ISSN 2279-9362

AI and Adversarial AI in insurance: Background, examples, and future implications

Behnaz Amerirad, Matteo Cattaneo, Ron Kenett, Elisa Luciano

No. 667
December 2021

Carlo Alberto Notebooks

www.carloalberto.org/research/working-papers

© 2021 by Behnaz Amerirad, Matteo Cattaneo, Ron Kenett, Elisa Luciano. Any opinions expressed here are those of the authors and not those of the Collegio Carlo Alberto.

AI and Adversarial AI in insurance:
Background, examples, and future implications

December 16, 2021

Behnaz Amerirad, McMaster University, Canada

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

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

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

Abstract

This paper describes the rapid and dynamic pace of Artificial Intelligence (AI) and Machine Learning (ML) developments that have revolutionized the insurance sector, and discusses their interpretability. AI content-based processing of information includes image and video analysis, facial recognition and automated decision making for claims management and fraud detection.

The paper focuses on adversarial AI, namely on the creation of input data slightly altered to mislead a machine learning system and make it produce incorrect predictions.

It provides a case study of the impact of adversarial AI in health insurance. Not only can the model be fooled for detecting the malignancy of patients, but a higher level of perturbation can increase the success rate of the attack by lowering the accuracy of the system. We deem important, for insurance companies ready to adopt AI technologies, to be aware of the consequences of adversarial attacks.

We conclude with policy recommendations, consistent with the current regulatory framework.

---

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

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

1. AI in insurance

The rapid and dynamic pace of Artificial Intelligence (AI) and Machine Learning (ML) developments have revolutionized the insurance sector. The EIOPA's Big Data Analytics thematic review in motor and health insurance, reported in 2018, that 31% of the participating European insurance firms were using AI and another 24% were at a "proof of concept" stage. (EIOPA, 2021).

AI applications in health insurance in conjunction with the COVID-19 pandemic is having a massive impact on the healthcare systems around the globe. Moreover, AI systems exhibit can potentially outperform humans in a wide range of insurance disciplines such as content-based processing of information, text analytics, and recognition of patterns, trends, and preferences (Eling, Nuessle, and Staubli, 2021).

AI content-based processing of information includes image and video analysis, facial recognition and automated decision making. For example, derivation and automatic application of rules can improve claim management across the world. Health, property, and casualty insurance claim applications use AI's image-based technology which makes them less complex and time-consuming. We provide below some examples.

E-claim is a process which provides an intuitive online portal to evaluate claim management by software solutions that nowadays have deployed by some insurance companies (e.g. TELUS Health (Startup)). In other words, analyzing and interpretation of uploaded applicant's images through automated AI facilitate the operations of insurance company (e.g., Lapetus, IBM Watson) to process high volumes of documents to extract relevant information in their claims management. For example, medical images and hospital bills can be electronically submitted by patients, and they receive immediate notification that their images or bills have been received, approved, and credited to their accounts (GovInsider).

In property and casualty insurance, automated AI image-recognition technology provide services such as damage detection and cost estimation processes which can save time compared to manual assessment. For instance, video and image analysis from satellite pictures can either provide early warning systems for natural catastrophes (e.g., Swiss Re) or risk monitoring/management after extreme events such as tornadoes, hurricanes, and wildfires. If a satellite image of a tornado track/flood zone is available, image recognition models can help identify the damaged areas as well as the degree of damage. A well-trained image recognition model can identify the tornado/flood path and determine whether insured homes are damaged and, if so, how severe that damage is.

Moreover, the Insurance Industry can benefit highly from the application of text analytics and natural language processing plus sentiment detection for pricing and customization of insurance

products. AI enables insurance companies (e.g., Lemonade, Allianz, PNB MetLife, AXA, Aetna, Geico (Startup)) to engage with consumers through their written or verbal requests faster and more consistently. Insurer can deploy virtual customer service agents in contact centers or use Robo-Advisors to understand, categorize and interpret written text and convert it into computer-readable structured data sets for consistent and rules-based advice at an affordable cost. Conversational agents, also called chatbots, can recommend and personalize products, handle complaints, improve communication with customers, and handle simple transactions (Eling et al., 2021). Besides, sentiment analytics as a powerful tool for understanding customer feeling can be employed to identify the polarity of specific aspects or general industry trends, analyze the brand perception, and gain information regarding reputational crises beforehand.

Lastly, but not least, the recognition of patterns and anomalies is extensively used in recommendation engines and fraud detection in the insurance sector. Recommending insurance products for customers accurately and efficiently can help to improve the competitiveness of insurance company. Automated fraud detection is another implementation of pattern recognition which helps the insurance company to recognize suspicious claims and minimize unnecessarily claim costs. Increasingly, insurers (e.g., Oscar, Fabric, Aegon, Ping An, AXA, Generali, Allianz (Eling et al., 2021)) are evaluating and implementing ML-based fraud solutions that augment internal data with new sources of information, including the Internet of Things (IoT) and third-party public data (Maull et al., 2019).

However, the AI adoption in the insurance sector, although it is growing rapidly, is in its non-advanced stages in many insurance companies. One of the key challenges that insurance carriers are facing today, within the AI domain, is how to design and execute an AI strategy. The difficulty lies in not only the novelty of the technology but also in the expected impacts and changes that the adoption of AI should have.

Beside the first moves towards the introduction of AI in insurance company, they often aim at initiatives that create positive business impact, while, progressively, building up all required foundations for a true AI-driven company (operating model, governance model, tools, etc.). Within the foundation of this topic and of increasing relevance and importance, it involves generally to AI's security and particularly to Adversarial Artificial Intelligence (AAI) aspects. The reason behind the importance of this topic is connected to two major factors:

- Its difference vis-à-vis the traditional cyber risk.
- The potential negative impacts that can be registered in many areas of an insurance company: underwriting, sales, claims, etc.

Despite of some insurance companies' efficient movement toward automation by application predictive models of ML systems in exclusive domain of human underwriters, actuaries and financial professionals, non-affirmative cyber exposures remain a source of concern.

To illustrate the opportunities and risk of AI, we concentrate on insurance fraud. In Section 2, we start by providing a quick review of AI application for insurance fraud. Section 3 presents the adversarial attack concept and how these attacks raise concerns about the extensive use of AI methodologies for the interaction between insurer and insured. Section 4 is a review of adversarial attacks in the underwriting process of health care insurance. Possible insurance fraud scenarios are reviewed. Section 5 presents a case study, which illustrates the powerfulness of adversarial attacks in this insurance type. Section 6 discusses possible policy interventions, faced to the AI and adversarial AI features. The paper ends contain two appendices which discuss the adversarial taxonomy and adversarial generation examples.

2. Fraud Detection in Insurance

Over recent years, insurers in Europe have experienced fraud on an increasing scale in their claims processing. Insurance Europe, the European (re)insurance federation, estimates that the total from all cases of fraud – both detected and undetected – amounts to 10 percent of overall claims expenditure (Europe, 2013). By its definition, insurance fraud refers to situations when the insureds submit illegitimate claims to the insurance company with the intent to obtain an improper payment from an insurer. This can be classified under the categories of soft and hard fraud, as describe below (Whitaker, 2019):

- Hard fraud: It is referred to a situation when a perpetrator plans to commit an act intentionally that leads to loss. This occurs when people unlawfully obtain money from insurance companies by a reporting a false injury or accident.

- Soft fraud: It is referred to a situation when a claimant overstates a legitimate claim to try and get a larger payout. This happens when people either exaggerate in the report to their insurance companies or hide certain information for financial gain.

The insurance fraud detection is important since it has been characterized as a "crime without victims". Extra costs due to fraud, if not detected and recovered, are passed to a large extent onto all policyholders by an increased in premium for future insured (Zanghieri, 2017). This might happen in different field of insurance (Sithic and Balasubramanian, 2013) including home insurance, life insurance, motor insurance, and medical insurance, which according to the survey of Insurance Fraud Detection, among them the last two categories are more common.

In detection of insurance fraud, a wide variety of statistical and data mining-based approaches have been proposed that can be categorized as

- Supervised learning: This learning category requires a labeled dataset or target variables which disclose whether the requested claim is fraud or not. K- Nearest Neighbor (KNN), Naive Bayes, Decision Trees (DT), Support Vector Machine (SVM), Neural Network (NN), Regression are some popular supervised algorithms. In (Sundarkumar and Ravi, 2015)

and (Li, Yan, Liu, and Li, 2016), for instance, different supervised learning algorithms have been used to deal with automobile insurance fraud detection.

- Unsupervised learning: This learning algorithms deal with the data where the target variable or the labeled data is not available. Clustering, Association rules, Principal Component Analysis (PCA) are commonly used in unsupervised learning. For instance, authors in (Tsoi, Zhang, and Hagenbuchner, 2005) and (Verma, Taneja, and Arora, 2017) have proposed pattern recognition to assess fraudulent insurance in health care systems.

- Hybrid Method: In this method, a combination of supervised and unsupervised methods is used to overcome the flaw of one learner by the other one. Data mining as a hybrid method is used in (Rawte and Anuradha, 2015) and (Rizk, Elgokhy, and Sarhan, 2015) for fraud detection claims in healthcare insurance sector.

- Ensemble Learners: The main idea of this learner is to improve the prediction performance that integrating various homogeneous or heterogeneous learners together to outperform the model than that of single classifier. As an example, ensemble combining classification is applied in (Xu, Wang, Zhang, and Yang, 2011) and (Kamil, Hassan, and Abraham, 2016) to detect insurance fraud in automobile claim datasets.

Although the previous studies develop algorithms to classify the fraudulent insurance claims through different ML methods, in this research, we have mainly focused on the vulnerability of Deep Learning (DL) caused by adversarial attacks in insurance sector. These attacks, as a form of hacking to ML systems, are developed to challenge the accuracy neural network in the realms of computer vision and in traditional predictive modelling. According to the fraud definition mentioned earlier in this section, adversarial attacks have potential capability to threaten the insurance sector and falsify its submitted documents to make an illegitimate profit for the attacker.

So far, this aspect of fraud is not well developed for several reasons, including the lack of privacy protection, ethical considerations, and regulatory compliance issues that require considerations from an integrated perspective of law and economics.

3. Interpretability in Artificial Intelligence

AI technology involves algorithmic tools to learn from numerical data, images, sounds, texts, etc. However, AI algorithms can provide predictions which turn out to be difficult to explain or interpret. From a general point of view, allowing an understandable explanation for any AI model output helps decision makers understand the underlying “reasoning” behind the AI results. Enhancing interpretability is making the debugging process easier and helps in model improvement and acceptability of the AI tool.

There is a lack of consensus about definitions of explainability and interpretability of AI models. Indeed, these notions refer to cognitive processes related to social sciences and their different fields of applications. Some authors also invoke other fundamental concepts (see, e.g., completeness, fairness, intelligibility, comprehensibility, transparency) to build a proper definition of what “explainable AI” is and what it is intended for (Gilpin et al., 2018); (Arrietaa et al., 2020). In this section, we focus on AI interpretability as the ability of an AI model (or any element related to this model, i.e., inputs, outputs, predictions) to be associated with concepts held by a human. with this view, interpretability is linked to the ability to understand representations of the AI model such as predictions and associated decisions.

Molnar (Molnar, 2020) offers a taxonomy of interpretability using criteria that are described below:

A first classification is following the type of results given by the method (e.g., visualization tools or summary statistics) which can be useful for a user-friendliness point of view but does not really help to understand the underlying similarities and differences between the methods.

The literature on the subject often makes a clear distinction between intrinsically interpretable models and model interpretability techniques. In the first case, these models are called transparent models, while in the second case, the techniques are referred to post-hoc interpretability (Arrietaa et al., 2020; Molnar, 2020).

Another dichotomy is model-specific vs. model-agnostic methods (Molnar, 2020). The first ones are specific to a given class of models while the other ones can be used on any AI model which has been trained on data. The idea is to characterize the genericity of the interpretability methods.

Finally, a strong distinction exists between global and local interpretability. On the one hand, a global interpretation aims at explaining the behavior of a model throughout his validity domain (i.e., for all the population of observations), and to identify the features that are the most influential, globally speaking. On the other hand, a local interpretation tries to catch the effect of a specific feature on one specific (or a group) of model predictions (Arrietaa et al., 2020; Longo, Goebel, Lecue, Kieseberg, and Holzinger, 2020; Molnar, 2020) Again, these notions show strong links with similar (but different) notions of “local” and “global” SA techniques (Veiga, Gamboa, Iooss, and Prieur, 2021).

Three different objectives are considered in the context of AI interpretability:

1. Measures of importance of explanatory variables, which consist in measuring the impact on the outputs of the inputs/features. It corresponds to the factor prioritization setting in SA. One can be interested in two types of impact. The first one consists in explaining how a feature or a set of features impact the output distribution. The second one

consists in explaining how the features impact the model's output for one particular instance.

2. Features identification. The objective of this setting is to find the links between the features and the targeted label, via the use of statistical techniques such as correlation measures (Pearson coefficient, Spearman's rank correlation coefficient, Kendall's tau, copula, Hoeffding's D) and kernel-based metrics. Even if their objectives are different, it corresponds to the factor fixing setting in SA (because it mainly concerns screening techniques).
3. Robustness of the decision boundary. This last setting concerns two topics: (a) The counterfactual examples (i.e., explaining the prediction related to one individual by another close individual with an opposite target prediction); (b) How the output label changes when the distribution of one input or group of inputs change (perturbations of the data distribution)?

Various methods have been proposed to help users interpret the predictions of complex models. Lundberg and Lee introduce SHAP (SHapley Additive exPlanations), a unified framework for interpreting predictions. SHAP assigns to each feature in the model an importance value for a particular prediction (Lundberg and Lee, 2017). It includes: (1) the identification of additive feature importance measures, and (2) theoretical results showing there is a unique SHAP solution with a set of desirable properties.

Local interpretable model-agnostic explanations (LIME) were proposed in (Ribeiro, Singh, and Guestrin, 2016) as a local surrogate interpretable model used to explain individual predictions of black box machine learning models. Surrogate models are trained to approximate the predictions of the underlying black box model. LIME focuses on training local surrogate models to explain individual predictions. First, LIME ignores the training data and uses the black box model to get predictions of the model. The goal is to understand why the machine learning model made a certain prediction. Then LIME tests what happens to the predictions for input data (instances) to the specific machine learning model. LIME generates a new dataset consisting of perturbed samples and the corresponding predictions of the black box model. On this new dataset, LIME trains an interpretable model, which is weighted by the proximity of the sampled instances to the instance of interest. The learned model is designed be a good approximation of the machine learning model predictions locally but does not have to be a good global approximation. Mathematically, local surrogate models with interpretability constraint, can be expressed as:

$$explanation(x) = \underset{g \in G}{\operatorname{argmin}} L(f, g, x) + \Omega(g)$$

The explanation model for instance x is the model g (e.g., linear regression model) that minimizes loss L (e.g., mean squared error), which measures how close the explanation is to

the prediction of the original model *f* (e.g., an XGBoost model), while the model complexity, $\Omega(g)$, is kept low (e.g., prefer fewer features). G is the family of possible explanations, for example all possible linear regression models. The proximity measure, πx, defines how large the neighborhood around instance x is that we consider for the explanation.

In statistical inference about a population parameter, from a sample, statistical generalizability and sampling bias are the focus. A key question of interest is "What population does the sample represent?" (Rao, 1985). In contrast, for predicting the values of new observations, the question is whether the analysis captures associations in the training data (i.e., the data used in model building) that generalize to the to-be-predicted situations, or out of sample conditions. Control charts present a good example. Assuming the process remains stable, we expect performance to vary within the upper and lower control limits (Kenett, Zacks, and Amberti, 2021).

Statistical generalizability is commonly evaluated using measures of sampling bias and goodness of fit. In contrast, scientific generalizability, used for predicting new observations, is often evaluated by the accuracy of prediction of a hold-out set from the to-be-predicted population. This assessment is a crucial protection against overfitting, which occurs when your model fits previously-collected data perfectly but does very poorly with new data (Sokolic, Giryes, Sapiro, and Rodrigues, 2017). Randomization is a key approach in interventions that enables statistical generalization.

As well as guarding against unknown biases, it provides the mathematical foundations that support calculation and interpretation of p-values. However, clinical trials with randomized allocation of patients to treatment or placebo or A/B testing of web applications may be subject to "sample selection-bias," since participation in a randomized trial cannot be mandated. Sample patients may consist of volunteers who respond to financial and medical incentives, leading to a distribution of outcomes in the study that differ substantially from the distribution of outcomes more generally. This sample selection bias is a major impediment in both the health and social sciences (Hartman, Grieve, Ramsahai, and Sekhon, 2015). It can also affect A/B testing (Kenett et al., 2021).
"Transportability" is another way to generalize. Transportability is defined as a transfer of causal effects learned in experimental studies to a new population, where only observational studies can be conducted (Pearl, 2015). In a study on urban social interactions, Pearl and Bareinboim used transportability to predict results in New York City, based on a study conducted in Los Angeles, accounting for differences in the social landscape between New York City and Los Angeles (Pearl and Bareinboim, 2011, 2014).

Online auction studies provide another example of the importance of generalization. A study of the effect of reserve price on eBay auction final price is reported in (Katkar and Lucking-Reiley, 2006). The authors designed an experiment that produced a representative sample of recorded auctions. Their focus was on statistical generalization of the impact of reserve price on auction

outcomes. In contrast, another study (Wang, Montgomery, and Srinivasan, 2008) forecasts the results of new auctions. In that study, the authors, evaluated predictive accuracy using a hold-out set as opposed to standard errors and sampling bias considered by the first study. A third study, on consumer surplus in eBay, dealt with statistical generalizability by inferring from a sample to all eBay auctions. Because the sample was not drawn randomly from the population, Bapna et al compare their sample with a randomly drawn sample. (Bapna, Jank, and Shmueli, 2008)

Domain-based (or scientific) expertise allows findings from specific data to be applied more generally (Kenett and Shmueli, 2016). Thus, a marketing manager might base his decisions on how to run a marketing campaign in location A, using a market study conducted in location B. He has no data on A, but his experience (soft data) tells him how to adopt the conditions in B to what is required in A. Similarly, a software development manager, in the face of a limited testing budget, might decide to release a version with minimal testing because its functionality is basic and the person who developed it has a good record. In other cases, he might decide to significantly increase the testing effort. Such decisions are often made without formal data analyses. The approach has benefits (i.e., speed), but carries risks that decision makers should bear in mind.

4. Adversarial Attack

An adversarial example is a sample of input data that has been very slightly altered in a way that is intended to mislead a machine learning system (Kurakin, Goodfellow, and Bengio, 2017a). In particular, the terms "adversarial input" and "adversarial example" are used interchangeably to refer to inputs that successfully cause a network to produce incorrect predictions in human point of view. Indeed, what is important in this definition is that an input is created with the intention of causing a system to produce a false result, whether the input succeeds in fooling the network. In general, the term is used only for inputs that have the goal of confusing a network (Warr, 2019).

Figure 1 shows an adversarial image where a pre-trained convolutional neural network has classified the adversarially perturbed "panda" image (right image) as "gibbon" with 99% confidence, even though it looks no different to the human eye than the original "panda" image (left image) (Goodfellow, Shlens, and Szegedy, 2015). In that study, Goodfellow et al. explained the adversarial attack problem by observing that the machine learning system misclassifies examples that differ only slightly from correctly classified examples from the data distribution.

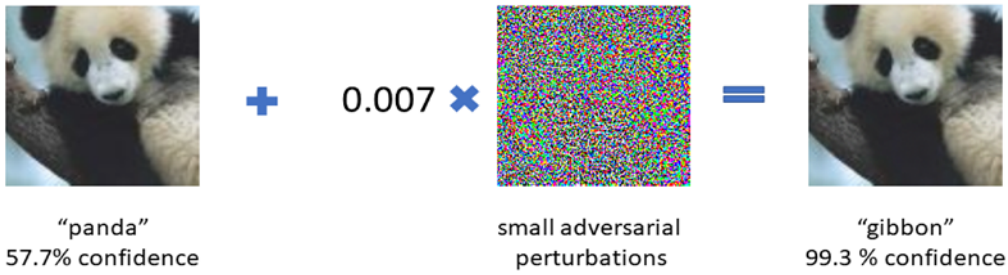


Figure 1 An Example of Adversarial Perturbations and Adversarial Example

After Szegedy (Szegedy et al., 2015) suggested that various machine learning systems are vulnerable to being attacked by adversarial examples, significant research efforts have been made to investigate adversarial attacks in various fields.

An often-cited scenario of an attack from the physical world is an environmental modification made to confuse an autonomous vehicle where decisions regarding steering, response, speed, etc. are based on processing image data captured by the camera. In (Evtimov et al., 2017), the authors showed that an adversarial graffiti on a "stop" sign can mislead a Deep Neural Network (DNN) to recognize it as a "Speed Limit 45" sign, and this could mislead a self-driving car to speed up at a stop sign. So, the vehicle's behavior could change based on interference with road markings, patterns of other vehicles, or road signs.

Another application of adversarial example is a novel algorithm proposed in (Chen, Zhang, Chen, Yi, and Hsieh, 2018). Their applied algorithm can successfully create visually similar adversarial examples for captions with randomly selected captions, as illustrated in Figure 2, that their algorithm is highly transferable to attack other neural caption systems. Meanwhile, Carlini and Wagner produced targeted audio examples of automatic speech-to-text transformation systems as used in "Apple Siri" and "Google Now" (Carlini and Wagner, 2018). In the adversarial speech recognition example, a slight bias - which is almost inaudible - constructs strong, iterative, optimization-based attacks that transform any audio waveform into any target transcription with a high success rate. (See Figure 3).

Much research-including this study-has focused on examining adversarial examples in image and audio data that are continuous, i.e., consist of pixels or audio frequencies with a continuous distribution of values. In contrast, other complex data, such as text, is discrete and does not consist of quantifiable values. In the discrete domain, it can be more difficult to create an adversarial example that goes undetected because it is difficult to quantify a "small" change. Creating adversarial examples for discrete structures such as text is much more challenging, as the semantic and syntactic properties of the original input should be preserved from a human perspective (Lei et al., 2018).

Original Top-3 inferred captions:

1. A red stop sign sitting on the side of a road.
2. A stop sign on the corner of a street.
3. A red stop sign sitting on the side of a street.

Adversarial Top-3 captions:

1. A brown teddy bear laying on top of a bed.
2. A brown teddy bear sitting on top of a bed.
3. A large brown teddy bear laying on top of a bed.

Figure 2 Image Captioning Adversarial Examples Crafted by Show-and-Fool Using the Targeted Caption Method

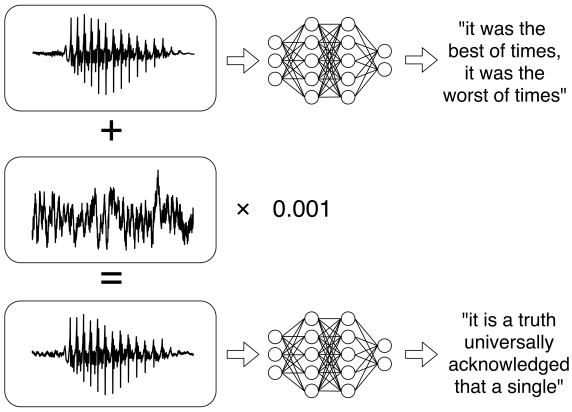


Figure 3 Adversarial Audio- any Audio Waveform Turns Into any Target Transcription with High Rates Success by Only Adding a Slight Distortion

In addition to the scenarios that focus on attacks with either small or imperceptible changes to the clean input, known as image-dependent perturbations, there is another perturbation known as universal patching. In this method, adversarial patches are determined to be added to clean images to generate a universal, robust, targeted adversarial image in the real world (Brown, Mané, Roy, Abadi, and Gilmer, 2017). An attacker could exploit generated adversarial patches in the form of 2D prints or even 3D objects within a scene.

In the presence of the universal patches, an autonomous vehicle can misidentify the stop sign as a speed limit sign and crash into other cars or people (Liab, Zhangab, and Huang, 2021) in contrast to image-based systems of autonomous vehicles misleading by a small perturbation of the input image. Additionally, a wearable object such as adversarial T-shirt (Xu, K. et al., 2020) can bypass the detection of a moving person in surveillance systems. Sharif et al. successfully demonstrate the idea that inconspicuous and physically implemented attacks allow an attacker to evade facial biometric systems or impersonate another individual (Sharif, Bhagavatula, Reiter, and Bauer, 2016) by using adversarial glasses. This leads wearers to confuse face recognition software and subsequently deceive the surveillance and security systems that work with the face recognition method.

While many applications of adversarial attacks in different field of study have already discussed, general taxonomy and generation methods of adversarial examples seem essential to provide a coherent overview of adversarial attacks that Appendices I and II discuss about them, respectively.

5. Impact of Adversarial Attack on Healthcare Insurance

Due to the novelty of the issue regarding adversarial attacks on automated ML systems, a list of attacks that have already been used in medical imaging of healthcare insurance is brought.

Afterward, some potential attacks are proposed based on the current research directions to implement in the breast cancer images. Finally, the impact of these attacks is studied from a fraud insurance point of view.

According to the mentioned definition of fraud, it may occur when the applicants make dishonest claims for medical care by falsifying their medical history to gain a profit. As part of the underwriting process, insurers determine the price of coverage by assessing the involved risk factors based on the applicant's medical history. For this evaluation, insurers are allowed to ask questions about applicants' pre-existing conditions and then decide who to offer coverage to, who to deny coverage to, and whether to charge additional fees for individually purchased coverage. Therefore, health care fraud occurs when at least one of the following conditions is met: (Europe, 2019)

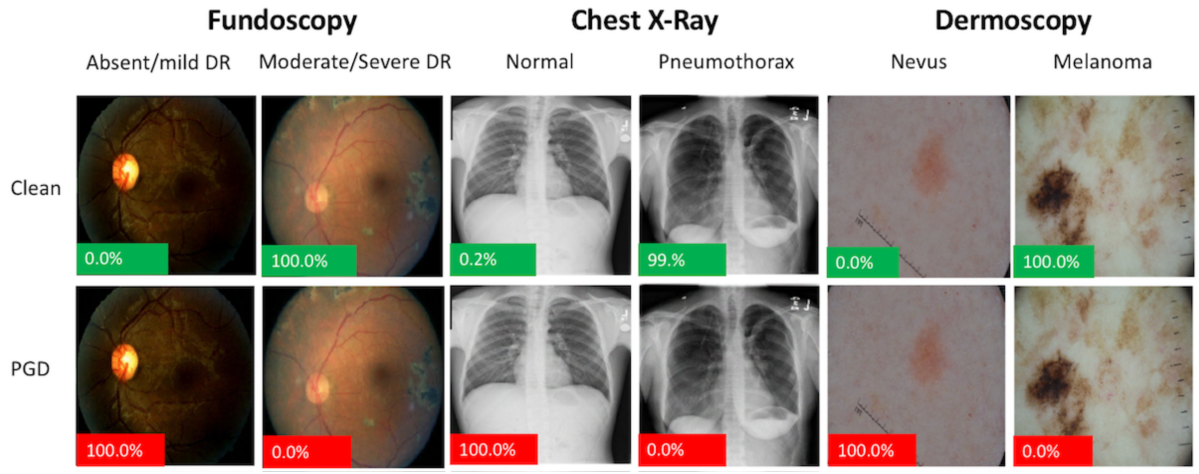
- Providing untrue or incomplete information in insurance applications or answers on an insurance application form.
- Submitting a claim based on misleading or untrue circumstances, including exaggerating a true claim.
- Misleading or untruthfulness in dealing with an insurer with the intention of obtaining a benefit under the insurance contract.

In accordance with the above situation, the purpose of this study is to review the current state of adversarial attacks in medical imaging and how these intentional craft attacks on patients' histories may affect insurance companies.

In healthcare insurance, adversarial attacks occur when an attacker with access to medical imaging material can alter the content to make a misdiagnosis. Specifically, the attacker can add or remove evidence of some 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) (Singh, Kumar, Singh, and Mohan, 2017) (Sadeghi, Dadkhah, Jalab, Mazzola, and Uliyan, 2018). For example, Mirsky manipulated the attacks by injecting and removing pixels on CT scans of patients' lung cancer. (Mirsky, Mahler, Shelef, and Elovici, 2019)

With a precise focus on the attack itself, adversarial examples in medical imaging can alter the behavior of classifiers and segmentation, highlighting the lack of robustness of the neural network. Finlayson et al discussed that the pervasive fraud in healthcare and the need for intelligent algorithms to diagnose the condition of insurance claimants for reimbursement and also for pharmaceutical decisions may create both the incentive and opportunity for a bad actor to perform an adversarial attack (Finlayson, Chung, Kohane, and Beam, 2019). In the same paper, first the model is developed to classify diabetic retinopathy using fundoscopy, pneumothorax using chest X-ray, and melanoma using dermatoscopic images (see Figure 4). Subsequently, the robustness of the model is tested using both PGD and patch attacks that are

imperceptible by humans. While the PGD attacks require digital access to the specific images to be sent into the model, the adversarial patch attacks are universal in the sense that they can be applied to any image.



An insurance company may require confirmation of suspected diagnoses to be reimbursed. It may also use separate methods to ensure that patient identities match previous admissions or have never been submitted by the same provider before.

Another study (Wetstein et al., 2020) evaluates several unexplored factors, including the degree of perturbation and the transmissibility of the adversarial attack, affecting the susceptibility of Deep Learning (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. The results of experiments with both attack methods, FGSM and PGD, at different levels of perturbation show that higher levels of perturbation lead to increased success of the attacks, but also to increased visual perceptibility, which could reduce their effectiveness in MedIA settings in the presence of a medical expert. Furthermore, pre-training MedIA networks on ImageNet can dramatically increase the transmission of adversarial examples; the greater the performance gain achieved through pre-training, the greater the transmission success.

An attack carried out with a conspicuous degree of perturbation could be easily detected by a (trained) person. Therefore, insurance applicants who use these medical images to conceal pre-existing disease (PED), chronic conditions, manipulated results of pre-policy health exams, and falsified documents to meet the terms of policy conditions (Code, 2005) could be neutralized. However, vulnerability assessment of cybersecurity-critical MedIA systems should also consider

attacks that are imperceptible to the human eye and highly transmissible due to their construction.

Universal Adversarial Perturbation (UAP) is a small, image-independent perturbation that can cause Deep Neural Network (DNN) failure in most image classification tasks. Hirano and et.al used UAP in clinical diagnosis for classification of skin cancer, diabetic retinopathy and pneumonia (Wetstein et al., 2020). Their results confirmed that DNNs are susceptible to both non-targeted UAPs that cause task failure, resulting in an input being assigned to an incorrect class, and targeted UAPs that cause the DNN to classify an input into a specific class.

6. A Case Study

To clarify the impact of adversarial attacks on insurance systems, we propose an approach that not only considers the impact of the attack on medical diagnostic systems, but also reviews how the perturbation may affect the insurer's decision to reimburse the treatment procedure.
We discuss countermeasures to increase the robustness of the AI model. It is indeed important, for insurance companies ready to adopt AI technologies, to be aware of the consequences of adversarial attacks and to be ready to make their adoption robust.

Dataset: While the previous section discussed some common examples of AAI application in the health insurance industry, here we explore the possibility of adversarial attacks on insurance claimant history input for breast abnormalities from 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 images (64 benign and 52 malignant) were obtained. In this study, only the abnormal images in the dataset are used to classify the "benign" and "malignant" classes.

This approach in the study can be divided into three main steps: Preprocessing as well as training of the model, adversarial attack on the abnormal images of breast cancer, and countermeasures to increase the robustness of the model.

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 DL, which refers to the process of generating new samples from existing data, used to improve data sparsity and prevent overfitting (Kooi et al., 2017). Transformations include rotations, translations, horizontal and vertical flips, crops, zooms, and

jittering. The main sources of variation in lesion-level mammograms are rotation, scale, translation, and the amount of occluding tissue.

Before we start the training, image preprocessing was performed specifically for the mammography image and the structure of Convolutional Neural Network (CNN) models. Due to the limited number of images in the dataset, 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.

Then, in the classification part of this study, we use a novel CNN model that has been previously proposed to classify abnormalities and benign or malignant tumors. Our method achieves the best accuracy rate, which is 98.3% and 97.0% on the train and test datasets, respectively. The high accuracy as well as other excellent evaluation indicators show that the CNN has a high performance as a key step in our mammography diagnosis.

Figures 5 and 6 provide an overview of the training process. As indicated in Figure 5., there are cases where the loss and accuracy of the validation set are better than their counterparts in the training set. To clarify the reason for this, we should mention that the loss and accuracy are measured following the training of each epoch. As the model improves in the learning process, the malignancy status of cancer is detected more accurately in the validation dataset compared to the training dataset.

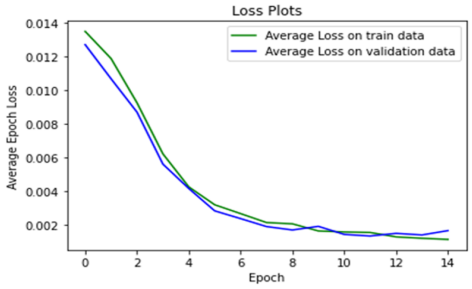


Figure 5. a. Loss Plots for Train and Test Dataset

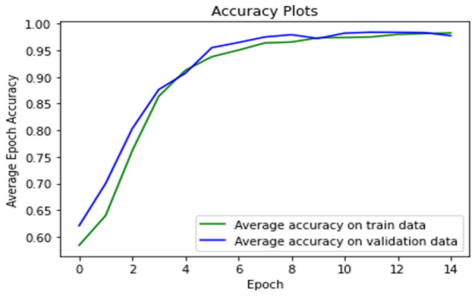


Figure 5. b. Accuracy Plots for Train and Test Dataset

Figure 5. Loss and Accuracy Plots for Training Phase

We then present the confusion matrix, whose sum of its main diagonal cells indicate the accuracy, for the train set on the left-hand side (Figure 6 a), and for the test set on the right-hand side (Figure 6 b) that we respectively get -precisely 98.3% and roughly- 99% for the train and 97% for the test set. An implication of this result is that the pre-attack model will detect 97% of all patients with the correct type of cancer. Moreover, based on the sensitivity definition, it represents the model's probability for predicting malignancy when the patient has the malignant cancer that in this case, the sensitivity for both train and test sets are 98.61% and 76.78%, respectively. Similarly, the specificity shows the model's probability of a predicting

benignity when the patient has benign cancer. Here, the specificity for both train and test sets are 99.18% and 98.27%.

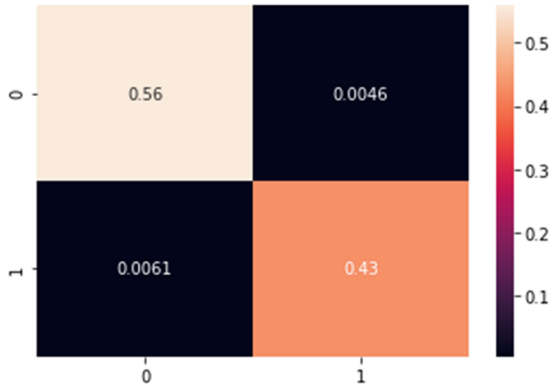


Figure 6. a. Confusion Matrix for Train Dataset

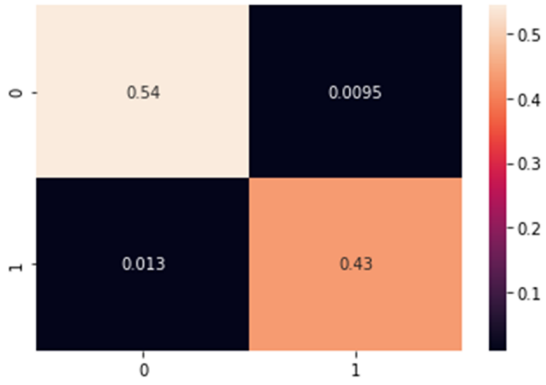


Figure 6. b. Confusion Matrix for Test Dataset

Figure 6. Confusion Matrices for Train and Test Dataset

To evaluate the model robustness of the diagnosis system, we simulate an adversarial attack with two widely used methods: PGD and Universal Patch (see Appendix II for an illustration). Then, we discuss possible fraud scenarios that may occur in the case of an insurance claimant with breast cancer symptoms.

First, Figure 7 shows an example of the results of PGD attacks on the clean images of mammography, where the first column represents the clean images, the second and third columns represent the corresponding perturbations and the results of the misclassification of the attack.

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

A possible scenario may be that a person has diagnosed a benign tumor based on their mammogram with 90% accuracy. However, if the system was attacked by a perturbation derived from a coefficient of their original image, this time their tumor will be detected as malignant, even with higher accuracy of the clean image. Note that lately insurance companies have been extending their policies to those diagnosed with cancer. An attacker, who could be an individual applying directly for insurance - or having access to patient data stored in the healthcare organization - has a financial motivation to portray their disease as malignant.

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.

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

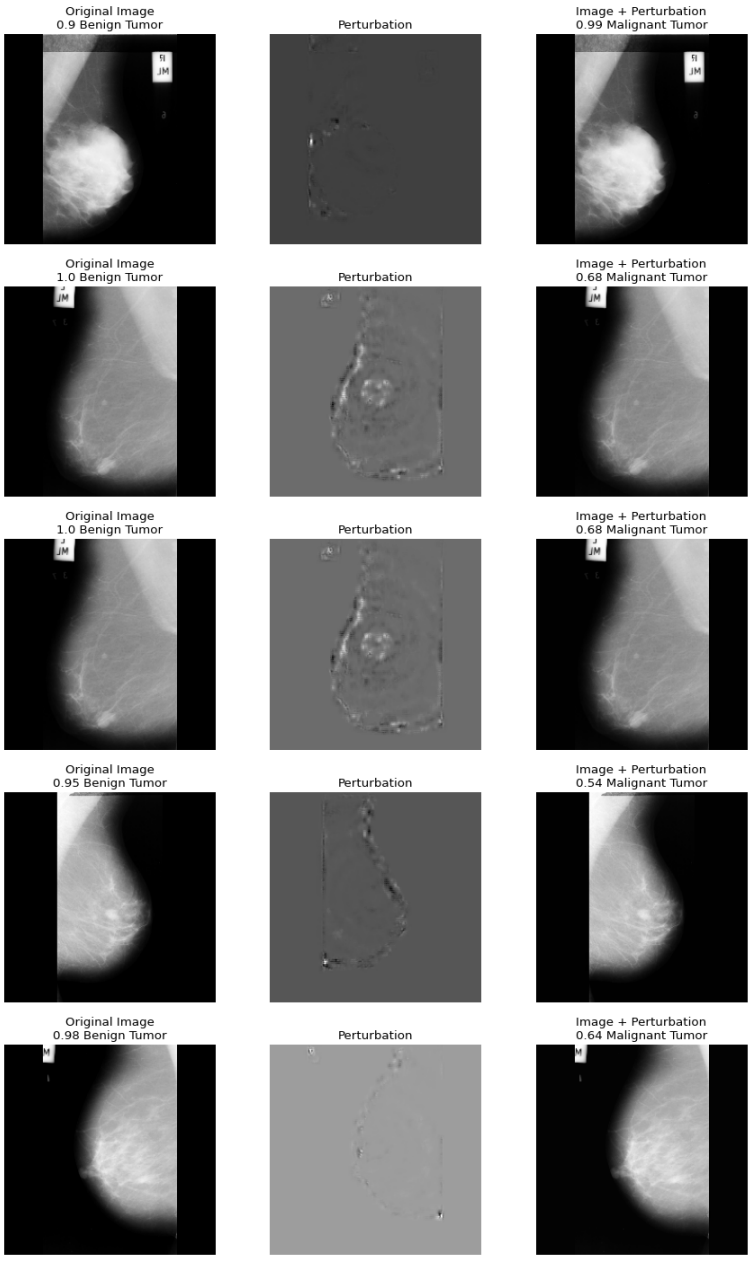


Figure 7. A Sample of Perturbation Caused by PGD and Misclassification Result

Let us discuss now another likely attack for medical images, the so- called Universal Patch,

whose pattern comes directly from the research of (Moosavi-Dezfooli, Fawzi, Fawzi, and Frossardy, 2017). Figure 8 shows the results of the universal attack on breast cancer image. Like the PGD attack, the benign tumor is detected as malignant cancer, so in some cases the adversarial image is even diagnosed with higher accuracy than the original image.

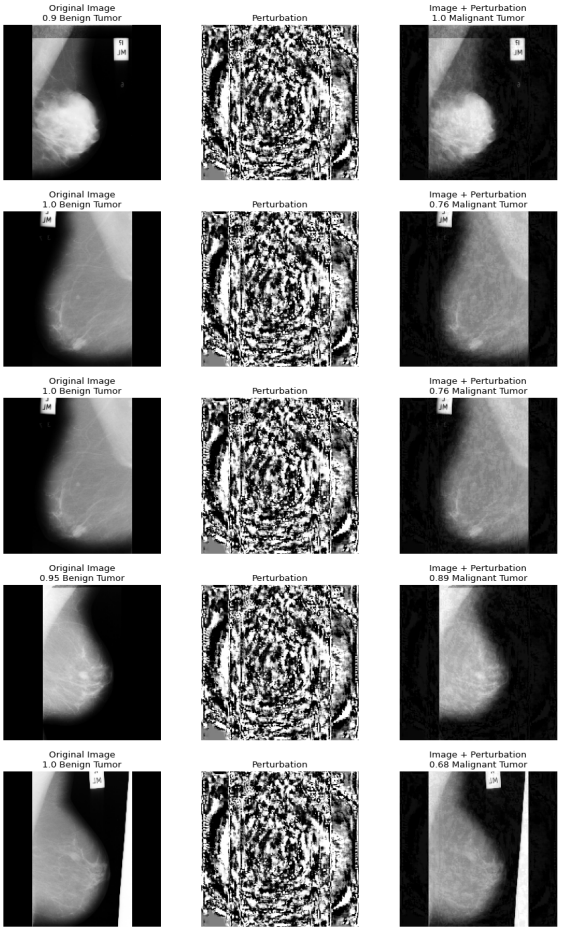


Figure 8. A Sample of Perturbation Caused by Universal attack and Misclassification Result

To show the dependence of the universal attack on the perturbation, there are two levels of perturbation ϵ=0.01, 0.04, the levels of model accuracy change correspond to 0.854 and 0.407. Also, there is a trade-off between the perceptibility and the success rate of the attack in PGD attack. A higher disturbance makes the attack a sure deception of the classification system with universal disturbance, but the attacked image can be easily detected by a trained insurer.

Conclusion of Case Study: This study examines the impact of adversarial attacks on PED diagnostic systems, focusing on breast cancer images as a case study for insurance companies. As mentioned earlier, insurance fraud can be committed by medical providers, patients, and others who intentionally deceive the training system to obtain illegitimate benefits or payments for both government and private insurance programs.

After training a model to diagnose breast cancer, two widespread attacks were used to deceive the systems. The results of our experiment show that not only can the model be fooled for detecting the malignancy of patients, but a higher level of perturbation can increase the success rate of the attack by lowering the accuracy of the system. However, it should be mentioned here that the more noticeable the perturbation, the higher the probability of a detected attack.

In addition, there are many obvious or hypothetical scenarios that can be developed based on the mentioned case study, such as the submission of fake claims for healthy individuals and, conversely, patients may hide their PED previously identified as having a higher risk of disease.

Making the model more robust against potential adversarial attacks requires a holistic view that is not only about the defenses in the technical solution, but also about understanding the broader impact that such attacks have on the organization and how to prevent inappropriate responses.

In this work, we found that medical image recognition requires multiple examinations even in the presence of a professional physician compared to other image-based recognition attacks. This is because viewing multiple medical image features simultaneously is time and budget consuming.

7. AI Robustness and Regulatory Developments

Regulation of the insurance sector is quickly evolving in order to take into account the growing adoption of AI techniques. AI regulation is not proper of the insurance industry, but of all industries, due to the wide adoption of AI across all business sectors. However, it is even more important for insurers, given their fundamental aim of consumer protection.

The EU intervened already in 2019 through the “Ethics Guidelines for Trustworthy AI” developed by the European Commission’s High Level Expert Group on AI (HLEG, 2019). Core to this intervention is the respect of ethics in AI adoption, and therefore the establishment of governance principles that make AI in insurance ethical and worth of being trusted.

Eiopa reinforced the EU interventions by issuing, early this year, a report titled “Artificial Intelligence Governance Principles: Towards Ethical and Trustworthy Artificial Intelligence in The European Insurance Sector” (EIOPA, 2021), in which they recommend an introduction of AI careful of respecting fairness to consumers’, and of keeping their trust high, so as to preserve or

increase inclusion, with the ultimate goal of improving the coverage and protection of EU citizens. To this end, Eiopa introduced six principles, namely the principles of

- proportionality
- fairness and non-discrimination
- transparency and explain ability
- human oversight
- data governance of record keeping, consistently with the national and European data protection laws
- robustness and performance.

The principle of robustness and performance is consistent with our focus on adversarial AI, and a careful examination by insurance companies on how to protect themselves against adversarial attacks we praise for. It is also consistent with the support, expressed by the European Commission since its 2018 White paper, of “a regulatory and investment-oriented approach with the twin objective of promoting the uptake of AI and of addressing the risks associated with certain uses of this new technology.” We do hope then that specific provisions against adversarial attacks will enter the EU recommendations and guidelines (Commission, 2020).

Appendix I. Adversarial Taxonomy

In the current and next appendix, we illustrate adversarial attack taxonomy and its generation method. Hence, in this section we first explain the goals and capabilities of such attacks and them its properties.

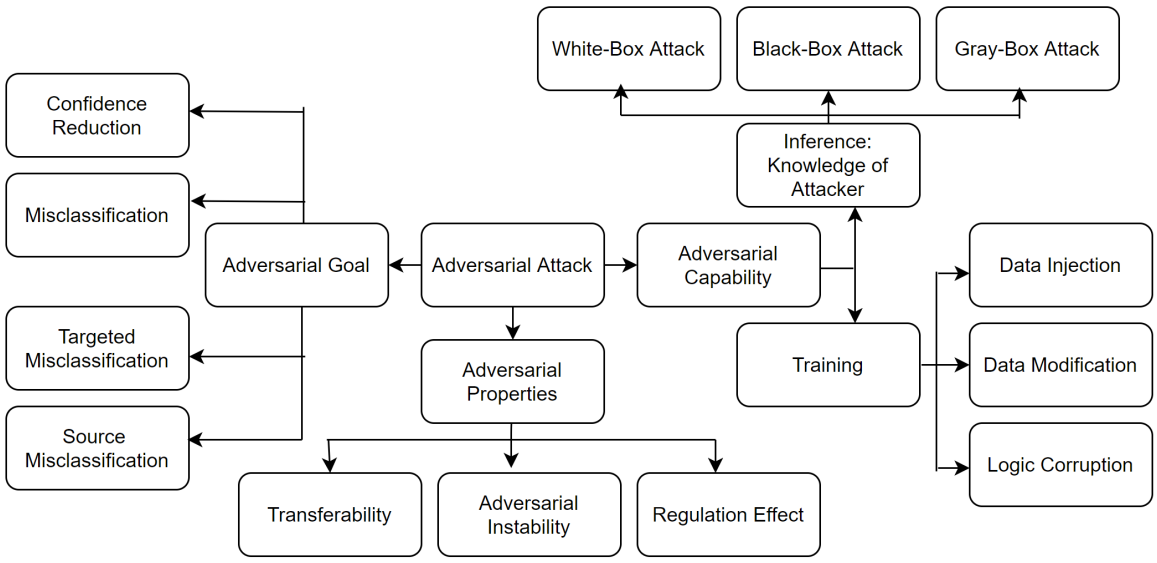


Figure I.1 Adversarial Attack Taxonomy

1. Adversarial Goal

The adversarial goals can be derived from the uncertainties of models. Depending on the impact on the integrity of the output of models, adversarial goals can be divided into four categories, i.e., confidence reduction, misclassification, targeted misclassification, and source/ target misclassification (Ren, Zheng, Qin, and Liud, 2020) (see Figure I.1). To illustrate each category, an example shows three algorithm-generated counterexamples of an ostrich image classified as "safe," "shoe shop" and "vacuum" (Chen, Sharma, Zhang, Yi, and Hsieh, 2018).

Figure I.2. Visual Illustration of Adversarial Examples. The original example is an ostrich image selected from the ImageNet dataset.

- Confidence Reduction: Adversaries try to reduce the confidence of the prediction for the target model, e.g., the adversary samples of an "ostrich" are initially correctly classified with high confidence, but the attack causes the model to detect it with very low confidence.
- Misclassification: The attackers attempt to change the initial classification of the input to an arbitrary class that is different from the original class - also known as untargeted

misclassification - e.g., an arbitrary class is predicted for an adversarial sample of an "ostrich" that is different from the true label. Similarly, in the figure above, an ostrich is predicted to be "safe", "shoe shop", or "vacuum".

- Targeted Misclassification: Adversaries attempt to change the output to a specific target class, for example, any adversarial pattern entered a classifier is predicted to be "vacuum".
- Source/ Misclassification: Adversaries attempt to change the output classification of a specific input to a specific target class. e.g., an "ostrich" image is predicted to be a "vacuum".

2. Adversarial Capabilities

Security is defined in terms of stronger or weaker adversaries who have access to the system and its data. The term *capabilities* refers to the what and how of the available attacks and defines the possible attack vectors on a threat surface (Papernot, 2018). Figure I.1 shows the range of attacker capabilities in ML systems as they relate to the inference and training phases.

2.1 Inference Phase: Attacks at the time of inference-exploratory attacks (Barreno, Nelson, Sears, Joseph, and Tygar, 2006) do not manipulate the targeted model, but either make it produce adversary-selected outputs or gather evidence about model properties. 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.

There are three common threat models in the inference phase for adversary attacks and defenses: the white-box, gray-box, and black-box models (Ren et al., 2020).

- **White-Box Attack:** In white-box attacks, the attackers have complete knowledge of the target model - including the model architecture (algorithm and structure of hypothesis H), model parameters θ (weights), training data, or combinations thereof. The attackers identify the most vulnerable feature space of the target model using the available information, and then modify an input using adversarial pattern generation methods. Accessing the model's internal weights for white-box attacks corresponds to a strong adversarial capability (Qiu, Liu, Zhou, and Wu, 2019).
- **Black-Box Attack:** In the black box model, an attacker does not know the structure of the target network or the parameters, but the attacker uses information about the environment or past inputs to infer the vulnerability of the system in these attacks. The attackers always create adversarial samples on a surrogate classifier trained with the acquired data and prediction pairs and other benign/adversarial samples. Due to the portability of adversarial samples, 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 (H is known) of the target model but does not have access to the model parameters (θ is unknown). The attacker can also interact with the DL algorithm. In this threat model, it is assumed that the attacker creates adversarial samples 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.

2.2 Training Phase: Attacks on training seek to learn, influence, or corrupt the model itself. The simplest and arguably weakest attack on training is to simply access a summary, part, or all the training data. Depending on the quality and scope of the data, the attacker may be able to create a substitute model (also known as a surrogate or auxiliary model) to perform attacks on the victim system. For example, the attacker may use a surrogate model to test potential inputs before sending them to the victim system (Rndic and Laskov, 2014).

The attack strategies used in the training phase can be divided into three categories:(Ren et al., 2020)

- **Data Injection:** The attacker does not have access to the training data and learning algorithms but can add new data to the training dataset. The adversary can falsify the target model by adding adversarial samples to the training dataset.
- **Data Modification:** The adversary does not have access to the learning algorithms, but it does have access to the full training data. The adversary can poison the training data by modifying the data before it is used to train the target model.
- **Logic Corruption.** The adversary has access to interfere with the learning algorithms of the target model.

3. Adversarial Properties

Adversarial examples has three basic properties, i.e., transferability, regularization effect, and adversarial instability (Zhang and Li, 2018).

- **Transferability:** Adversarial examples are not limited to attacking a particular neural network. Adversarial examples generated by one model H_1 can fool another model H_2 with a similar probability. Therefore, an attacker can construct adversarial examples in a known machine learning model and then attack related unknown models.
- **Adversarial Instability:** In the physical world, it can easily happen that after physical transformations, such as translation, rotation, and illumination of a specific adversarial example, its ability is lost. In this case, the adversarial example is correctly classified by the model. This instability property challenges attackers to construct robust adversarial examples and makes it difficult to deploy attacks in the real world.

- **Regularization Effect.** Adversarial training (Goodfellow et al., 2015) is a regularization method that can reveal the defects of neural network systems and improve the robustness of adversarial examples. Compared to other regularization methods, the cost of constructing adversarial samples is expensive.

Appendix II- Adversarial Attacks Generation

Although adversarial attacks can be generated in different ways, two main categories have been proposed (Zhang et al., 2021):

a) minimizing the perturbation size - distance between the adversarial example and the instance to be manipulated - assuming that the image is misclassified

b) maximizing the attack success rate assuming a limited perturbation budget.

As explained above, 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, which are described in the next section.

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:

$$minimize \left\|x_{adv} - x\right\|_p \; s.t. \; H(x_{adv}) \neq y \tag{II.1}$$

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:

$$minimize \left\|x_{adv} - x\right\|_p \; s.t. \; H\left(x_{adv}\right) = y_{target} \neq y \tag{II.2}$$

In general, targeted adversarial examples are more difficult to generate than untargeted adversarial examples. Different ways to solve both (II.1) and (II.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)

$$x_{adv} = x - \epsilon \, . \, sign\left(\nabla_x J(\theta,\ x,\ t)\right) \tag{II.3}$$

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

$$x_{adv} = x - \epsilon \, . \, sign\left(\nabla_x \ J(\theta,\ x,\ y)\right) \tag{II.4}$$

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., $H\left(x_{adv}\right) \neq y$). 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 (II.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

$$x_{adv}^0 = x,\ x_{adv}^i = clip_{x,\epsilon}(x_{adv}^{i-1} + \epsilon sign(\nabla_{x_{adv}^{i-1}}\ J(\theta,\ x_{adv}^{i-1},\ y))) \quad \text{(II.5)}$$

The perturbation process can stop in two cases: first, when the misclassification $H\left(x_{adv}\right) \neq y$ 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, Goodfellow, and Bengio, 2017b) and it iteratively performs the FGSM attack. In each iteration, only a fraction of the allowed noise limit is added; this contributes to its higher attack effectiveness compared to FGSM. Essentially, I-FGSM is the same as PGD, the only difference being that the PGD attack initializes the perturbation with 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. Another advantage of initialization is that it can help to further improve the success rate of the attack with multiple attempts.

1.3 DeepFool: The DeepFool algorithm (Moosavi-Dezfooli, Fawzi, and 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. Moreover, the authors propose DeepFool as an untargeted attack, but the algorithm can in principle be easily modified for the targeted setting.

By considering DNNs as affine classifiers, 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 and Wagner Attack (CandW): CandW's attack attempts (Carlini and Wagner, 2017) to find the minimally biased perturbation problem - similarly to the DeepFool algorithm - as follows:

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

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 (Zhang et al., 2021). Then H is defined as $H(x', t) = (max_{i \neq t} Z(x')_i - Z(x')_t)^+$.Minimizing $H(x', t)$ 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 (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 et al., 2017) was the first to present an 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. In addition, 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, RNN).

The authors in (Zhang et al., 2021) exploit the “transferability” property (Section I.3) 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 (Xu, H. et al., 2020):

1. Synthesize a substitute training dataset: Create a "replica" training set. For example, to attack handwritten digits recognition, create an initial substitute training set 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 proposed by Chen et al. (Chen, Zhang, Sharma, Yi, and Hsieh, 2017) for black-box attacks. 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 "scrape" the gradient information around the victim sample by observing the changes in the prediction confidence $H(x)$ as the pixel values of x are tuned.

Equation (II.7) shows that for each index i of sample x, we add (or subtract) x_i by ϵ. If ϵ is small enough, we can extract the gradient information from the output of $H(\ .\ )$ by

$$\frac{\partial h(x)}{\partial x_i} \approx \frac{H\left(x + \epsilon e_i\right) - H\left(x - \epsilon e_i\right)}{2\epsilon} \tag{II.7}$$

3. Universal Attack:

Adversarial attacks described so far always manipulate a single image to fool a classifier with the specific combination of image and 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 (i.e., image-independent) perturbations. Universal perturbations can pose a greater threat because they allow an attacker to create attacks where it does not matter what underlying pattern is used for the attack. The goals of a universal perturbation have to be at a minimum distance from the class boundary of a class $y' \neq y_{true}$ to ensure imperceptibility of the perturbation and to fool as many images as possible from a separate subset X_{test}. The authors have described these goals in terms of two constraints for their minimization problem such that:

$$\begin{aligned} \|\eta\|_p &\leq \epsilon \\ P\Big(H\big(x+\eta\big) \neq H(x)\Big) &\geq 1-\delta \end{aligned} \tag{II.8}$$

Recall that η is the adversarial perturbation with $p-$norm constraint ϵ and $H()$ is the classifier. δ is a newly introduced parameter that quantifies the fooling rate.

References

Arrietaa, A. B., Díaz-Rodríguezb, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., . . . Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82 – 115.

Bapna, R., Jank, W., and Shmueli, G. (2008). Consumer Surplus in Online Auctions. Information Systems Research, 19(4), 400-416.

Barreno, M., Nelson, B., Sears, R., Joseph, A. D., and Tygar, J. D. (2006). Can Machine Learning Be Secure? ACM Symposium on Information, Computer and Communications Security, 16-25.

Brown, B., Mané, D., Roy, A., Abadi, M., and Gilmer, J. (2017). Adversarial Patch. 31st Conference on Neural Information Processing Systems (NIPS 2017). doi:arXiv:1712.09665v2

Carlini, N., and 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

Carlini, N., and Wagner, D. (2018). Audio Adversarial Examples: Targeted Attacks on Speech-to-Text.

Chen, H., Zhang, H., Chen, P. Y., Yi, J., and Hsieh, C. J. (2018). Attacking Visual Language Grounding with Adversarial Examples: A Case Study on Neural Image Captioning. 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018).

Chen, P. Y., Sharma, Y., Zhang, H., Yi, J., and Hsieh, C. J. (2018). EAD: Elastic-Net Attacks to Deep Neural Networks via Adversarial Examples. doi:arXiv:1709.04114v3

Chen, P. Y., Zhang, H., Sharma, Y., Yi, J., and 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

Code, A. C. (2005). Definition of Insurance Contract in European Commission. Retrieved from https://ec.europa.eu/info/sites/info/files/definition_of_insurance_contract_en.pdf

Commission, E. (2020). White Paper On Artificial Intelligence: A European Approach to Excellence and Trust.

EIOPA. (2021). Artificial Intelligence Governance Principles: Towards Ethical and Trustworthy Artificial Intelligence in the European Insurance Sector. Retrieved from

Eling, M., Nuessle, D., and Staubli, J. (2021). The impact of artificial intelligence along the insurance value chain and on the insurability of risks. The Geneva Papers on Risk and Insurance - Issues and Practice. doi:https://doi.org/10.1057/s41288-020-00201-7

Europe, I. (2013). The impact of insurance fraud. Retrieved from www.insuranceeurope.eu/sites/default/files/attachments/The%20impact%20of%20insurance%20fraud.pdf

Europe, I. (2019). Insurance Fraud: Not a Victimless Crime. Retrieved from

Evtimov, I., Eykholt, K., Fernandes, E., Kohno, T., Li, B., Prakash, A., . . . Song, D. (2017). Robust Physical-World Attacks on Deep Learning Models. .

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

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. The 5th IEEE International Conference on Data Science and Advanced Analytics (DSAA 2018).

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

GovInsider. AI Is Changing Healthcare and Insurers Are Taking Notice. Retrieved from https://govinsider.asia/inclusive-gov/ai-changing-healthcare-insurers-taking-notice/

Hartman, E., Grieve, R., Ramsahai, R., and Sekhon, J. (2015). From SATE to PATT: Combining Experimental with Observational Studies to Estimate Population Treatment Effects. Royal Statistical Society, 757–778.

HLEG. (2019). Ethics Guidelines for Trustworthy AI. Retrieved from https://digital-strategy.ec.europa

Kamil, A., Hassan, I., and Abraham, A. (2016). Modeling Insurance Fraud Detection Using Ensemble Combining Classification. Computer Information Systems and Industrial Management Applications, 8.

Katkar, R., and Lucking-Reiley, D. (2006). Public Versus Secret Reserve Prices in eBay Auctions: Results from a Pokémon Field Experiment. Advances in Economic Analysis and Policy, 6(2).

Kenett, R. S., and Shmueli, G. (2016). Information Quality: The Potential of Data and Analytics to Generate Knowledge: John Wiley and Sons.

Kenett, R. S., Zacks, S., and Amberti, D. (2021). Modern Industrial Statistics: with applications in R, MINITAB and JMP (3rd ed.): John Wiley and Sons.

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., and Bengio, S. (2017a). Adversarial Examples in the Physical World. ICLR, 14.

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

Lei, Q., Wu, L., Chen, P. Y., Dimakis, A. G., Dhillon, I. S., and Witbrock, M. (2018). Discrete Adversarial Attacks and Submodular Optimization with Applications to Text Classification.

Li, B., Ge, Y., Zhao, Y., Guan, E., and 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

Li, Y., Yan, C., Liu, W., and Li, M. (2016). Research and Application of Random Forest Model in Mining Automobile Insurance Fraud. 12th International Conference on Natural Computation and 13th Fuzzy Systems and Knowledge Discovery.

Liab, D., Zhangab, J., and Huang, K. (2021). Universal Adversarial Perturbations Against Object Detection. Pattern Recognition, 110.

Longo, L., Goebel, R., Lecue, F., Kieseberg, P., and Holzinger, A. (2020). Explainable Artificial Intelligence: Concepts, Applications, Research Challenges and Visions. Springer, 12279.

Lundberg, S., and Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS), 4768–4777.

Maull, R., Collomosse, J., Brewer, S., Bordon, A., Jones, K., and Breeze, J. (2019). Taking Control: Artificial Intelligence and Insurance. Retrieved from

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

Molnar, C. (2020). Interpretable Machine Learning, A Guide for Making Black Box Models Explainable.

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

Moosavi-Dezfooli, S. M., Fawzi, A., and 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. (2018). Characterizing The Limits And Defenses of Machine Learning in Adversarial Settings. (Doctor of Philosophy),

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and 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

Pearl, J. (2015). Generalizing Experimental Findings. Causal Inference, 3(2), 259–266. doi: 10.1515/jci-2015-0025

Pearl, J., and Bareinboim, E. (2011). Transportability Across Studies: A Formal Approach. Retrieved from Los Angeles, USA:

Pearl, J., and Bareinboim, E. (2014). External Validity: From Do-Calculus to Transportability Across Populations. Statistical Science, 29(4), 579–595. doi:10.1214/14-STS486

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

Rao, C. R. (1985). Weighted distributions arising out of methods of ascertainment: what population does a sample represent? New York, NY: Springer.

Rawte, V., and Anuradha, G. (2015). Fraud Detection in Health Insurance using Data Mining Techniques. International Conference on Communication, Information and Computing Technology (ICCICT).

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

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why Should I Trust You? Explaining the Predictions of any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. doi:http://dx.doi.org/10.1145/2939672.2939778

Rizk, H., Elgokhy, S., and Sarhan, A. (2015). A Hybrid Outlier Detection Algorithm Based On Partitioning Clustering And Density Measures. 2015 Tenth International Conference on Computer Engineering and Systems (ICCES). doi:10.1109/ICCES.2015.7393040

Rndic, N., and Laskov, P. (2014). Practical Evasion of a Learning-based Classifier: A Case Study. IEEE Symposium on Security and Privacy, 197–211.

Sadeghi, S., Dadkhah, S., Jalab, H. A., Mazzola, G., and 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.

Sharif, M., Bhagavatula, S., Reiter, M. K., and Bauer, L. (2016). Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. doi:http://dx.doi.org/10.1145/2976749.2978392

Singh, A. K., Kumar, B., Singh, G., and 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

Sithic, H. L., and Balasubramanian, T. (2013). Survey of Insurance Fraud Detection Using Data Mining Techniques. International Journal of Innovative Technology and Exploring Engineering, 2(3).

Sokolic, J., Giryes, R., Sapiro, G., and Rodrigues, M. R. D. (2017). Generalization Error of Invariant Classifiers. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), 54.

Startup, I. https://www.telus.com/en/health/health-professionals/allied-healthcare-professionals/eclaims.

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

Sundarkumar, G. G., and Ravi, V. (2015). A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance. Engineering Applications of Artificial Intelligence, 37. doi:10.1016/j.engappai.2014.09.019

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2015). Intriguing Properties of Neural Networks. ICLR.

Tsoi, A. C., Zhang, S., and Hagenbuchner, M. (2005). Pattern Discovery on Australian Medical Claims Data: A Systematic Approach. IEEE Transactions On Knowledge And Data Engineering, 17(10).

Veiga, S. D., Gamboa, F., Iooss, B., and Prieur, C. (2021). Basics and Trends in Sensitivity Analysis : Theory and Practice in R: Society for Industrial and Applied Mathematics.

Verma, A., Taneja, A., and Arora, A. (2017). Fraud Detection and Frequent Pattern Matching in Insurance claims using Data Mining Techniques. Proceedings of 2017 Tenth International Conference on Contemporary Computing (IC3).

Wang, X., Montgomery, A., and Srinivasan, K. (2008). When Auction Meets Fixed Price: a Theoretical and Empirical Examination of Buy-It-Now Auctions. Quantitative Marketing and Economics, 6. doi: https://doi.org/10.1007/s11129-008-9041-0

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.

Whitaker, J. E. (2019). Insurance Fraud Handbook. Association of Certified Fraud Examiner.

Xu, H., Ma, Y., Liu, H., Deb, D., Liu, H., Tang, J., and 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

Xu, W., Wang, S., Zhang, D., and Yang, B. (2011). Random Rough Subspace based Neural Network Ensemble for Insurance Fraud Detection. Fourth International Joint Conference on Computational Sciences and Optimization.

Zanghieri, P. (2017). What Causes Insurance Fraud? Evidence on Motor third Part Liability in Italian Provinces. Generali Investment, Macro and Market Research.

Zhang, C., Benz, P., Lin, C., Karjauv, A., Wu, J., and 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., and Li, C. (2018). Adversarial Examples: Opportunities and Challenges. IEEE Transactions on Neural Networks and Learning Systems, 16. doi:10.1109/TNNLS.2019.2933524