misclassification (Qiu et al. 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.[2]

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.

### *4.2. Adversarial Capability*

While attacks in the training phase seek to learn from, 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 et al. 2020).

**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 dataset in order to falsify the target model.

**Data Modification:** Without access to the learning algorithm, but to all the training data, the attacker can poison the target model by manipulating the training data.

**Logic Corruption:** The attacker has access to and can 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, the model parameters, and the 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, the attackers do not know the structure of the target networks and parameters but exploit system vulnerabilities using information about the environment or past inputs. The black-box attacks can always compromise a naturally trained non-defensive system.

**Gray-Box Attack:** In the gray-box model, it is assumed that an attacker knows 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 A, more details on the further split of the white- and black-box attack models, together with their intuition.

### 4.3. Adversarial Properties

Adversarial attacks have three basic properties,: (i) transferability, (ii) adversarial instability, and (iii) the possibility of reaching a 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 the 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:** This consists in the training of the AI so as to reveal its defects—especially neural network systems—and consequently to improve its resilience. Adversarial training (Goodfellow et al. 2015), which we discuss in Section 6, is an example of a regularization method.

## 5. A Health Insurance Example and the Assessment of Damages

The adversarial attack example we provide below consists in 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 when households rely on insurance coverage. Unfortunately, fraud can be committed by a diverse set of individuals, including professionals in healthcare together with their patients. The amount of consequent damage runs the risk of being inflated by fraud. We know from Section 2 that fraud is extremely costly for European insurers.

Fraud in health insurance may occur both in the underwriting phase and in the claim filing one. In the first case, at least in Europe, it may occur when the applicants make false or misleading claims or at least provide 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 status, such as being a smoker or not. For this evaluation, insurers are allowed to ask questions about applicants' pre-existing conditions and then decide 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 et al. 2017; Sadeghi et al. 2018). For example, the attacks may consist in injecting and removing pixels on the CT scans of patients' lung cancer (Mirsky et al. 2019).

Finlayson et al. discuss how pervasive fraud is 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 Appendix A).

Another study (Wetstein et al. 2020) evaluated several unexplored factors, including the degree of perturbation and the transmissibility of the adversarial attack, which affect the susceptibility of DL, in Medical Image Analysis systems (MedIA) mainly focused on diabetic retinopathy detection, chest X-rays for thoracic diseases, and histopathological images of lymph node sections.

(Hirano et al. 2021) used universal attack in clinical diagnosis to classify skin cancer, diabetic retinopathy, and pneumonia. Their results confirmed that Deep Neural Networks (DNNs) are susceptible to universal attacks, resulting in an input being assigned to an incorrect class and causing the DNN to classify a specific input into a specific class.

In another study (Joel et al. 2021), the authors emphasize the susceptibility of Deep Learning (DL) models to adversarial attacks for three common imaging tasks within oncology, and they found that DL models for medical images are more vulnerable to adversarial attacks compared to DL models for non-medical images. In particular, attacks on medical images require a more negligible perturbation than attacks on non-medical images in general. Thus, they concluded that the adversarial sensitivity of individual images could improve model performance by identifying the images most at risk of misclassification.

Moreover, adversarial attacks are reviewed for ultrasound (US) imaging in the fatty liver by small perturbation, which is reconstructed based on radio-frequency signals by applying zeroth-order optimization (Byra et al. 2020). Then, the authors illustrate the accuracy of their proposed approach using the DL model, which was previously designed for fatty liver disease diagnosis and has been reduced to less than a 50% success rate. In this regard, we examine an example that is built on public data and proceed as follows, to see how these attacks will affect the insurance company.

*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 1024 $\times$ 1024 pixels. In the MIAS database, 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**

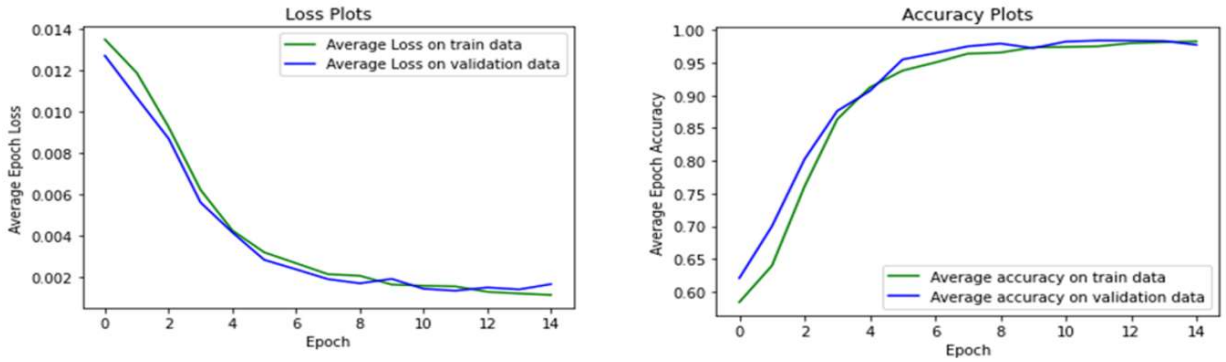
A common step in computer-aided diagnosis systems is preprocessing, which improves the characteristics of the image by applying a series of transformations to improve performance (Li et al. 2018). An applied approach in this research is data augmentation, often used in the context of Deep Learning (DL), which refers to the process of generating new samples from existing data and is used to improve data sparsity and prevent overfitting, as in (Kooi et al. 2017). 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 training and test datasets, respectively. The high accuracy as well as other excellent evaluation indicators show that CNN has a high performance. This is key in our mammography diagnosis.

Figure 1a,b 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. The loss is the error one tries to minimize in the AI process. As shown in Figure 1, in some cases the loss function and accuracy of the validation set are

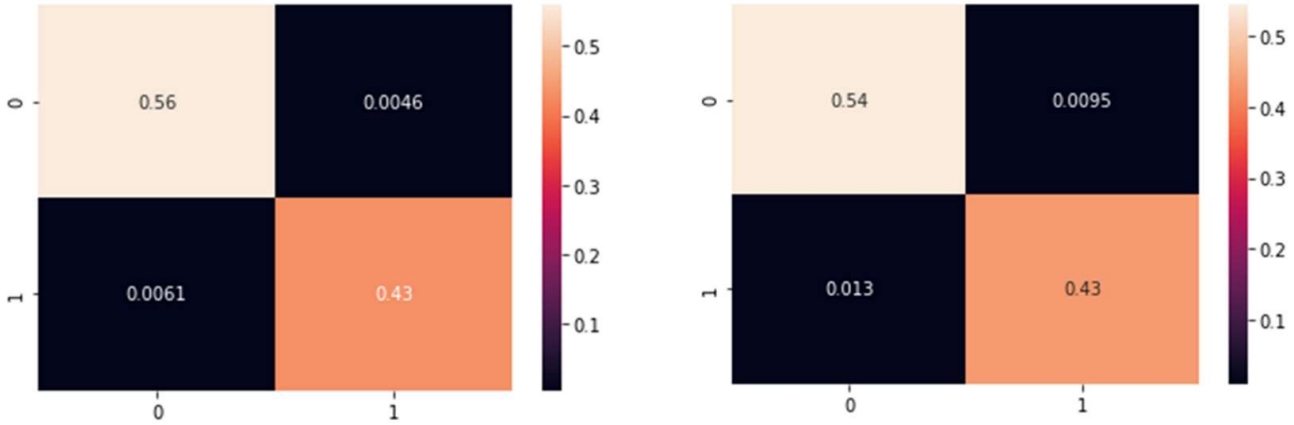
better than the training set's counterparts. To clarify the reason, it should be noted that loss and accuracy are measured after each period of training. As the model improves in the learning process, the malignancy status of the cancer is more accurately detected in the validation dataset as compared to the training dataset.




a b

**Figure 1.** Loss and accuracy plots for the training phase: (**a**) loss plots for training and test datasets; (**b**) accuracy plots for training and test datasets.

Figure 2 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 training set is 99% (Figure 2a) and for the test set it is 97% (Figure 2b). An implication of this result is that the pre-attack model will detect 97% of all patients with the correct type of cancer.




a b

**Figure 2.** Confusion matrices for training and test dataset: (**a**) confusion matrix for training dataset; (**b**) confusion matrix for 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 training 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 3 shows the results of a PGD attack on mammography images, where the first column shows clean images and the second and third columns show the perturbations and the results of the misclassification of the attack.

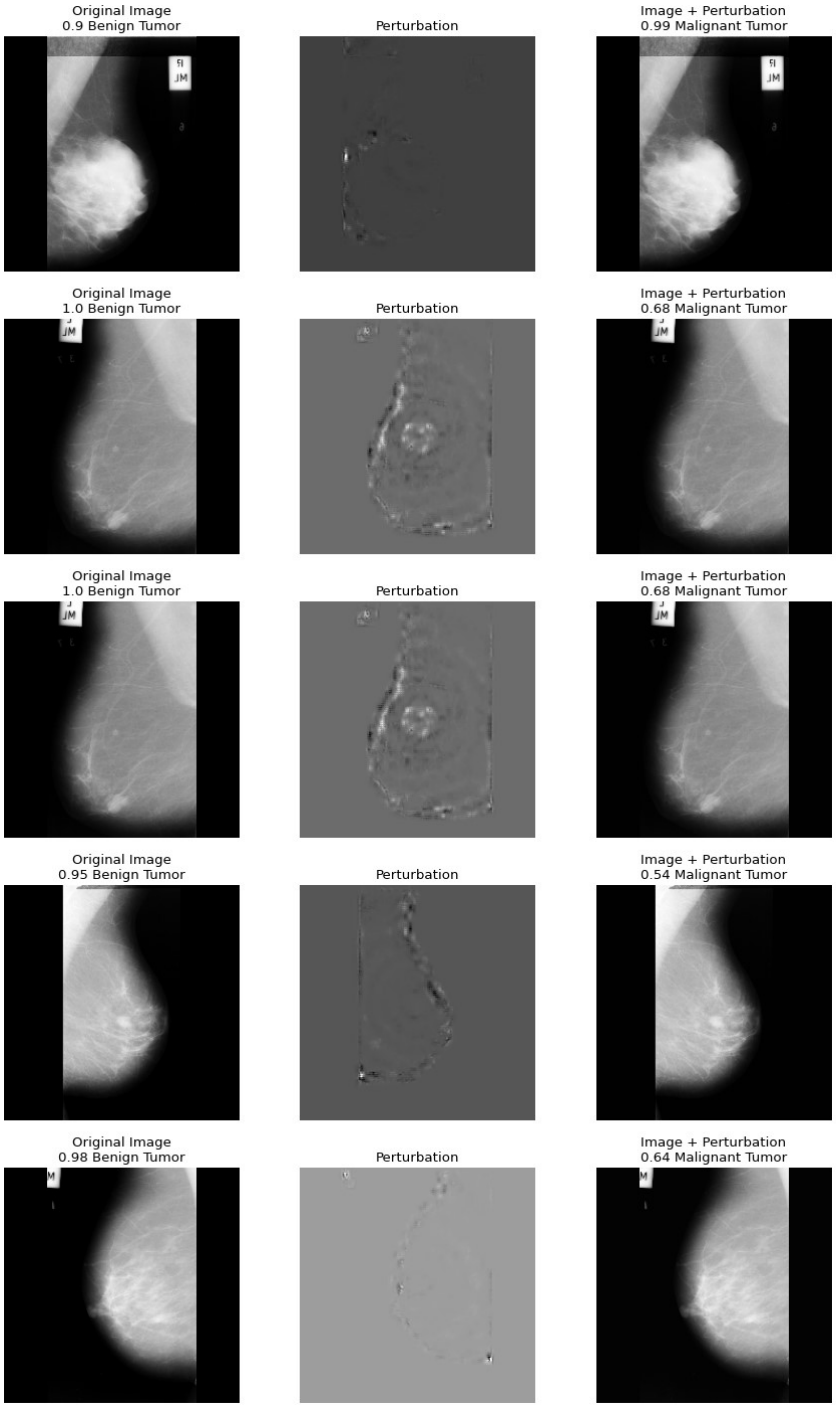



**Figure 3.** A sample of perturbation caused by PGD and misclassification result.

Table 2 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 a 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. Therefore, for a patient submitting a false claim, 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 the 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.

We also report another attack on medical images, the so-called universal attack, which comes directly from the research of (Moosavi-Dezfooli et al. 2017). Figure 4 shows the results of the universal attack on the breast cancer images. Not only is a benign tumor detected as malignant cancer, as with the PGD attacks, but in some cases, the adversarial images are even diagnosed with higher accuracy than the original image.

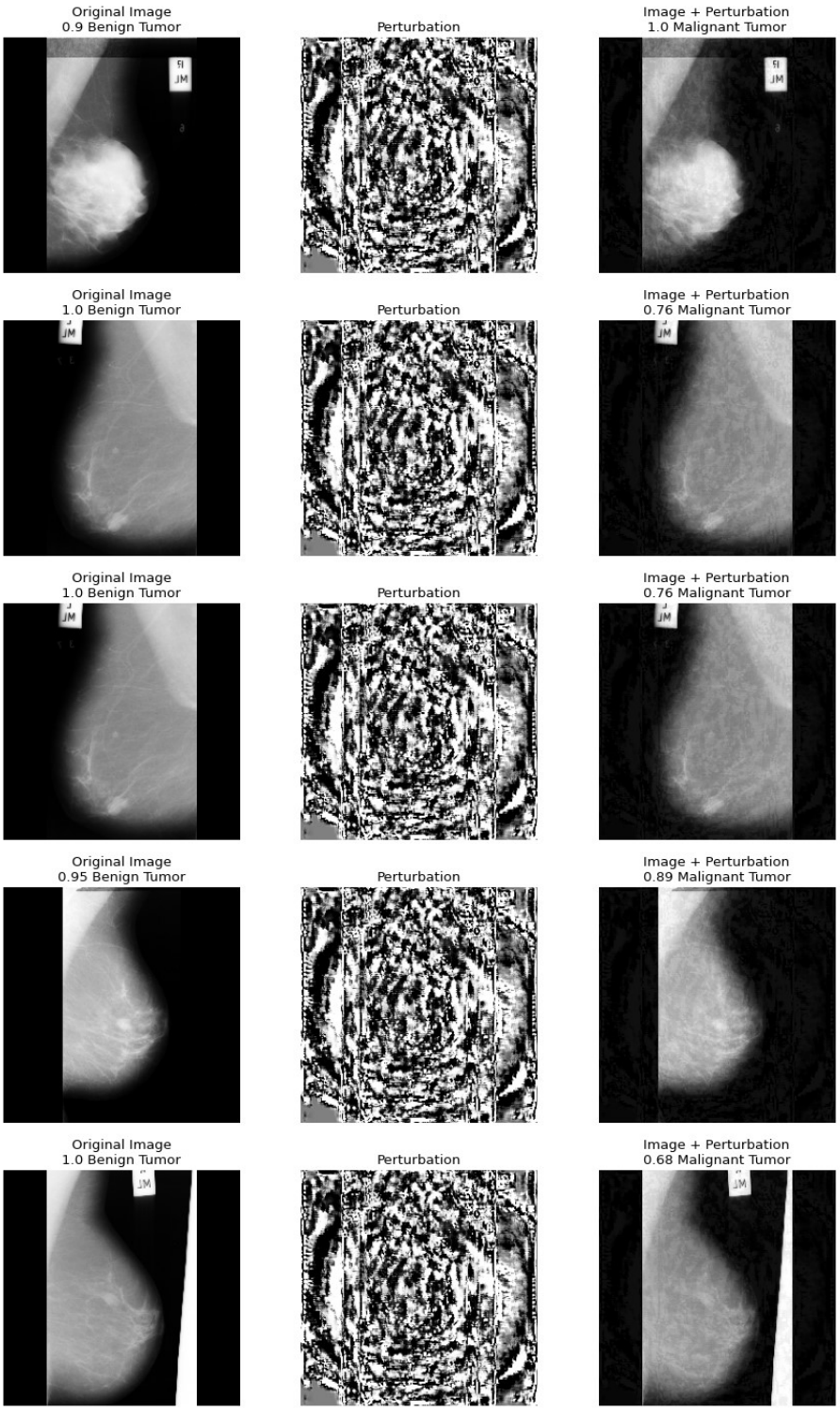



**Figure 4.** A sample of perturbation caused by universal attack and misclassification result.

**Table 2.** The perturbation impact on the accuracy level of the target model.

| Perturbation $\epsilon$ | Model Accuracy | Perturbation $\epsilon$ | 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 |

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 4).

Overall, the results of the experiments show that the model can be fooled for detecting the type of patients' cancer. Based on the above case studies, healthy individuals may submit fake claims.

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 with even higher accuracy than the pre-attacked image. An 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.

## 6. Preparing for Adversarial Attacks

A number of papers and books suggest how to make AI systems more robust to adversarial attacks. Given that writing an attack to an AI application may also be quite inexpensive (we built the mammogram attack on our own) and can provide the attacker with great benefits, such as being refunded for a malign rather than a benign tumor, attacks are easy to receive. Symmetrically, the cost of suffering them is much bigger than the cost of producing them. Defense against them is cost-efficient as long as it is inferior to the magnitude (severity times frequency) of the attacks, which is the result of a case-by-case evaluation. In the fraud case, it is akin to fraud entity evaluation.

The 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 the training of the whole AI approach. So-called certified defenses not only provide 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 et al. (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, autoencoder-based input denoising, and feature denoising. Among the provable defenses, they distinguish the semidefinite programming-based certificated defense, distributional robustness certification, weight-sparse DNNs, dual approach-based provable defense, KNN-based defenses, and Bayesian model and consistency-based defenses. As for DNNs, which we applied above, the book by Ware (2019) explains how to make applications of DNNs to image processing more resilient. Xu et al. (2020) extend the analysis to the DNNs applied to graphs and text.

It is obvious from the rich taxonomy above that the artillery at one's disposal to prevent adversarial attacks is rich. Sometimes the defense methods are also quite powerful. Kurakin et al. (2017b), for instance, show, using ImageNet, that adversarially trained models perform better on adversarial examples than on non-attacked ones, as 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, like any AI model.

Most of the current literature, however, focuses on specific attacks and on how to strengthen 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 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 such as 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 detecting where attacks can hurt more so as to prioritize the search for resiliency.

The first countermeasure one can take in 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 the model results used in impact assessments and elsewhere depend upon the information fed into them, their structure, and their 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.[3]

Given the wide experience that actuaries, risk managers, and tariff producers have of the data and their 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.

## 7. Conclusions

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[4]. This recourse has been quite pervasive in the US, because there the total cost of insurance fraud (excluding health insurance) is estimated to exceed USD 40 billion per year, which means an increase in premiums of USD 400 and USD 700 per year and per household. Europe is following, with European regulators paying a lot of attention to AI. In the era of AI, the defenses against adversarial AI could save a lot of money.[5] How to do this with a holistic, sensitivity-based approach as a first, universal defense, together with a balance between autonomy and human oversight in AI applications, is still an open issue. The deployment of AI in the insurance ecosystem requires the enhancement of analytic maturity. For a high-level review of the role of data science and data scientists in modern organizations, with a stress on this maturity, see Kenett and Redman (2019). For a technical introduction to AI and statistical analysis with Python, which helps in developing attacks and counterattacks such as those in our example, see Kenett et al. (2022). Enhancing analytic maturity is both a management and a technical challenge. Addressing adversarial AI in insurance companies is an example of such a challenge. We cannot provide a one-size-fits-all monitoring solution in this paper, and the prevention list cannot go beyond the lines of Section 6. However, following the suggestions of the EIOPA, we consider higher attention to adversarial attacks on AI and analytic maturity an enhancement of the current situation of insurers and, consequently, a reassuring solution for consumers.

**Author Contributions:** Conceptualization, B.A., E.L. and R.S.K.; methodology, B.A. and R.S.K.; software, B.A.; validation, E.L. and M.C.; formal analysis, B.A. and R.S.K.; investigation, All; writing—original draft preparation, B.A. and R.S.K.; writing—review and editing, All; visualization, B.A. and E.L.; supervision, E.L.; project administration, B.A.; funding acquisition, E.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** Data from open source was used in analysis.

**Conflicts of Interest:** Authors declare no conflict of interest.

## Appendix A

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.

### *Appendix A.1. White-Box Attacks*

To theoretically explain the adversarial attack of group "a", let the input domain be $X \in R^d$ and the class domain be $Y \in \{0,1\}^C$, and let $H(x) : X \to Y$ be a functional mapping of 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 $\theta$ 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:

$$minimize\ ||x_{adv} - x||_p\ s.t.\ H(x_{adv}) \neq y \tag{A1}$$

where $||.||_p$ is the norm that measures the extent of perturbation. Some commonly used $L_p$ norms are $L_0$, $L_2$, or $L_\infty$. 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:

$$minimize\ ||x_{adv} - x||_p\ s.t.\ H(x_{adv}) = y_{target} \neq y \tag{A2}$$

In general, targeted adversarial examples are more difficult to generate than untargeted adversarial examples. Different ways to solve both (A1) and (A2) 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 $\theta$ 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.

(1) Fast Gradient Sign Method
(2) Projected Gradient Descent
(3) DeepFool
(4) Carlini and Wagner

#### Appendix A.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 $y$) and a single-gradient ascent step in the case of untargeted attack.

$$x_{adv} = x - \epsilon.sign\ (\nabla_x\ J(\theta,\ x,\ t)) \tag{A3}$$

$$x_{adv} = x + \epsilon.sign\ (\nabla_x\ J(\theta,\ x,\ y)) \tag{A4}$$

FGM is a fast method that only perturbs the input once, with an $\epsilon$—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 $\epsilon$, although this may result in large perturbations that are perceptible to human observers.

Appendix A.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 (A5), $clip_{x,\epsilon}(.)$ 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

$$x^0_{adv} = x,\ x^i_{adv} = clip_{x,\epsilon}\left(x^{i-1}_{adv} + \epsilon sign\left(\nabla_{x^{i-1}_{adv}}\ J\left(\theta,\ x^{i-1}_{adv},\ y\right)\right)\right) \tag{A5}$$

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 et al. 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.

Appendix A.1.3. DeepFool

The DeepFool algorithm (Moosavi-Dezfooli et al. 2016) was developed with the goal of providing an efficient yet accurate method for computing minimal perturbations with respect to the $l^p-$norm. As 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).

Appendix A.1.4. Carlini and Wagner Attack (C and W)

C and W's attack (Carlini and Wagner 2017) attempts to find the minimally biased perturbation problem—in a similar manner to the DeepFool algorithm—as follows:

$$min\ ||x - x'||^2_2 + c.H(x', t),\ s.t\ x' \in [0,\ 1]^m \tag{A6}$$

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 a superior performance (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 the 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 and 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.

*Appendix A.2. Black-Box Attacks*

While the definition of a "white-box" attack on DNNs is clear and precise, i.e., it provides 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, the attacks 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

Appendix A.2.1. Substitute Model

The paper by (Papernot et al. 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, and human faces) and (2) the classifier's architecture (e.g., CNN and DNN).

The authors in (Zhang et al. 2021) exploit the "transferability" property (defined in Section 4.3 above) of adversarial examples: an example $x'$ can attack $H_1$; it is also likely to attack $H_2$, which has a 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:

(a). Synthesize a substitute training dataset: create a "replica" training set. For example, to attack handwritten digit recognition, create an initial substitute training set $X$ by: (a) requiring samples from the test dataset or (b) creating handcrafting samples.
(b). 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.
(c). 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 $H'$.
(d). 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.

Appendix A.2.2. Gradient Estimation

Another approach for black-box attacks is the gradient estimation method ZOO, proposed by (Chen et al. 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 (A7) shows that for each index $i$ of sample $x$, we add (or subtract) to an $\epsilon$ multiple of another vector $e_{i,}$ to obtain $x_i = x \pm \epsilon e_i$. If $\epsilon$ is small enough, we can extract the gradient information for $H(.)$ by

$$\frac{\partial H(x)}{\partial x_i} \approx \frac{H(x + \epsilon e_i) - H(x - \epsilon e_i)}{2\epsilon} \tag{A7}$$

*Appendix A.3. Universal Attack*

The 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 by (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 A. The goal of a universal perturbation is to make the classifier classify the perturbed image differently from what it should, on at least a percentage of $1 - \delta$ of cases. Let $H()$ be the classifier, $\eta$ be the adversarial perturbation, and $P$ denote the probability. The universal goal is

$$P(H(x + \eta) \neq H(x)) \geq 1 - \delta \tag{A8}$$

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

$$||\eta||_p \leq \epsilon \tag{A9}$$

In the constraint, the $p-$norm is required to be smaller than a constant $\epsilon$.

## Notes

1 https://www.capgemini.com/ca-en/wp-content/uploads/sites/10/2017/07/Fraud_Detection_in_Healthcare_from_Capgemini_and_Palantir.pdf (accessed on 1 December 2022).

2 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.

3 https://ec.europa.eu/search/?QueryText=sensitivity+audit&op=Search&swlang=en&form_build_id=form-WZ65edbU064IlfvfZtaOeEFLhj5IOLUbYfEFLNJ707Q&form_id=nexteuropa_europa_search_search_form (accessed on 5 June 2021).

4 See, for instance, https://www.pinow.com/articles/305/insurers-on-the-alert-for-false-claims-turn-to-private-investigators, (accessed on 10 December 2022).

5 See https://www.fbi.gov/stats-services/publications/insurance-fraud (accessed on 5 June 2021).

## References

Ai, Jing, Patrick L. Brockett, Linda L. Golden, and Montserrat Guillén. 2013. A robust unsupervised method forfraud rate estimation. *Journal of Risk and Insurance* 80: 121–43. [CrossRef]

Artís, Manuel, Mercedes Ayuso, and Montserrat Guillén. 1999. Modelling different types of automobile insurance fraud behavior in the Spanish market. *Insurance: Mathematics and Economics* 24: 67–81.

Artís, Manuel, Mercedes Ayuso, and Montserrat Guillén. 2002. Detection of automobile insurance fraud with discrete choice models and misclassified claims. *The Journal of Risk and Insurance* 69: 325–40. [CrossRef]

Brockett, Patrick L., Richard A. Derrig, Linda L. Golden, Arnold Levine, and Mark Alpert. 2002. Fraud classification using principal component analysis of RIDITs. *Journal of Risk and Insurance* 69: 341–71. [CrossRef]

Byra, Michał, Grzegorz Styczynski, Cezary Szmigielski, Piotr Kalinowski, Lukasz Michalowski, Rafal Paluszkiewicz, Bogna Wróblewska, Krzysztof Zieniewicz, and Andrzej Nowicki. 2020. Adversarial attacks on deep learning models for fatty liver disease classification by modification of ultrasound image reconstruction method. *IEEE International Ultrasonics Symposium (IUS)*, 1–4. [CrossRef]

Carlini, Nicholas, and David Wagner. 2017. Towards Evaluating the Robustness of Neural Networks. Paper presented at IEEE Symposium on Security and Privacy, San Jose, CA, USA, May 22–24; pp. 39–57. [CrossRef]

Caron, Louis, and Georges Dionne. 1999. *Insurance Fraud Estimation: More Evidence from the Quebec Automobile Insurance Industry. Automobile Insurance: Road Safety, New Drivers, Risks, Insurance Fraud and Regulation. Huebner International Series on Risk, Insurance, and Economic Security*. Boston: Springer, vol. 20.

Chen, Pin-Yu, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. 2017. ZOO: Zeroth Order Optimization based Black-Box Attacks to Deep Neural Networks without Training Substitute Models. Paper presented at 10th ACM Workshop on Artificial Intelligence and Security, Dallas, TX, USA, November 3; pp. 15–26. [CrossRef]

Dionne, Georges, Florence Giuliano, and Pierre Picard. 2008. Optimal Auditing with Scoring: Theory and Application to Insurance Fraud. *Management Science* 55: 58–70. [CrossRef]

Finlayson, Samuel G., Hyung Won Chung, Isaac S. Kohane, and Andrew L. Beam. 2019. Adversarial Attacks against Medical Deep Learning Systems. *Science* 363: 1287–89. [CrossRef] [PubMed]

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. Paper presented at Conference ICLR, San Diego, CA, USA, May 7–9.

Grize, Yves-Laurent, Wolfram Fischer, and Christian Lützelschwab. 2020. Machine learning applications in nonlife insurance. *Applied Stochastic Models in Business and Industry* 36: 523–37. [CrossRef]

Hirano, Hokuto, Akinori Minagi, and Kazuhiro Takemoto. 2021. Universal adversarial attacks on deep neural networks for medical image classification. *BMC Medicial Imaging* 21: 1–13. [CrossRef]

Joel, Marina Z., Sachin Umrao, Enoch Chang, Rachel Choi, Daniel Yang, James Duncan, Antonio Omuro, Roy Herbst, Harlan M. Krumholz, and Sanjay Aneja. 2021. Adversarial Attack Vulnerability of Deep Learning Models for Oncologic Images. *medRxiv*. [CrossRef]

Kenett, Ron S., and Thomas C. Redman. 2019. *The Real Work of Data Science: Turning Data into Information, Better Decisions, and Stronger Organizations*. Hoboken: John Wiley & Sons.

Kenett, Ron S., Shelemyahu Zacks, and Peter Gedeck. 2022. *Modern Statistics: A Computer-Based Approach with Python*. Berlin/Heidelberg: Springer Nature.

Kooi, Thijs, Geert Litjens, Bram Van Ginneken, Albert Gubern-Mérida, Clara I. Sánchez, Ritse Mann, Ard den Heeten, and Nico Karssemeijer. 2017. Large scale deep learning for computer aided detection of mammographic lesions. *Medical Image Analysis* 35: 303–12. [CrossRef] [PubMed]

Kurakin, Alexey, Ian J. Goodfellow, and Samy Bengio. 2017a. Adversarial Examples in the Physical World. Paper presented at ICLR, Toulon, France, April 24–26; p. 14.

Kurakin, Alexey, Ian J. Goodfellow, and Samy Bengio. 2017b. Adversarial Machine Learning at Scale. Paper presented at ICLR, Toulon, France, April 24–26.

Li, Bin, Yunhao Ge, Yanzheng Zhao, Enguang Guan, and Weixin Yan. 2018. Benign and malignant mammographic image classification based on Convolutional Neural Networks. Paper presented at 2018 10th International Conference on Machine Learning and Computing, Macau, China, February 26–28.

Mirsky, Yisroel, Tom Mahler, Ilan Shelef, and Yuval Elovici. 2019. CT-GAN: Malicious Tampering of 3D Medical Imagery using Deep Learning. Paper presented at 28th USENIX Security Symposium, Santa Clara, CA, USA, August 14–16.

Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi, and Pascal Frossard. 2016. A Simple and Accurate Method to Fool Deep Neural Networks. Paper presented at IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, June 27–30; pp. 2574–82. [CrossRef]

Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2017. Universal Adversarial Perturbations. Paper presented at IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, July 21–26.

Papernot, Nicolas, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. 2017. A Practical Black-Box Attacks against Machine Learning. Paper presented at 2017 ACM Asia Conference on Computer and Communications Security, Abu Dhabi, United Arab Emirates, April 2–6; pp. 506–19. [CrossRef]

Qiu, Shilin, Qihe Liu, Shijie Zhou, and Chunjiang Wu. 2019. Review of Artificial Intelligence Adversarial Attack and Defense Technologies. *Applied Science* 9: 909. [CrossRef]

Ren, Kui, Tianhang Zheng, Zhan Qin, and Xue Liu. 2020. Adversarial Attacks and Defenses in Deep Learning. *Engineering* 3: 346–60. [CrossRef]

Sadeghi, Somayeh, Sajjad Dadkhah, Hamid A. Jalab, Giuseppe Mazzola, and Diaa Uliyan. 2018. State of the Art in Passive Digital Image Forgery Detection: Copy-Move Image Forgery. *Pattern Analysis and Applications* 21: 291–306. [CrossRef]

Singh, Amit Kumar, Basant Kumar, Ghanshyam Singh, and Anand Mohan. 2017. *Medical Image Watermarking Techniques: A Technical Survey and Potential Challenges*. Cham: Springer International Publishing, pp. 13–41. [CrossRef]

Suckling, John. 1996. The Mammographic Image Analysis Society Digital Mammogram Database. Available online: https://www.kaggle.com/datasets/tommyngx/mias2015 (accessed on 10 February 2019).

Ware, Colin. 2019. *Information Visualization: Perception for Design*. Burlington: Morgan Kaufmann. ISBN 9780128128756.

Wetstein, Suzanne C., Cristina González-Gonzalo, Gerda Bortsova, Bart Liefers, Florian Dubost, Ioannis Katramados, Laurens Hogeweg, Bram van Ginneken, Josien P. W. Pluim, Mitko Veta, and et al. 2020. Adversarial Attack Vulnerability of Medical Image Analysis Systems: Unexplored Factors. *Medical Image Analysis* 73: 102141.

Xu, Han, Yao Ma, Haochen Liu, Debayan Deb, Hui Liu, Jiliang Tang, and Anil K. Jain. 2020. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing* 17: 151–78.

Zhang, Jiliang, and Chen Li. 2018. Adversarial Examples: Opportunities and Challenges. *IEEE Transactions on Neural Networks and Learning Systems* 16: 2578–93. [CrossRef] [PubMed]

Zhang, Chaoning, Philipp Benz, Chenguo Lin, Adil Karjauv, Jing Wu, and In So Kweon. 2021. A Survey On Universal Adversarial Attack. Paper presented at Thirtieth International Joint Conference on Artificial Intelligence, Montreal, QC, Canada, August 19–26.