UNECE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Confidentiality

1-3 December 2021, Poland

Statistical disclosure control for machine learning models.

Felix Ritchie (University of the West of England)

felix.ritchie@uwe.ac.uk

Abstract

AI models are trained on large datasets. Where the training data is sensitive, the data holders need to consider risks posed by access to the training data and risks posed by the models that are released.

The first problem can be considered solved: there are multiple tested solutions delivering secure access to sensitive data for research purposes. These include robust 'statistical disclosure control' (SDC) procedures for checking the confidentiality risk in outputs released from the secure environment. However, these SDC procedures are designed for statistical outputs. It is not clear how they relate to AI model specification created within the secure environment.

Similarly, there is a small but growing literature on re-identification and other risks from AI models trained on personal data. However, this does not consider the operational circumstances which might limit opportunities for misuse.

We bring these two fields together to consider

- Is there any conceptual risk from releasing AI model specifications from a controlled environment?
- If so, is there any practical risk?
- If so, are there effective controls to minimise that practical risk without excessive cost or damage to the data/models?

We present case studies using specific intruder scenarios, develop response mechanisms, and suggest what lessons can be learned for the wider class of ML models.

Statistical disclosure controls for machine learning models

Susan Krueger*, Esma Mansouri-Benssassi*, Felix Ritchie**, Jim Smith**

* University of Dundee SKrueger001@dundee.ac.uk, EMansouribenssass001@dundee.ac.uk
** University of the West of England, Bristol felix.ritchie@uwe.ac.uk, James.Smith@uwe.ac.uk

Abstract
Artificial Intelligence (AI) models are trained on large datasets. Where the training data is sensitive, the data holders need to consider risks posed by access to the training data and risks posed by the models that are released.
The first problem can be considered solved: there are multiple tested solutions delivering secure access to sensitive data for research purposes. These include robust 'statistical disclosure control' (SDC) procedures for checking the confidentiality risk in outputs released from the secure environment. However, these SDC procedures are designed for statistical outputs. It is not clear how they relate to AI model specification created within the secure environment.
Similarly, there is a small but growing literature on re-identification and other risks from AI models trained on personal data. However, this does not consider the operational circumstances which might limit opportunities for misuse.
We bring these two fields together to consider

- Is there any conceptual risk from releasing AI model specifications from a controlled environment?
- If so, is there any practical risk?
- If so, are there effective controls to minimise that practical risk without excessive cost or damage to the data/models?

We show that there is certainly a theoretical risk, which also seems to have practical validity. There exist both statistical/technical controls to reduce risk, as well as operational controls which might be relevant for restricted environments. However, there remains a very large degree of uncertainty, including such fundamental questions as what exactly is 'disclosive' in ML models.

1 Introduction

In recent years, the availability of large datasets, growing computer power and complex machine learning (ML) techniques has made the use of AI models for service delivery increasingly practical and efficient. These models often require large volumes of sensitive eHealth data to train them effectively. Correspondingly, this century has also seen an explosion in the availability and effectiveness of 'trusted research environments' (TREs) – facilities which allow researchers to work with very few restrictions on highly sensitive data, in an environment which ensures no 'leakage' of confidential data.

One of the components of the TRE security model is checking that any outputs released from the TRE do not release confidential data. Output statistical disclosure control (SDC) techniques have been developed for these environments, but they are designed to deal with statistical output such as tabulations, coefficient estimates, or graphs. While there are some similarities between ML models and regression models, ML models present a number of new challenges: different representation of data, different attack scenarios, different reconstruction possibilities, a different scale of parameter extraction.

ML researchers have already identified a number of attack scenarios allowing source data to be reconstructed, or other confidential information to be gleaned (such as whether an individual is in the source file). However, to date the operational context of the models has not been considered: given that the training data is held within a secure environment and only the model is released, can that model alone be a source of confidentiality breach? And if so, what are the range of statistical and non-statistical protection measures that could be employed to limit the risk?

This paper considers this problem, using the example of image recognition modelling carried out in the Scottish National Safe Haven (SNSH). As the SNSH is typical of a modern research data centre (the most common form of accredited TRE) the results are easily generalizable. We have not covered the full range of ML models; instead we focus on illustrating how to approach the assessment of ML models. The choice of image processing is to demonstrate risks in relatable terms (recognizable objects rather than internal hard and soft tissue scans) and for which human inspections and control methods alone do not suffice (due to complexity of the models). We distinguish between theoretical and practical risks, and suggest statistical and non-statistical solutions to manage the identified risks.

The next section describes the context for this paper in more detail. Section 3 reviews the literature on outputs SDC, TREs, and confidentiality risks in ML models. Section 4 outlines the methods to be used. Section 5 details specific attacks, and the factors which determine the likelihood of success. Section 6 considers how the TRE design affects these factors, and develops guidelines to limit risk exposure and formal measures to assess risk. Section 7 concludes.

2 Operational context

Healthcare research has traditionally involved statistical analysis on structured eHealth data recorded in flat files. Recent advances in medical imaging and genomics make these data types now also available for research, driving the need for advanced tools and approaches (Nind et al. 2020). The discipline of Radiomics, for instance aims to enhance the existing data available to clinicians through advanced mathematical analysis of medical images, uncovering disease characteristics before they are visible to the naked eye (van Timmeren, J. et al. 2020). Machine learning and deep learning are two such advanced mathematical techniques used in the field of Artificial Intelligence (AI) for healthcare.

Complex AI models generally require large volumes of training data to 'learn' effectively; however, research cohorts are usually composed of relatively small, narrow subsets of people with specific conditions. Consented medical data are typically collected using specific acquisition protocols under ideal conditions thus generalising findings and repurposing data is problematic (Nind et al. 2020).

In order for AI to benefit a whole population, not just narrow subsets, it needs to be trained on population-level data. This can be achieved through the use of *routinely collected* health data where consent for research purposes may not have been explicitly granted at the time of collection and is not feasible to acquire post collection.

In Scotland, the use of *unconsented* data is allowed in healthcare research for public benefit; however a common law right to privacy does exist and must be protected through safeguards that prevent misuse and identification of individuals (Charter for Safe Havens in Scotland. 2015). Machine learning and AI methods of research introduce new risks to individual privacy that we seek here to understand and mitigate.

The Charter for Safe Havens in Scotland sets out the agreed principles and standards for handling unconsented health data to support research and statistics in the public benefit. Safe Havens are Trusted Research Environments (TRE) that enforce the "5 Safes" (Ritchie, 2017):

1. safe projects – researchers must show their research delivers clear public benefits and that appropriate governance is in place such as ethics and data controller approvals
2. safe people – researchers must have the technical skills to use the data, approved training in Information Governance, and agree to protect confidentiality of data at all times
3. safe places – data is held in a secure environment, accessed under restricted conditions and analysed on controlled systems with built in security controls
4. safe data – researchers are only given access to the minimum data required to answer their research question, stripped of personal identifiers
5. safe outputs – research results extracted from the secure environment are checked and assessed to ensure they don't contain potentially identifiable or disclosive information.

TREs are mostly implemented as research data centres (RDCs). RDCs allow researchers full access to the data as if it was on their local machines, but in an environment that limits the ability of researchers to upload or download data or results without the approval of the RDC manager. A rarer type of TRE is the remote job server (RJS) which allows researchers to request statistical analysis, either from a menu or by sending in code, without generally being able to see the microdata.
The 'five safes' framework (Ritchie, 2017), described above, illustrates the security controls available to TRE managers. Of particular relevance to TREs are "safe people" and "safe outputs"; the first addresses the likelihood of researchers breaching TRE rules deliberately or accidentally; the second describes the measures in place to check outputs. Green and Ritchie (2015) and Green et al (2020) show there are a wide variety of practices in TREs internationally: user training varies from non-existent to intensive face-to-face training; output checking also ranges from no checks (and assuming the researchers make no mistakes) to everything being checked by at least two pairs of eyes. Those who do not check outputs rely on training and/or written instructions on how to check outputs.
The Scottish National Safe Haven (SNSH) is typical of modern TREs. It uses well-established technical, procedural and statistical controls, allied to good-practice training models for researchers and TRE staff. As such, the discussion here is directly applicable to other TREs allowing machine learning models to be generated within their secure environment.

3 Literature review

3.1 Output SDC

Output SDC is the checking of statistical outputs to ensure that they present no meaningful risk of disclosing confidential information used to generate them (in contrast, input SDC is concerned with reducing the detail of the source data). As the literature on output SDC has been dominated by the needs of statistical agencies, almost all of it focused on simple linear statistics, such as tabulations of frequencies, medians or means. However, Ritchie (2016) discusses a generalised approach to output SDC (see section 4 below). Ritchie (2016) then applies this to linear regression and frequency tables, showing that there is no practical disclosure risk in the former, and considerable risk in the latter.
The 'no practical risk' arises even on the part of malicious action by the researcher for TREs, because a malicious researcher can observe information directly on a data point; there is therefore no incentive to fake regressions. In theory this provides a useful example: both linear regression and ML models generate representations of the key features of the data, not exact reproduction. However, there are two significant differences. First, regression models generate few parameters relative to the number of observations, and often estimate 'incidental' parameters which are not published. Second, regression source data is likely to contain single values of interest to

malicious researchers, and so remembering them is feasible. In contrast, the data input to an ML model, such as a facial image, is only of value in its entirety; humans memorising data points is not feasible, and so there is more incentive to falsify models to enable the extraction of an entire image or other input.
There is no clear agreement in the literature on what constitutes 'acceptable' risk in outputs. Traditional generalist SDC texts such as Hundepool et al (2010) are content to explain the theoretical basis of SDC, but provide little indication of how an acceptable level of risk may be chosen in practice; they also tend to assume that such evaluations are objective. In contrast, manuals for output checkers such as Brandt et al (2010) or SDAP (2019) do discuss the need to evaluate outputs in context, and are more explicit about the subjectivity of decision-making. Hafner et al (2015) point out, that once subjectivity is acknowledged, the position of the organisation as 'default-open' (release unless problem identified) or 'default-closed' (do not release unless no problems identified) has a major effect on release decisions.

3.2 Machine Learning Models

With the growing availability and complexity of data in various domains, there is an increasing use of machine learning for analysis inside TREs and a growing demand for release of trained models from those environments due to generalisation of ML application and opportunity for clinical impact in practice. Release requests for machine learning applications consist mainly of 'learnt models' comprising a model's architecture and learned weights. Learnt models also include various hyperparameters, enabling correct inference on unseen data.
There are multiple types of machine learning algorithms (Aiswariya et al. 2020). The most popular types are cited as follows:

- Classical machine learning: Classical models are the most popular machine learning models. They include models such as linear regression, logistic regression, Support Vector Machines, decision trees or k-means clustering.
- Ensemble Models: Ensemble models consist of aggregating multiple learning models; they generally achieve a better performance than individual and traditional machine learning models. Gradient boosting and random forest are examples of ensemble models.
- Neural networks: neural networks are designed for more complex tasks especially for unstructured data such as images, videos or audio. There are different types of models such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, Recurrent neural networks (RNNs) or autoencoders, which are used to encode input data in an unsupervised manner.

'Classical' algorithms, and (relatively) simpler Multi-Layer Perceptron's rely on input data being first processed or 'encoded' so that each example is described by defined values for a fixed number of features. In contrast, so-called "Deep Learning"

approaches typically use many-layered architectures and take as inputs ‘raw’ data (images, audio, etc), where the first few layers automate the process of feature creation. While more powerful, not surprisingly there is a trade-off here: the early layers are all defined by (large numbers of) parameters, and there is a consequent risk that these effectively ‘remember’ the training data, rather than generalising from it. Also, the autoencoding process can make the model very difficult if not impossible for humans to interpret or explain.
ML models are trained in various ways including supervised, semi-supervised, unsupervised, or reinforcement learning. Supervised learning relies on the algorithm having labelled data to learn patterns from, whereas unsupervised learning relies on the algorithm identifying discriminative features in the data without prior knowledge of labels.

3.3 Risks of disclosure through ML models

Machine learning models are prone to various attacks. There exist two main threats on machine learning, the first during training and the second during production and deployment (He et al 2020). Attacks such as data poisoning, and evasion happen at the training level and are not meaningful within TREs. On the other hand, attacks at a production and deployment level can occur after models are released from those environments.
In this section we describe what we identified as the most significant risks on machine learning models affecting their release from TREs.

3.3.1 Model inversion attacks

Model inversion attacks are the most common attack on machine learning. Various types of inversion attack have been published in the literature, such as (Fredrikson et al, 2019).
Model inversion attacks are also referred to as reconstruction attacks, aimed at reconstructing part or full training datasets, data labels or both. These attacks can be particularly dangerous for private and confidential data such as medical or facial recognition applications (Kaissis et al. 2020). However, the most popular ones use existing knowledge of the models such as some known features, labels or weights and aim to recover sensitive features or some of the training data itself.

3.3.2 Membership inference attacks

Membership inference attacks aim at determining whether a sample data point was used for training a machine learning model. These kinds of attacks are usually applied in a black-box fashion where there is no prior knowledge of model architecture, parameters, or weights (Chakraborty et al. 2018).
The following section describes case studies and examples of model inversion and membership inference attacks demonstrated with a worst case scenario in the models’ design.

4 Methods

In this paper we follow the approach outlined in Ritchie (2016) for assessing the risk in particular types of output. This consists of

1. Identifying the statistical form of the output
2. Identifying theoretical risks:
    a. inherent in the output, such as unique observations
    b. by comparing the output with similar outputs that, for example, differ by a single observation or parameter/variable
3. Identifying the necessary conditions for the risk to manifest itself
4. Reviewing the likelihood of those conditions occurring in practice
5. Identifying remedial action to reduce disclosure risk

To identify theoretical risk we use an 'ultra-intruder' model: that is, we assume that the intruder has access to almost unlimited external knowledge, and has no objective other than to uncover confidential information. We also focus on the worst cases within each group of results; that is, if a more complex ML procedure makes it harder to uncover confidential data, we ignore it. For ML modelling, we assume that the output takes the form of all the model parameters and any auxiliary information that is produced as part of the model.

The intruder model has been criticised for its use in decision-making (eg Hafner et al, 2015) but it is useful for defining an overall worst-case theoretical position. However, for practical guidance we need to identify what are the genuine risks in realistic (even if unlikely) scenarios. Identifying necessary and/or sufficient conditions for confidentiality breaches to occur allows their likelihood to be assessed in the context of a particular access regime – such as a TRE.

As noted above, there are also metrics which can quantify the risk. These can be used to provide quantitative support for what seems 'likely', as well as providing a handy metric to be applied to outputs.

Finally, confidentiality breaches can arise from both insider risk (an authorised user deliberately creating misleading outputs to fool output checking) and outsider risk (an external individual with access to the released output and knowledge of the technique used to create it, who then tries to uncover the source data). As noted above, while insider risk is not an issue for regression modelling (except in RJSs), for ML models there is at least a theoretical incentive for insider risk. The distinction is important, as different remedies may be appropriate to insider and outsider risk.

As this is the first paper to consider disclosure risk in ML models, we do not try to cover the full range of ML models. Instead we focus on an AlexNet model to illustrate how we can approach the assessment of ML models. Nor do we pursue 2(b), above.

Note that the measure of 'disclosure risk' is still undetermined. In traditional SDC modelling, risk is measured as the probability of extracting an exact value (or within a given tolerance for dominance checks). In the data sources used for AI, a large degree of 'fuzziness' may still not protect against a disclosure risk – a picture may be easily

recognisable, even if blurred. We are exploring this, but for now we use subjective measures of ‘close enough’.

5 Risk assessment

In this section we describe two case studies with experiments and analysis of the two identified risks for machine learning models.

5.1 Case 1: model inversion attack

In model inversion attacks, the attacker has prior knowledge of the model and its parameters. In the first case we experimented with an attack model based on the work of Fredrikson et al (2019). The attack model aims at optimizing the image to minimize the loss given a set of fixed weights, instead of optimizing the weights to minimize the loss as occurs during conventional model training. In other words, it reverse-engineers the model process. In conventional SDC terms, this is the equivalent of trying to reconstruct an observation from the estimated coefficients.
The attack model exploits the vulnerability of the target model as follows:

- The attack model has access to the model architecture, parameters and target labels.
- For each class label of the target model, the adversary first creates a noise image, feeds this sample to the model, and computes the posteriors.
- The attack model uses backpropagation over the target model’s parameters to optimize the input sample so that the corresponding posterior of the class can exceed a pre-set threshold.
- Once the threshold is reached, the optimized sample is the representative sample of that class, i.e., the attack output.

5.1.1 Target model architecture

AlexNet is chosen as a target model (Krizhevsky et al. 2017), which consists of eight layers with five convolution layers and three fully connected layers. The model is modified and trained without dropout or regularisation layer in order to simulate an overfitting model, which is more prone to attacks.
The model is trained on the CIFAR10 dataset where the training set consist of 60,000 images and the test set consists of 10,000 images. The model is trained to classify 10 class labels of different objects. Figure 2 shows samples of images from the training set. The model is trained in 25 epochs.

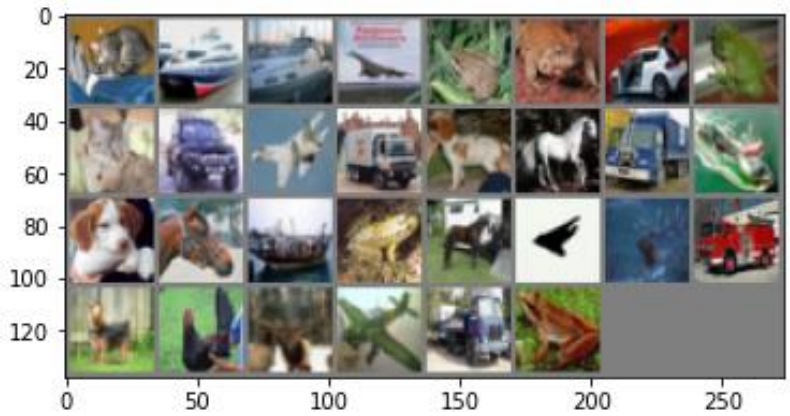

Figure 1 - sample from CIFAR10

5.1.2 Results

The target model overall accuracy on test data is 62% overall with individual accuracies by class label ranging from a minimum of 43% to a maximum of 81%. The model inversion attack is applied on a random test sample of 100 images. The attack model exploits the knowledge of the model and label class. New re-constructed images are created, then tested on the target model to assess the accuracy of the reconstruction. 16% are accurately reconstructed. Figure 3 shows samples of the inversion attack results.

We have also modified the model and added training with Differential Privacy (DP) using the Opacus library. DP addresses privacy risk by incorporating random noise at the training level. We have used DP-SGD (Differentially-Private Stochastic Gradient Descent) which is a modification of the stochastic gradient descent algorithm (Van der Maater et al. 2020). The model is trained in PyTorch through access to its parameter gradients, i.e., the gradients of the loss with respect to each parameter of the model. This access preserves differential privacy of the training data, hence the resulting model is more secure.

The inversion attack is applied to 100 random samples, and the test results in 9% of images accurately reconstructed. Training the model using DP has reduced the risk in the model inversion attacks. There is a known privacy-utility trade off with DP and model performance is scarified.

More experiments need to be applied for fine tuning parameters to obtain optimum results. The main aim of the experiments presented in this section is a demonstration of the most common risks of disclosure of machine learning models.

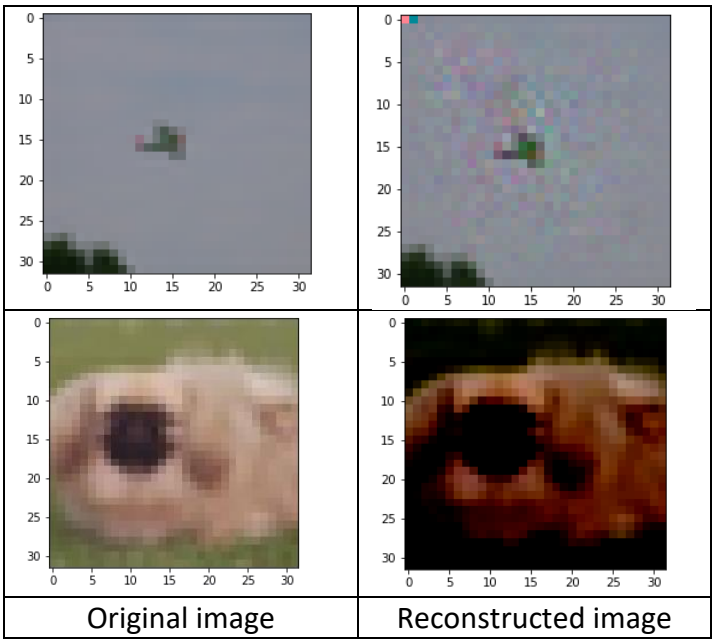

Figure 2 – samples of reconstructed images using model inversion attack

5.2 Case 2: membership inference attack

This second case study describes an example of the second risk of disclosure through machine learning models; that is membership inference attack in a black box fashion. Here, the attack model aims at determining whether a given sample belongs to a training set with black-box access to the model and no prior knowledge of parameters or architecture. Machine learning models are designed for the objective of being able to generalise on unseen data, however the learning process means they tend to perform better on the data they are trained on. The confidence scores on the training samples are always higher than any other input. A Membership inference attack takes advantage of this characteristic to compute the likelihood of membership of a sample and therefore recover some or all of the training sets (He et al 2021). There are other ways of performing membership inference attacks such as using the output of intermediate layers or distance to the decision boundary.
Although these kinds of attacks are possible, they are usually very difficult to apply (depending on the type of machine learning model) due to the complexity of the feature space in some tasks such as image recognition for example. Membership inference attacks are mainly associated with model overfitting (He et al 2021). Overfitting results usually from poor model architecture, training method or the training dataset.

5.2.1 Experiment

In this experiment we use a confidence level attack as it is one of the simplest and most common membership inference attacks (Shokri et al. 2017). Here the attack exploits the confidence level to determine if a sample is a member of the training set. We have used the same target model implemented during the model inversion attack based on Alexnet architecture on CIFAR10 dataset.
The attack model is based on the model presented in (Shokri et al. 2017) with the implementation based on (Bogdan et al . 2018). First the target model is trained on the CIFAR10 dataset. The second step is to train a number of shadow models that will be used for the attack. The attack sends a query to the target attack and obtains a vector probability of 10 values, obtaining a probability for each class. The probability vector is passed to the attack model in addition to the label class. The attacker then infers whether the data point is a member of the training set or not.

5.2.2 Results

The overall accuracy of the target model is 62%. The attack overall accuracy is 56.13% where more than half of the attack attempts were successful in using the confidence probabilities in inferring if a data point belongs to the training set used to train the target model.

6 Responses and metrics

The previous section described some of the most common risks in disclosing machine learning models. We have presented some typical cases studies on neural networks and in particular deep neural networks, where architectures are more complex than other classical machine learning models, and where applying human inspections and disclosure control methods is impossible. In this section we introduce the use of automatic risk assessment methods to help in the disclosure of such models.

6.1 Formal risk measure

Addressing disclosure control for machine learning models is a challenging task due to the ambiguity, multiple parameters and privacy risks from different attacks as described in the previous section. Quantifying risks from machine learning models is impossible using human controls only. Providing TREs and researchers with metrics tools that helps assess the risk factors for models can help quantify and manage risks in disclosure control of such models.
Various machine learning attacks metrics tools have been proposed in the literature such as (Liu et al. 2021), (Murakonda et al. 2020) and (ART toolbox). The main aim of these tools is to provide risk metrics and assess machine learning models robustness against various attacks. (Schwerdtner et al. 2020) have also proposed a framework by applying the definition of risk from the statistical risk theory to machine learning models. The framework is used for risk assessment of deployed models.

In this section, we describe experimentation using one of the assessment tools for risk scoring machine learning models.
We have assessed the usability of ML_Privacy_Meter (Murakonda et al. 2020) and how it can be applied for disclosure control on ML models.

6.1.1 ML_Privacy_tool

The ML_Privacy_tool is a tool to assess robustness of machine learning models by applying various types of attacks such as model inversion attacks or membership attacks and provide metrics for quantifying individual risks in models. The quantification of risks is useful during model development or at the disclosure control point in order to assess and quantify data protection and privacy impacts. The tool helps identify potential privacy risks and appropriate risk mitigation measures.
Functionality of the tool can be summarised as follows

- ML_Privacy_Meter analyses the vulnerability of a machine learning model to membership inference attacks.
- The tool generates attacks on a trained target model assuming black box or white box access to the model.
- White box attacks can exploit the target model parameter's gradients, intermediate layer outputs or prediction of the model to infer training set membership of the input.
- Black box attacks only use the target model predictions to identify membership. The attack is performed by generating an inference model using the target model components which can be exploited for some data and returns the probability of membership of the training set for that data.

6.1.2 Case study

In this section we describe the use of ML_privacy-tool for the assessment of a machine learning model.
We use Alexnet model trained in CIFAR10 dataset as a target model. The attack is defined either as a white box (access to the model architecture, parameters) or a black box (access to results only).
Using the tool we define an attack object as describe in Figure 4. The attack model takes the target model, the training data, the network layers where to apply attacks on, and the number of epochs the attack model is trained. Using the attack initialization, we define the type of attack as well.

```
attackobj = ml_privacy_meter.attacks.meminf.initialize(
            target_train_model=cmodel,
            target_attack_model=cmodel,
            train_datahandler=datahandler,
            attack_datahandler=datahandler,
            optimizer="adam",
            layers_to_exploit = [3,4,5],
            gradients_to_exploit = [5],
            exploit_loss=True,
            exploit_label=True,
            learning_rate=0.0001,
            epochs=100, mode_name='whitebox')

# Begins training the attack model
```

Figure 3 - white box attack parameters

6.1.3 Results

The overall target model's accuracy is 43.84% and 46.42% on the test data. We test two attack models, the black box model accuracy is 74.78% and the white box attack is 74.42%. The attack models' accuracy provides a useful metric on the robustness of the target models. The tool also provides the following metrics:

6.1.3.1 Histogram of the membership probabilities for training set member data and non-member data from the population.

A higher membership probability shows that the model has predicted a higher probability that the data is part of the training data. Figure 5 shows results of the privacy risk probability for both white box and black box attacks. The attack identified around 15% of the training data with a probability or 1 for the black box attack. Whereas it identified 18% of the data as being member of the training set with a probability of 0.8 for the white box attack.

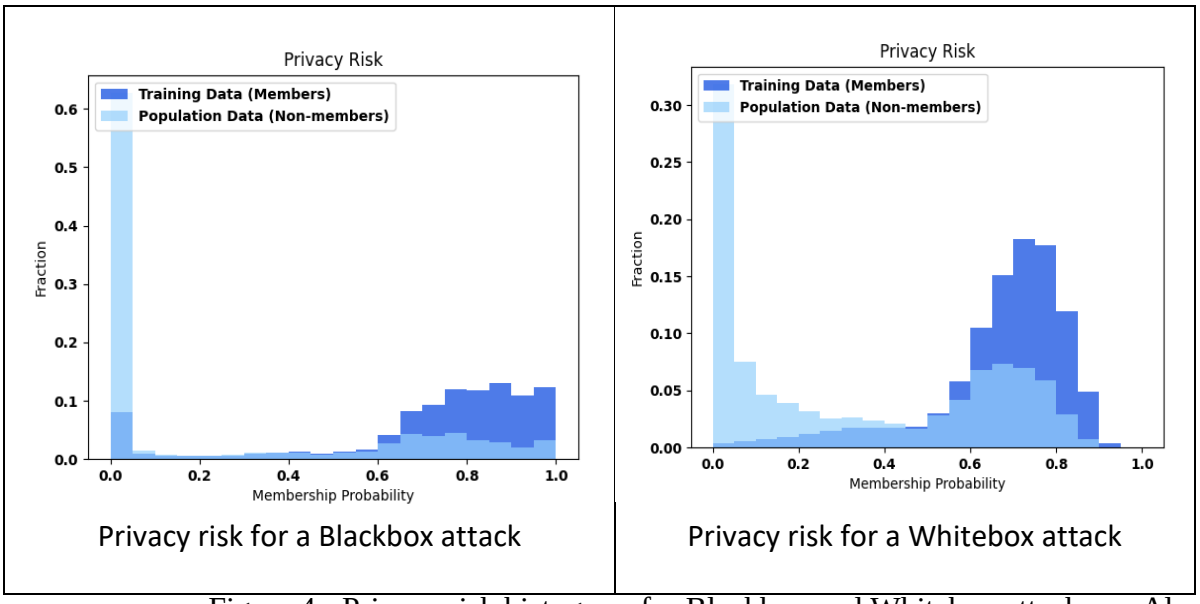

Figure 4 - Privacy risk histogram for Blackbox and Whitebox attacks on AlexNet network

6.1.3.2 Receiver Operating Characteristic (ROC) curve for the membership inference attack.

An attack is successful if it can achieve larger values of True Positive rate and small values of False Positive rate. Success of the attacker can be quantified by an ROC curve representing the trade-off between False Positive Rate and True Positive Rate of the attacker. True positive represents correctly identifying a member as present in the data and False positive refers to identifying a non-member as member. An attack is successful if it can achieve larger values of True Positive rate at small values of False Positive rate. Figure 6 shows the ROC curves for both types of attacks. The area under those curves quantifies the aggregate privacy risk to the data posed by the model. The black box attack area is larger than the white box attack. For this sample target model, the risk of a black box attack is higher than a white box attack.

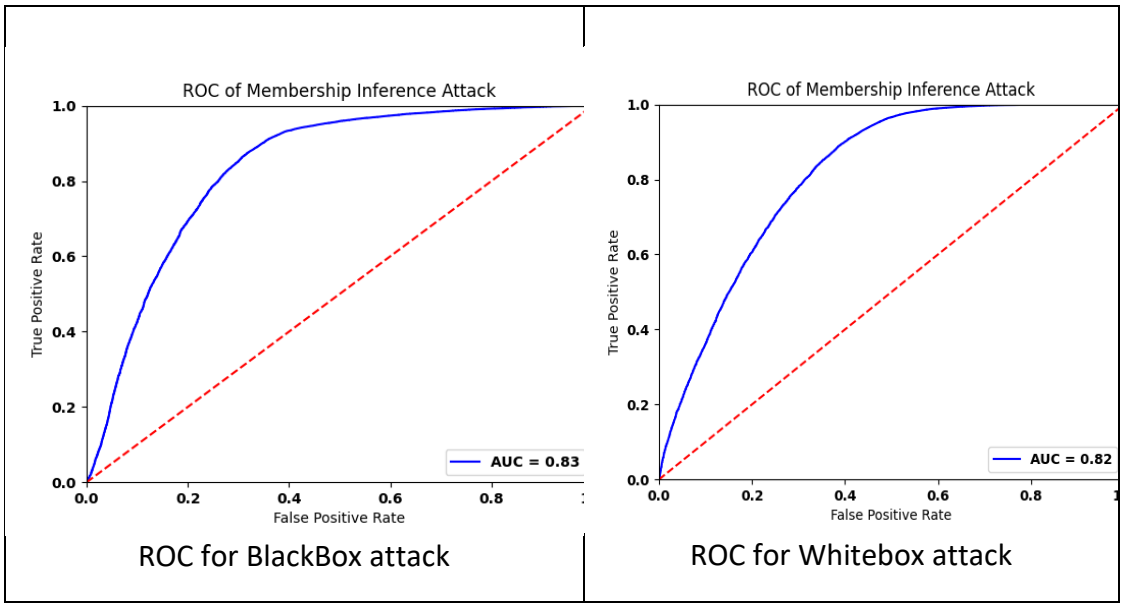

Figure 5 - Receiver Operating Characteristic (ROC) curve for black box and white box attacks

6.1.4 Discussion

This section describes a use case for the ML-Privacy-model. Running the attack metric on a sample target model provides us with an example of the benefits of such tools for assessing risks of disclosure through machine learning models. The tool provides clear metrics and risk probability for each attack and provides useful information about the data privacy risks. In addition to providing clear statistical measure for a model's robustness, the tool can be also used by model owners to adjust different parameters such as differential privacy parameters to reach optimum results. In the context of TREs, the tool can be used by controllers to assess risk measures for disclosure control applications.

7 Recommendations

Unlike conventional SDC process, applying disclosure control on machine learning models is a challenging task and cannot be performed using the existing methods. Machine learning models and architectures are very complex and hard to apply human controls only.

In this section we propose some provisional recommendations and solutions that can constitute a path to SDC disclosure control for machine learning models. We have identified two types of solutions, the first one is data centric, where more controls are applied on the data released to users. The second one is user and model centric, where control and guidelines are applied on users and the machine learning models.

7.1 Data centric solutions

- Watermarking data released to users in order to facilitate leak source tracking and identification when models are released.
- Keeping a held-out dataset to test difference in distribution of predicted probabilities for membership inference (eg 10% of rows are not provided to the researcher but are used by disclosure controllers for a final independent evaluation).
- Developing blunt matching tools that look for straight data 'copies' contained in the model outputs either deliberately or accidentally.

7.2 Model/User centric solutions

- Introduce best practice guideline for the use of data and design architecture of machine learning models. This will help users to design robust and safe algorithms. For neural network models this can include: regularisation, dropout, deferential privacy, model ensemble learning or adding noise to confidence score vectors to avoid membership inference attacks (Yang et al. 2020). Other principles and guidelines can be adopted such as the ones introduced by the Committee of Standards in Public Life.
- Introduce automated metric tools for both TRE controllers and users, such as the tools described in section (Response). These kinds of tools can be designed and customized within TRE environments. They can enable quantification of different risks and vulnerabilities of machine learning models. They can also help users to assess their models internally and build more robust ones. The previous section shows some results in using such tools to predict and assess the risks of attacks in machine learning.
- Introduce the notion of data privacy and security by design. This can be applied with the introduction of explainable AI (Arrieta et al. 2020). Explainable AI are a set of frameworks and machine learning suites that enable machine learning models to be explainable, interpretable by humans, hence auditable. Explainable AI aims at creating safe and private model security by design. Users of TREs can adopt some features of explainable AI such as the following:
    - Rationale: expansibility and description of the process that led to decision made by ML models.
    - Data: Giving clear description of what data have been used in training ML models. This helps the traceability of the data provided to TRE users.
    - Responsibility: Define clear responsibility roles on the models' owners.

- Safety: describing the steps in the design to make models safe and privacy preserving.

8 Conclusion

This paper is a first attempt to structure the ML-SDC problem. We have concentrated on a simple, common case to illustrate some potential issues, and outlined a method of addressing the problem that other researchers may find useful.

In this paper we have presented two case studies representing the most commonly identified risks for machine learning models, focusing on neural networks. The first experiment described a white box attack, where an attacker has access to the model characteristics, architectures, weights and class labels. The second attack describes a black box attack where the attacker has no access to the model's architectures but infers the membership through confidence scores.

The attacks are applied on a common state-of-the-art network architecture of convolutional neural networks, AlexNet. Results show that, although the attack is not 100% accurate, there are still risks of some data being recovered. Risks can be reduced by adopting several good practice guidelines in a robust design of the models.

This is a very early analysis of a large problem with many dimensions and many unknowns. Some do have analogies in traditional SDC: does 'differencing risk' still exist when models are, by construction, non-linear representations of the data? But other unknowns are very fundamental: what counts as disclosure risk when the source data is an object, rather than the single data point being targeted in traditional SDC models?

We have also not considered control measures beyond metrics for risk and the encouragement of ethical practices. User controls are an important consideration for TREs; and statistical controls such as requiring a subset of parameters to be withheld could prove an effective, simple way to manage releases.

This is the first output in a new programme of research being conducted by the authors and others. We welcome contributions from other researchers at the above address.

9 References

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. and Chatila, R., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, pp.82-115.

ART toolbox https://github.com/Trusted-AI/adversarial-robustness-toolbox

Artificial Intelligence and Public Standards A Review by the Committee on Standards in Public Life, 2020,

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/868284/Web_Version_AI_and_Public_Standards.PDF
Aiswariya Milan, K. and Kumar, N.P., 2020. Machine Learning Techniques in Healthcare—A Survey. *Journal of Computational and Theoretical Nanoscience*, *17*(9-10), pp.4276-4279.
Bogdan Kulynych and Mohammad Yaghini. mia: A library for running membership inference attacks against ML models. 2018.
Brandt M., Franconi L., Guerke C., Hundepool A., Lucarelli M., Mol J., Ritchie F., Seri G. and Welpton R. (2010), Guidelines for the checking of output based on microdata research, Final Report of ESSnet Sub-group on Output SDC http://neon.vb.cbs.nl/casc/ESSnet/guidelines_on_outputchecking.pdf
Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A. and Mukhopadhyay, D., 2018. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*.
Green, E., and Ritchie, F. (2016) Data Access Project: Final Report. Australian Department of Social Services. June. http://eprints.uwe.ac.uk/31874/
Hafner H-P., Lenz R., Ritchie F., and Welpton R. (2015) Evidence-based, context-sensitive, user-centred, risk-managed SDC planning: designing data access solutions for scientific use. In: Worksession on Statistical Data Confidentiality 2015, Eurostat
He, Y., Meng, G., Chen, K., Hu, X. and He, J., 2020. Towards Security Threats of Deep Learning Systems: A Survey. *IEEE Transactions on Software Engineering*.
He, X., Wen, R., Wu, Y., Backes, M., Shen, Y. and Zhang, Y., 2021. Node-Level Membership Inference Attacks Against Graph Neural Networks. *arXiv preprint arXiv:2102.05429*.
Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Schulte Nord-holt, E., Seri, G. and De Wolf, P-P. (2010). Handbook on Statistical Disclosure Control. ESSNet SDC. http://neon.vb.cbs.nl/casc/.\SDC_Handbook.pdf
Kaissis, G.A., Makowski, M.R., Rückert, D. and Braren, R.F., 2020. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, *2*(6), pp.305-311.
Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2017. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), pp.84-90.
Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology*. 2017;14:749.
Liu, Y., Wen, R., He, X., Salem, A., Zhang, Z., Backes, M., De Cristofaro, E., Fritz, M. and Zhang, Y., 2021. ML-Doctor: Holistic Risk Assessment of Inference Attacks Against Machine Learning Models. *arXiv preprint arXiv:2102.02551*.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Mode Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In ACM SIGSAC Conference on Computer and Communications Security(CCS),pages1322–1333.ACM,2015. 1,3,4,5,7, 12

Murakonda, S.K. and Shokri, R., 2020. Ml privacy meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning. *arXiv preprint arXiv:2007.09339*.

Nind T, Sutherland J, McAllister G, et al. *An extensible big data software architecture managing a research resource of real-world clinical radiology data linked to other health data from the whole Scottish population*, GigaScience, Volume 9, Issue 10, October 2020.

Opacus, https://github.com/pytorch/opacus

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In 2017 IEEE Symposium on Security and Privacy (SP), pages 3–18. IEEE, 2017.

Ritchie F. (2017) "The 'Five Safes': a framework for planning, designing and evaluating data access solutions". Data For Policy Conference 2017. September. https://zenodo.org/record/897821#.WceTWMZryUk

Ritchie F. and Green E. (2020) "Frameworks, principles and accreditation in modern data management ", Working papers in Economics no 202002. https://uwe-repository.worktribe.com/output/6790882

Rigaki M, Garcia S. A survey of privacy attacks in machine learning. arXiv preprint arXiv:2007.07646. 2020 Jul 15.

Schwerdtner, Paul & Greßner, Florens & Kapoor, Nikhil & Assion, Felix & Sass, René & Günther, Wiebke & Hüger, Fabian & Schlicht, Peter. (2020). Risk Assessment for Machine Learning Models.

SDAP (2019) Statistical Disclosure Control Handbook v1.0. Secure Data Access Professionals. August. https://securedatagroup.org/sdc-handbook/

The Scottish Government. *Charter for Safe Havens in Scotland: Handling Unconsented Data from National Health Service Patient Records to Support Research and Statistics*. 2015 https://www.gov.scot/publications/charter-safe-havens-scotland-handling-unconsented-data-national-health-service-patient-records-support-research-statistics/

van der Maaten, L. and Hannun, A., 2020. The Trade-Offs of Private Prediction. *arXiv preprint arXiv:2007.05089*.

van Timmeren, J., Cester, D., Tanadini-Lang, S. et al. Radiomics in medical imaging—"how-to" guide and critical reflection. *Insights Imaging* 11, 91 (2020).

Yang Z, Shao B, Xuan B, Chang EC, Zhang F. Defending model inversion and membership inference attacks via prediction purification. arXiv preprint arXiv:2005.03915. 2020 May 8.