Machine learning models disclosure from trusted research environments (TRE), challenges and opportunities

Mansouri-Benssassi, Esma; Rogers, Simon; Smith, Jim; Ritchie, Felix; Jefferson, Emily

Publication date:
2021

Machine Learning Models Disclosure from Trusted Research Environments (TRE), Challenges and Opportunities

Esma Mansouri-Benssassi(1)*, Simon Rogers(2)*, Jim Smith(3), Felix, Ritchie(3), Emily Jefferson(1)

(*) These authors contributed equally to this work

(1) University of Dundee

(2) NHS National Services Scotland

(3) University of the West of England

Abstract

Trusted Research environments (TRE)s are safe and secure environments in which researchers can access sensitive data. With the growth and diversity of medical data such as Electronic Health Records (EHR), Medical Imaging and Genomic data, there is an increase in the use of Artificial Intelligence (AI) in general and the subfield of Machine Learning (ML) in particular in the healthcare domain. This generates the desire to disclose new types of outputs from TREs, such as trained machine learning models. Although specific guidelines and policies exists for statistical disclosure controls in TREs, they do not satisfactorily cover these new types of output request.

In this paper, we define some of the challenges around the application and disclosure of machine learning for healthcare within TREs. We describe various vulnerabilities the introduction of AI brings to TREs. We also provide an introduction to the different types and levels of risks associated with the disclosure of trained ML models. We finally describe the new research opportunities in developing and adapting policies and tools for safely disclosing machine learning outputs from TREs.

Keywords: TREs , AI, Machine Learning, Data Privacy

INTRODUCTION

Trusted Research environments (TRE)s are secure and safe platforms enabling access to and analysis of sensitive data, [1]. TREs must follow specific principles when providing researchers with access to data, defined around five distinct areas as outlined by [2]. These are:

- Safe people: Individuals allowed to access data through the TRE must sign a term of use committing them to follow certain guidelines and rules. Users agree not to re-identify individuals and not to share their TRE access with non-authorised persons. They also commit to reporting any safety or security weakness identified within TREs.
- Safe projects: TRE operators need to ensure the projects' use of data is appropriate and in line with public benefits.
- Safe settings: sensitive data is held only within a secure environment, accessed through restricted areas.
- Safe outputs: TREs must set up and implement process to only allow appropriate output data to be released.
- Safe data: In general, TREs operate with highly sensitive/identifiable data: the other elements (people, projects, settings and outputs) provide a high degree of protection so that even the most sensitive data can be analysed lawfully and ethically

TRE output guidelines are typically aimed at specific outputs such as aggregated results, graphs, and tables. With the increased interest in the application of AI and machine learning (ML) on sensitive data such as healthcare and medical data, new guidelines, systems, and platforms are required within TREs to ensure that the above principles are preserved and to mitigate any risk of direct or indirect data privacy breach from disclosure of trained models,. We have interviewed 14 UK and 6 international TREs, [3] to

discover current processes within TREs for AI algorithms disclosure. We have discovered that across the board TREs did not have mature processes, tools or an understanding of disclosure control for AI algorithms. The only processes consist of manual checking.

We present an overview of ML and its applications in healthcare, outline challenges faced by TREs in releasing trained machine learning models and describe potential risks linked to those challenges.

This paper is organized as follows:

- Introduction to ML models and its applications in medical and healthcare
- Identification of new types of outputs generated through machine learning
- Outline of the security threats and vulnerabilities introduced by different types of machine learning models
- Identification of potential risks and risk levels associated with the disclosure of machine learning models
- Identification of opportunities for research to answer some of these challenges.

MACHINE LEARNING IN HEALTHCARE TREs

With the recent surge in the availability of data and enhances in the field there has been an increase in the development and adoption of AI, and in particular ML, in the medical and healthcare fields.

ML is a subset of Artificial Intelligence, that automatically *learns* patterns from datasets. It can be used to help humans better understand complex data, or make predictions based upon new, unseen data. ML is now a mature field, but in recent years particular technological advances, coupled with the increasing availability of very large datasets has seen a surge in popularity. Modern ML encompasses many types of learning and computational algorithms, [3].

ML models can be categorised based on the type of problem they are attempting to solve, summarized below:

Supervised learning:

When performing supervised learning, algorithms are supplied with a corpus of training examples, each associated with an output category or value (a *label*). The goal is to learn a mapping between the input and output. This type of learning is one of the most common and has applications such as the classification and diagnosis of medical conditions, [4] and predicting risks of future health events. The output is typically either one of a number of distinct categories (e.g. case vs controls; known as classification), or a real value (regression). Supervised learning has been used with a wide range of healthcare and medical data, such as medical images, [5], Electronic Health Records (EHR), [6] and genomic data, [6] .

Unsupervised learning:

In unsupervised learning, the model uses unlabeled data to discover patterns in a dataset. One of the most common examples of unsupervised learning is clustering, [7]. The goal of clustering is to group data instances into clusters such that the instances sharing a cluster are similar with one another. Unsupervised learning techniques such as clustering are often used for exploratory data analysis in healthcare and medical applications. For example, [8] used k-means clustering for identifying several subtypes of Alzheimer's using Electronic Health Records (EHR). More recently unsupervised learning was used to learn appropriate features for COVID-19 diagnosis from CT medical imaging, [9] .

Semi-supervised learning:

Semi-supervised learning methods are applied to datasets in which both labelled and unlabeled examples co-exist, with the volume of unlabeled data typically exceeding that of the labeled data. This scenario is common when the cost of labeling data is high which is common in the healthcare and

medical domains, [10]. Semi-supervised methods use patterns present in the unlabeled data to improve performance over models trained on the labeled data alone. One type of semi-supervised learning is Active Learning which starts with of small set of labelled data and works iteratively, [11].

Reinforcement learning:

Reinforcement learning (RL) involves training machine learning models to make decisions sequentially, based upon periodically received feedback, [12]. RL has had high profile success in learning game strategies, where rewards are received at the end of a (variable length) game, [12]. There are various applications of reinforcement learning in healthcare, such as in medical imaging, [13], diagnosis systems, [14] and precision medicine, [15].

MACHINE LEARNING MODEL TYPES

Machine learning models can be grouped into various categories according to their purpose, algorithm type and how they work and learn, [3]. Table 1 summarizes the different and most common types of machine learning models categorized by the type of algorithms they use, the learning category and how they handle data.

Learning type	**Algorithm type**	**Definition**	**Models**	**Examples**	**Reference**
Supervised	Traditional regression / parametric based methods.	Aim at modelling a link between input variables and an output. This enables the prediction of outputs for new examples. Typically, a single parameter needs to be learnt per input feature.	Linear Regression Logistic Regression Stepwise Regression Multivariate Adaptive Regression	Medical images landmark detection	[16]
Supervised	Instance / non-parametric based	These methods use similarity between data instances.	K-nearest Neighbor (KNN)	Decision making in mental health	[17]

	classification methods.	Predictions are made by summarizing the outputs for training examples. Weights are increased for data that is more similar to new observation.	Self-Organising Map (SOM) Support Vector Machines SVM	system using SOM	
Supervised	Tree-based Algorithms	Methods that make decisions based on traversing a tree. The value of features at each node is used to decide the direction of the decision. Typically, recursive search methods are used to successively grow trees, or select nodes.	Classification and Decision Tree Chi-squared Automatic Interaction Detection (CHAID) Conditional Decision Trees	Analysis of adverse drug reaction using CHAID	[18]
Supervised	Bayesian Algorithms	Probabilistic algorithms based upon Bayesian statistical principles. These methods operate by using data to update prior probabilities (of e.g. class membership) to posterior probabilities.	Naïve Bayes Bayesian Belief Network (BBN) Gaussian Naïve Networks	Modelling fetal mortality	[19]
Unsupervised	Non-parametric Clustering Algorithms	Group the data observations according to similarity between them. This results in a partition of the observations.	Hierarchical Clustering Kernelised K-means	Alzheimer structural imaging phenotype Detection using hierarchical clustering. Diabetes diagnosis using K-means clustering	[20]

Unsupervised	Parametric Clustering Algorithms	Defines group of data using parametric model. Data is assigned to cluster using the largest prior probability	K-means K-medians Statistical Mixture models Gaussian Mixture Models	Diabetes diagnosis using K-means clustering	[21]
Supervised / Unsupervised	Neural Networks and Deep learning	Neural Networks including Deep learning methods are inspired by biological neural networks. They consist of an input layer, connected to hidden layers, followed by an output later. Each layer consists of a number of neurons. Deep neural networks have architectures with many hidden layers and provide state-of-the-art performance for complex data such as images, text and audio.	Convolution Neural Networks (CNN), Recurrent Neural Networks (RNNs) - for example Long Short-Term Memory networks (LSTM), Auto-Encoders, Deep Belief Network (DBN)	CNN for the classification of skin cancer LSTM for predictive medicine from EHR. Genomic data imputation using auto encoders	[22] [23]
Unsupervised	Dimensionality reduction	Dimensionality reduction algorithms compress high-dimensional data into a low-dimensional representation. This is achieved by preserving all useful structure of the data in an unsupervised way. The resulting low dimensional representations are	Principal Component Analysis (PCA) Multidimentional Scaling (MDS) T-Stochastic Neighbour Embedding(TSNE)	Brain tumour segmentation	[24]

		used for data exploration and visualization. They can also be used as features for other machine learning models			

MACHINE LEARNING PIPELINE

The development of a machine learning model requires various steps, starting from data collection and problem definition as shown in Figure 1. The process consists of choosing relevant data, extracting useful features, before choosing an appropriate machine learning model before evaluation and deployment.

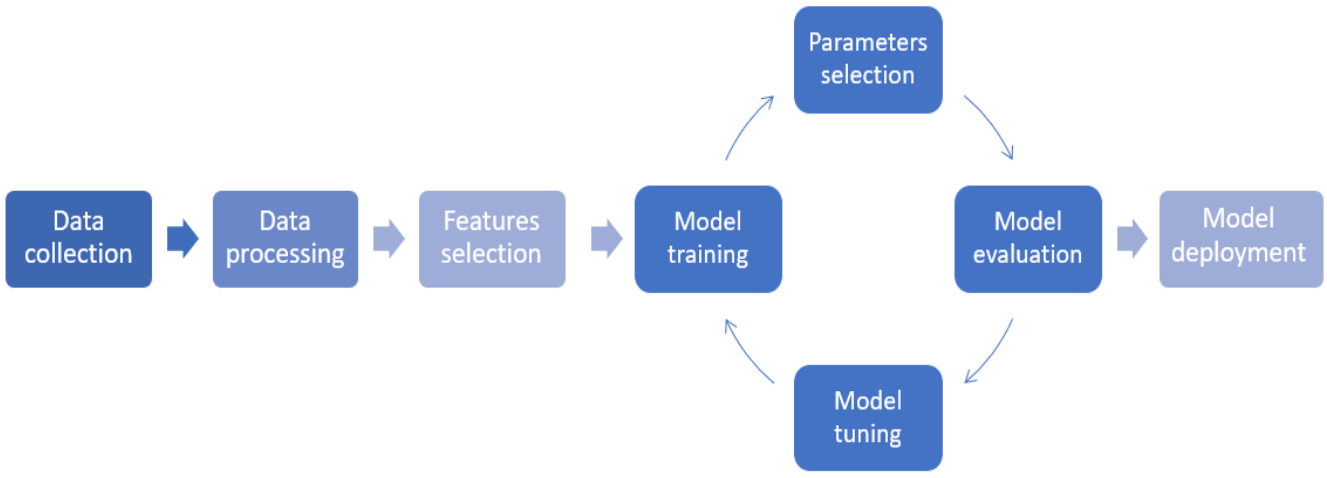

Figure 1 - Machine Learning Model Development Pipeline

All ML models typically have parameters that are optimized via training, as well as hyper-parameters that are typically set via domain knowledge or through trial and error (e.g. setting learning rates). Thus, finding the model with the best estimated performance for a desired task is often an iterative process.

A key step in model development, particularly for supervised models, is validation of the model's performance. Model performance should not be evaluated only on training data but also on a testing set unseen during training. Performance on test data is used to estimate how well the model can generalise to unseen data, and hence how well it might perform once deployed. The validation process also highlights when models have simply memorized the training data rather than learning useful structure, an issue known as overfitting, [27].

MACHINE LEARNING OUTPUT

The main aim in using ML models is to solve specific problems, and for doing so, researchers and data scientists need to deploy models into production systems. This will be done outside TREs and therefore trained models must be disclosed from the TRE.

Trained ML models are often saved by serializing their state to a file to be reused in production or research. Saved ML model files usually include the following:

Architecture:

This is most relevant in neural networks (and therefore deep learning) models where the architecture defines the number of nodes in each layer, how the nodes are connected, and the activation functions used in each layer. Knowledge of the architecture is useful to a potential attacker (e.g. knowledge of the structure of the input layer in a neural network is required to make use of the model). In our definitions

we consider architecture to be set *before* training so, e.g., the definitions of the nodes (features and thresholds) in a decision tree are *parameters* rather than *architecture*.

Parameters:

A model's parameters are those variables within a model that are optimised during training. In the example of a logistic regression, these are the weights that each input variable is multiplied by in the decision function. For a neural network, these are the weights on the connections between neurons.

Configuration:

Many models save the details of any additional parameters that were set by the user for training, possibly including some information regarding the dataset that was used for training, for example variable names.

Optimizer and its state:

Models that require optimisation in the training phase will often store the state of the optimiser at the optimum. This can include which optimizer was used, any associated parameters (e.g. learning rate), how many iterations the optimiser ran for, and why it terminated (convergence, the maximum number of iterations was reached, etc). Saving the optimizer state can be useful if further trained if desired.

Subset of training data:

Some algorithms require training examples to make predictions (e.g. K-nearest neighbours and Support Vector Machines). Where this is the case, these training examples would be saved within the model file.

MACHINE LEARNING MODEL SECURITY THREATS AND VULNERABILITIES

There exist various threats and vulnerabilities introduced by trained ML models regardless of their type of learning or training methods. Threats exist throughout training, production, and deployment. For example, the training phase can be compromised if training sets are poisoned and deployed systems can

be subject to adversarial attacks. Other threats consist of recovery of sensitive data such as membership inference or model inversion attack, [25], [29].

In this section we describe the most common ML attack types that could constitute a privacy risk if successfully applied to a deployed ML model trained within a TRE.

Model membership inference

In a membership inference attack, an attacker attempts to determine whether particular data instances to which they have access belonged to the data used to train the model. This attack model is one of the most popular and was first introduced by, [26].

Membership inference attacks leverage the observation that models will often make more confident predictions on data that they were exposed to in training than previously unseen data. Therefore, high predictive confidence can help to infer the likelihood of a data point belonging to a training set. Overconfidence on training examples is associated with model overfitting, [30] which can be the result of poor model architecture, inappropriate training or too few training examples. This has been demonstrated by (Nasr et al, 2018), who have also experimented on membership inference attacks on federated learning systems, [30].

Model inversion attacks

Model inversion attacks aim at reconstructing part or full training datasets, data labels or both. These attacks can be particularly dangerous for private and confidential data, [27].

Various types of inversion attack have been published in the literature. Fredrikson, [28] was the first to introduce the concept of model inversion by showing how machine learning models could be inverted to leak sensitive data.

Model inversion attacks use model extraction, where attackers attempt to reproduce the parameters or functionality of a model. For example, if an attacker knows that the model is a logistic regression classifier, then they know the structure and, by presenting sufficient diverse inputs to the model and recording the outputs, can construct a series of equations from which the regression weights can be reverse-engineered. If attackers are just interested in mimicking the functionality, they can present many input examples, record the outputs and train a completely new model.

Other types of attacks exist such as property inference, model extraction and adversarial attacks. However, these attacks are out of the scope of this paper as they would not allow the extraction of individually identifiable data.

CHALLENGES IN THE DISCLOSURE OF MODELS FROM TREs

It is clear that the ability to train ML models on the rich data held within TREs could be incredibly beneficial. However, there remain considerable unresolved challenges in the disclosure of trained models from TREs. These challenges include many aspects and are summarized as follows:

Users/actors challenges

Threats around trained ML model disclosure can emerge from different users and actors. We can identify three main actors. The first is a TRE user training models inside a TRE with little awareness of possible threats faced by those models. For example, it is easy to imagine a researcher training a Support Vector Machine (SVM), unaware that the saved SVM has to include a copy of at least part of the training dataset for it to operate, rendering it prone to membership inference attacks and therefore a considerable disclosure risk.

The second type is a malicious user who deliberately hides data inside disclosed models and outputs. To a certain extent, malevolent behavior is guarded against through the existing TRE safeguarding

procedures, [1]. However, these guidelines were designed for aggregated results (tables, plots, or summary statistics), in which the possibility of hiding large quantities of data was insignificant. This changes with the disclosure of ML models, where files being disclosed could be large and not necessarily human readable.

The third actor is an external attacker who has access to trained models after they have been disclosed from the TRE, through either model deployment (e.g. via an accessible application programming interface; API) or model sharing. Attacks can be carried out by these actors in various forms as described in section (Threats and vulnerabilities of ML).

Data challenges

The application of machine learning in healthcare represents more challenges when disclosing models from TRESs. Healthcare data is very heterogenous and unstructured ranging from Electronic Health Records (EHR), medical images to large genomic databases. Within TREs data is safe-guarded and follows strict protocols such as anonymization, pseudonymization or de-identification. However, these methods are not infallible leaving a residual risk of re-identification of individuals. Depending on the data types being considered, TRE operators need to adopt appropriate methods for removing private and personal data. Another challenge related to data, represents the risk to re-identify some anonymized data using different types of information through triangulation.

Challenges around ML models

As described in the previous sections ML models are vulnerable to attacks that can potentially lead to data privacy breaches. Attacks can occur in a *black box* setting, where attackers do not have access to the model itself but only an interface or API. Attackers can therefore *use* the model (present inputs and be provided with outputs) but cannot observe the inner workings of the model. The second type is a

white box attack where attackers have access to the model parameters and architecture. White box attackers can therefore both use and inspect the model, [29]. In general, white box access confers greater risk. However, black box attacks can also represent a significant risk but require more effort to be successful.

For example, someone with white-box access to an SVM could potentially perform a membership inference attack by reading training data directly from the model file. Someone with black-box access would have to rely on more complex optimisation approaches to perform a membership inference attack, with no guarantee of success.

Attacks on deployed models constitute a potential risk for data privacy. Veale et al, [30] present a review of different attacks on machine learning and the possible implication on data protection and GDPR. They have discussed machine learning models and personal data and drew a parallel between pseudonymization and model inversion attacks in the GDPR law.

The challenge of disclosing ML models varies between different models and training regimes. Some models are more prone to attacks than others although model configuration (architecture and setting of model hyper-parameters) also plays a significant role. Disclosing an SVM or a tree-based algorithm may be riskier than disclosing more complex models, [31]. More work is required to better understand the risk of a wide range of models and configurations.

RISKS FACTORS IN DISCLOSING MACHINE LEARNING MODELS

In the previous sections, we have given an overview of machine learning models, their different types, applications in healthcare and their potential vulnerabilities. We have also described challenges faced when disclosing machine learning models from TREs.

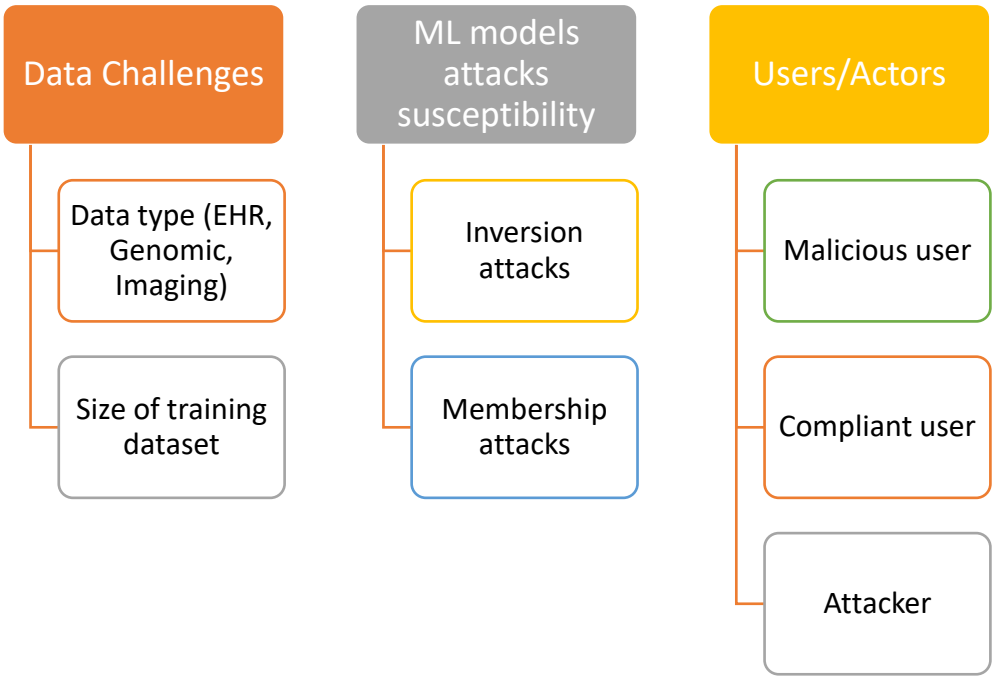

Figure 2 - Machine Learning Challenges Types

It is clear that preserving data privacy when disclosing ML models is challenging. For a model to be effective, it must retain some aspects of the data on which it was trained, [29].

Disclosing ML models from safe environments therefore carries risks that must be accounted for and mitigated. We have identified various factors that may affect the level of risks such as data type, machine learning type or the attack type. Figure 2 summarizes the above-mentioned challenges. To assess the risk of disclosure, TREs need to consider the severity of the different factors (alone or in combination) and the impact of a data breach, which would primarily concern re-identification of individuals and subsequent loss of trust in TREs.

Figure 3 illustrates the potential link between the level of risks and the different factors such as model type, attack type or user type. We have used the strength of the connections to highlight what is, in our opinion, the strength of the relationship. For example, disclosing an SVM model trained on brain MRI data represents a higher risk of re-identification than disclosing a deep learning model using genomic data. Medical images such as brain images can contain hidden biometrics, [32], [33] and pose a higher risk to privacy if reconstructed via a membership inference attack.

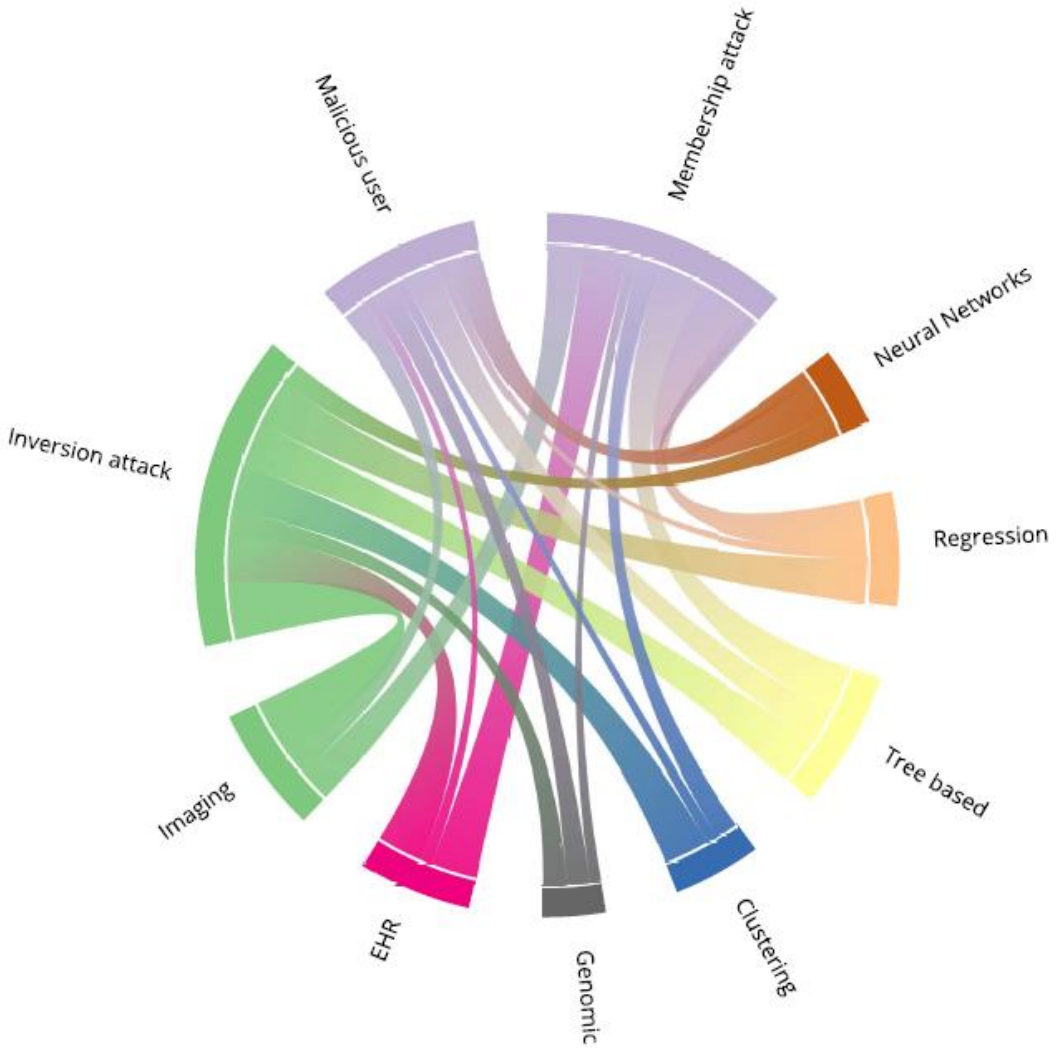

Figure 3 - Relationship between different factors in risk assessment

CONCLUSIONS AND FUTIRE RESEARCH OPPORTUNITIES

In this paper, we have presented an overview of the different types of ML, some of their applications in the healthcare and medical domains, the challenges of disclosing ML models from TREs and finally the introduction of different risk factors associated with the disclosure of trained ML models.

Detailed risk analysis needs to be conducted to quantify risks using the different factors and challenges outlined in this paper. Various risk model frameworks have been investigated in the literature in domains such as financial, privacy or security, [34]. Model frameworks such as FAIR (Factor Analysis of Information Risk), [37] could be applied to conduct a risk assessment of ML models disclosure from TREs. FAIR is a risk management model that divides different aspects of risks into different factors that quantify risk into different probabilities. However, this is not a strictly statistical discussion. The nature of TREs means that multiple measures of control are potentially available, and so a high baseline risk may still be acceptable. This is particularly relevant when considering the motivation of attackers.

The challenges and risks described in this paper create new opportunities for new interdisciplinary research. Investment in the following open areas of research will play a role in ensuring that the next generation of TRE's can facilitate safe disclosure of ML models:

- ***Data privacy research:*** Various opportunities exist in providing more secure processes at the data level. This could include encryption of the data, enabling the application of machine learning algorithm directly on encrypted data. Methods such as homomorphic encryption allow this. Adding digital watermarks enables data tracking that could help detect privacy leaks. ML models could also be trained on synthetic data. However, this method can only be applied in very limited types of data and applications applying synthetic data in the medical field can lead to algorithm bias, [36].

- ***Automatic ML model Privacy assessment*** Quantifying risks from machine learning models is impossible using human controls only. Providing TREs and researchers with tools that helps assess risk factors for models can help quantify, manage and mitigate risks in disclosure of such models. Various machine learning attack metrics tools have been proposed in the literature such as those in, [6].
- ***Automatic output privacy breach detection:*** Providing automatic assessment of the presence of hidden data in output requests, for example using generative models.
- ***Federated learning and privacy preserving by design:*** Providing privacy preserving frameworks for TREs users. There exist various opportunities and research into privacy preserving using federated learning such as the work presented by, [34]. Other research focused on providing privacy preserving platform, [43].

- **AI responsibility and accountability** Exploring areas such as accountability, explainability, fairness from both a technical and legal point of view.
- **Model based mitigation strategies and guidelines:** Develop methodologies for mitigating risks linked to model disclosure such as restricting types of models that can be used. Some model types such as KNN and SVM represent a higher risk and need a tighter control.
- **Effective manual disclosure procedures**: This could include model / code inspections prior to disclosure. This requires experts within the disclosure teams, and the technical infrastructure to generate and store code snapshots for inspection. Ideally, the disclosure team would also use the model, independently evaluating model performance on both the data that was provided to the researcher and a subset of data

that they had never used. For example, in a large dataset, 10% of the examples could be withheld completely from the researcher for use by the disclosure team.

Funding

This project was in part supported by MRC and EPSRC programme grant: InterdisciPlInary Collaboration for efficient and effective Use of clinical images in big data health care Research: PICTURES [grant number MR/S010351/1].

This work was in part supported by Health Data Research UK (HDR UK: 636000/ RA4624) which receives its funding from HDR UK Ltd (HDR-5012) funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation (BHF) and the Wellcome Trust.

This work was in part supported by the Industrial Centre for AI Research in digital Diagnostics (iCAIRD) which is funded by Innovate UK on behalf of UK Research and Innovation (UKRI) [project number: 104690].

REFERENCES

[1] T. Hubbard, G. Reilly, S. Varma and D. Seymour, "Trusted Research Environments (TRE) Green Paper (2.0.0).," Zenodo, 2020.

[2] F. Ritchie, "Five Safes': a framework for planning, designing and evaluating data access solutions," in *Data For Policy Conference*, 2017.

[3] W. Zheng, L. Yan, C. Gou, Z.-C. Zhang, J. J. Zhang, M. Hu and F.-Y. Wang, "Learning to learn by yourself: Unsupervised meta-learning with self-knowledge distillation for COVID-19 diagnosis from pneumonia cases," *International Journal of Intelligent Systems,* 2021.

[4] K. Yu and X. Xie, "Predicting hospital readmission: a joint ensemble-learning model," *IEEE journal of biomedical and health informatics,* vol. 24, p. 447–456, 2019.

[5] X. Ying, "An overview of overfitting and its solutions," in *Journal of Physics: Conference Series*, 2019.

[6] M. Xue, C. Yuan, H. Wu, Y. Zhang and W. Liu, "Machine learning security: Threats, countermeasures, and evaluations," *IEEE Access,* vol. 8, p. 74720–74742, 2020.

[7] L. Wang, Q. Qian, Q. Zhang, J. Wang, W. Cheng and W. Yan, "Classification model on big data in medical diagnosis based on semi-supervised learning," *The Computer Journal,* 2020.

[8] M. Veale, R. Binns and L. Edwards, "Algorithms that remember: model inversion attacks and data protection law," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences,* vol. 376, p. 20180083, 2018.

[9] C. Song, T. Ristenpart and V. Shmatikov, "Machine learning models that remember too much," in *Proceedings of the 2017 ACM SIGSAC Conference on computer and communications security*, 2017.

[10] R. Shokri, M. Stronati, C. Song and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017.

[11] P. M. Shakeel, S. Baskar, V. S. Dhulipala and M. M. Jaber, "Cloud based framework for diagnosis of diabetes mellitus using K-means clustering," *Health information science and systems,* vol. 6, p. 1–7, 2018.

[12] F. Ritchie, "The 'Five Safes': a framework for planning, designing and evaluating data access solutions," 2017.

[13] Y. L. Qiu, H. Zheng and O. Gevaert, "Genomic data imputation with variational auto-encoders," *GigaScience,* vol. 9, p. giaa082, 2020.

[14] T. Pham, T. Tran, D. Phung and S. Venkatesh, "Deepcare: A deep dynamic memory model for predictive medicine," in *Pacific-Asia conference on knowledge discovery and data mining*, 2016.

[15] B. K. Petersen, J. Yang, W. S. Grathwohl, C. Cockrell, C. Santiago, G. An and D. M. Faissol, "Precision medicine as a control problem: Using simulation and deep reinforcement learning to discover adaptive, personalized multi-cytokine therapy for sepsis," *arXiv preprint arXiv:1802.10440,* 2018.

[16] O. Obulesu, M. Mahendra and M. ThrilokReddy, "Machine learning techniques and tools: A survey," in *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2018.

[17] F. Navarro, A. Sekuboyina, D. Waldmannstetter, J. C. Peeken, S. E. Combs and B. H. Menze, "Deep reinforcement learning for organ localization in CT," in *Medical Imaging with Deep Learning*, 2020.

[18] M. Nasr, R. Shokri and A. Houmansadr, "Machine learning with membership privacy using adversarial regularization," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, 2018.

[19] M. Nasr, R. Shokri and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *2019 IEEE symposium on security and privacy (SP)*, 2019.

[20] A. Nait-Ali, "Hidden biometrics: Towards using biosignals and biomedical images for security applications," in *International Workshop on Systems, Signal Processing and their Applications, WOSSPA*, 2011.

[21] R. Mishra and P. Bhanodiya, "A review on steganography and cryptography," in *2015 International Conference on Advances in Computer Engineering and Applications*, 2015.

[22] R. Miotto, F. Wang, S. Wang, X. Jiang and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in bioinformatics,* vol. 19, p. 1236–1246, 2018.

[23] R. Liu, C. A. Mancuso, A. Yannakopoulos, K. A. Johnson and A. Krishnan, "Supervised learning is an accurate method for network-based gene classification," *Bioinformatics,* vol. 36, p. 3457–3465, 2020.

[24] L. Liu, F.-X. Wu, Y.-P. Wang and J. Wang, "Multi-receptive-field CNN for semantic segmentation of medical images," *IEEE Journal of Biomedical and Health Informatics,* vol. 24, p. 3215–3225, 2020.

[25] J. Li, Y. Wang, J. Mao, G. Li and R. Ma, "End-to-end coordinate regression model with attention-guided mechanism for landmark localization in 3D medical images," in *International Workshop on Machine Learning in Medical Imaging*, 2020.

[26] M. Lapan, Deep Reinforcement Learning Hands-On: Apply modern RL methods, with deep Q-networks, value iteration, policy gradients, TRPO, AlphaGo Zero and more, Packt Publishing Ltd, 2018.

[27] M. Kholghi, L. Sitbon, G. Zuccon and A. Nguyen, "Active learning: a step towards automating medical concept extraction," *Journal of the American Medical Informatics Association,* vol. 23, p. 289–296, 2016.

[28] M. Kärkkäinen, M. Prakash, M. Zare, J. Tohka, A. D. N. Initiative and others, "Structural brain imaging phenotypes of mild cognitive impairment (MCI) and Alzheimer's disease (AD) found by hierarchical clustering," *International Journal of Alzheimer's Disease,* vol. 2020, 2020.

[29] H.-C. Kao, K.-F. Tang and E. Chang, "Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[30] T. Kang, K.-I. Oh, J.-J. Lee, B.-S. Park, W. Oh and S.-E. Kim, "Measurement and Analysis of Human Body Channel Response for Biometric Recognition," *IEEE Transactions on Instrumentation and Measurement,* vol. 70, p. 1–12, 2021.

[31] G. Kaissis, A. Ziller, J. Passerat-Palmbach, T. Ryffel, D. Usynin, A. Trask, I. Lima, J. Mancuso, F. Jungmann, M.-M. Steinborn and others, "End-to-end privacy preserving deep learning on multi-institutional medical imaging," *Nature Machine Intelligence,* vol. 3, p. 473–484, 2021.

[32] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen and Y. Wang, "Artificial intelligence in healthcare: past, present and future," *Stroke and vascular neurology,* vol. 2, 2017.

[33] S. Imai, T. Yamada, K. Kasashi, M. Kobayashi and K. Iseki, "Usefulness of a decision tree model for the analysis of adverse drug reactions: Evaluation of a risk prediction model of vancomycin-associated nephrotoxicity constructed using a data mining procedure," *Journal of evaluation in clinical practice,* vol. 23, p. 1240–1246, 2017.

[34] T. Hunt, C. Song, R. Shokri, V. Shmatikov and E. Witchel, "Chiron: Privacy-preserving machine learning as a service," *arXiv preprint arXiv:1803.05961,* 2018.

[35] T. Hubbard, G. Reilly, S. Varma and D. Seymour, *Trusted Research Environments (TRE) Green Paper,* Zenodo, 2020.

[36] O. Gottesman, F. Johansson, M. Komorowski, A. Faisal, D. Sontag, F. Doshi-Velez and L. A. Celi, "Guidelines for reinforcement learning in healthcare," *Nature medicine,* vol. 25, p. 16–18, 2019.

[37] P. Ghosh, S. Azam, M. Jonkman, A. Karim, F. J. M. Shamrat, E. Ignatious, S. Shultana, A. R. Beeravolu and F. De Boer, "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques," *IEEE Access,* vol. 9, p. 19304–19326, 2021.

[38] K. Ganju, Q. Wang, W. Yang, C. A. Gunter and N. Borisov, "Property inference attacks on fully connected neural networks using permutation invariant representations," in *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, 2018.

[39] C. Gambella, B. Ghaddar and J. Naoum-Sawaya, "Optimization problems for machine learning: A survey," *European Journal of Operational Research,* vol. 290, p. 807–828, 2021.

[40] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page and T. Ristenpart, "Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing," in *23rd {USENIX} Security Symposium ({USENIX} Security 14)*, 2014.

[41] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *nature,* vol. 542, p. 115–118, 2017.

[42] A. Coronato, M. Naeem, G. De Pietro and G. Paragliola, "Reinforcement learning for intelligent healthcare applications: A survey," *Artificial Intelligence in Medicine,* vol. 109, p. 101964, 2020.

[43] Y. Chung, L. Salvador-Carulla, J. A. Salinas-Pérez, J. J. Uriarte-Uriarte, A. Iruin-Sanz and C. R. García-Alonso, "Use of the self-organising map network (SOMNet) as a decision support system for regional mental health planning," *Health research policy and systems,* vol. 16, p. 1–17, 2018.

[44] M. E. Celebi and K. Aydin, Unsupervised learning algorithms, Springer, 2016.

[45] I. Campero-Jurado, D. Robles-Camarillo and E. Simancas-Acevedo, "Problems in pregnancy, modeling fetal mortality through the Naïve Bayes classifier.," *International Journal of Combinatorial Optimization Problems & Informatics,* vol. 11, 2020.

[46] K. R. Babu, P. V. Nagajaneyulu and K. S. Prasad, "Brain tumor segmentation of T1w MRI images based on clustering using dimensionality reduction random projection technique," *Current Medical Imaging,* vol. 17, p. 331–341, 2021.

[47] O. M. Araz, T.-M. Choi, D. L. Olson and F. S. Salman, "Role of analytics for operational risk management in the era of big data," *Decision Sciences,* vol. 51, p. 1320–1346, 2020.

[48] N. Alexander, D. C. Alexander, F. Barkhof and S. Denaxas, "Using unsupervised learning to identify clinical subtypes of Alzheimer's disease in electronic health records," *Studies in health technology and informatics,* vol. 270, p. 499–503, 2020.

[49] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *Ieee Access,* vol. 6, p. 14410–14430, 2018.

[50] D. E. Adkins, *Machine learning and electronic health records: A paradigm shift,* Am Psychiatric Assoc, 2017.