Privacy Concerns and Solutions in Machine Learning

Gokula Rani Vallabhu
CIISE. Concordia University
Montreal, Canada
40161606

Sai Chandra Sekhar Reddy Dwarampudi
CIISE. Concordia University
Montreal, Canada
40189233

Preetha Nadipally
CIISE. Concordia University
Montreal, Canada
40193731

Abstract— **A wide number of sectors, including smart healthcare, financial technology, and surveillance systems, are being revolutionized by the usage of machine learning models, which has developed tremendously in recent years. It is now more crucial than ever to safeguard sensitive data against privacy intrusions. The work on protecting privacy with ML is still in its infancy because most existing solutions only handle privacy problems during the machine learning process. Consequently, a full examination of the problems with privacy protection and machine learning is required. Three broad fields are primarily the focus of the paper. Machine learning includes private machine learning, privacy protection assisted by machine learning, machine learning-based privacy assault, and related security measures. The survey offers a summary of the various threat models and assumptions typically used to evaluate the security of machine learning algorithms. Several protection strategies that were used to safeguard the models were also examined.**

Keywords - Secure Multi-Party Computation, Protection Schemes, Generative Adversarial Networks, Artificial Datasets, High Privacy, Neural Networks

I. Introduction

Artificial intelligence has brought in new growth chances along with the quick advancement of science and technology. The functional qualities of artificial intelligence are strengthened by the incorporation of interdisciplinary theoretical knowledge into machine technology that is based on computer technology, such as statistics and algorithm complexity. It is possible to improve the applicability of machine learning algorithms and give more convenience for the economic growth of the industry by conducting a realistic analysis of machine learning algorithms and giving direction reference for later machine learning development. [10].

The Classification of Machine Learning can be described as follows

1. Supervised Learning: Supervised learning is a very simple learning technique used in the process of machine learning. The setting of appropriate learning objectives by individuals prior to learning is referred to as this learning strategy. The machine uses information technology to learn the requirements of learning throughout its first training. We are meant to gradually finish the necessary learning content in a supervised environment to gather fundamental data information. Compared to other learning techniques, supervised learning may fully activate the machine's inherent capacity for generalized learning. After finishing the system learning, it can assist individuals in solving some highly systematic classification or regression issues.[10]
2. Unsupervised learning: Unsupervised learning is a counterpart to supervised learning. During the whole learning process, the machine does not mark the material in a certain direction, as is the case with so-called supervised learning. Instead, it relies on the machine to finish the analysis of the data and information. In practice, this means letting the machine pick up on the fundamental ideas and information before giving it the flexibility to acquire a variety of other ideas and information that are related to the fundamental ideas, like tree roots. The range of machine learning material has generally risen due to learning that is continuously improved via phases. Now, deep belief networks and autoencoders are examples of unsupervised learning algorithms.[10]
3. Reinforcement Learning: There are application techniques for reinforcement learning in machine learning in addition to supervised learning and unsupervised learning. The systematic learning of a particular topic is what is referred to as reinforcement learning. The information gathered throughout the preceding time will be applied in the specific application procedure. It compiles and organizes the feedback data from a specific component to create a closed data processing loop. In general, the data gathering based on statistics and dynamic learning is expanded using reinforcement learning, a sort of learning methodology. These techniques are mostly employed to address the issue of robot control. The Q-learning algorithm and the Temporal Difference Learning Algorithm are examples of its representative learning techniques.

The widespread use of machine learning (ML) models has brought about a new era of privacy concerns. While ML can serve as a powerful tool for privacy protection, it can also be used as an attack tool to extract sensitive information. The emergence of deep learning techniques has made this even more apparent, as it enables the automatic collection and processing of millions of photos and videos to extract private information from social networks. This poses a challenge to traditional privacy-preserving methods, which may not be sufficient to combat deep learning-based attacks. As such, there is a pressing need to revisit the definition and scope of privacy and to develop new solutions to address the new threats posed by deep learning-based attacks.[2]

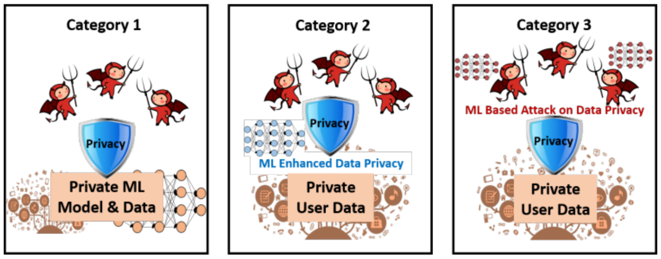


Figure 1: Various categories of problems in privacy and ML [1]

This report provides a comprehensive overview of the current state of research on privacy attacks and protection schemes in machine learning. The insights provided in the paper can be used to guide the development of more robust and secure machine-learning models

II. ML-BASED PRIVACY

Social networks have become one of the riskiest sources of personal information leakage. Although social network platforms have enriched people's interactivity and relationships, shared posts, including check-ins, activities, thoughts (tweets, status updates, etc.), pictures, and videos, often contain sensitive information that poses high privacy risks. Users may unintentionally hand over their privacy by sharing such information. Moreover, [3] companies and start-ups are increasingly specializing in analyzing shared pictures on social media to exploit them for commercial purposes or selling them to other companies. As a result, advanced deep neural networks (DNNs) have been used to launch privacy attacks on these platforms. The increasing use of DNNs has made it more challenging to protect users' privacy on social networks, and traditional privacy-preserving methods may not be sufficient to defend against these attacks. The figure 2 depicts the attacks and protection schemes of private machine learning.

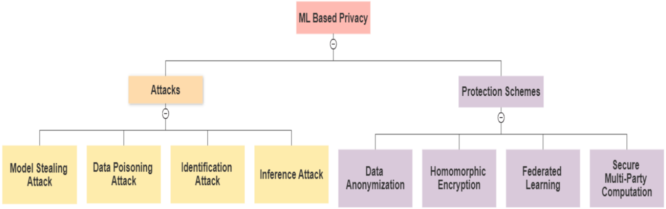


Figure 2: Attacks and Protection Schemes of ML based Privacy

A. Model Stealing Attack

This attack aims to steal the parameters of a machine learning model by querying it. The attacker can use queries to the model to construct a copy of the model. Protection schemes against model stealing attacks include watermarking, obfuscation, and model aggregation. The process is as follows

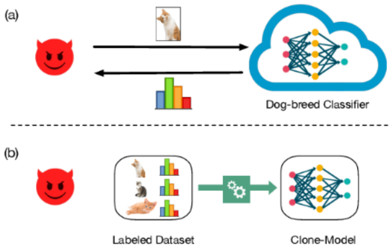


Figure 3: Model Stealing Attack [4]

- Selecting the Target Model: Choosing the target machine learning model is the initial stage in a model theft assault. The attacker can select the target model by doing a web search or by looking at how it is used in a specific application.
- Making a Substitute Model: The attacker then makes a substitute model that is trained using either a dataset that is available to the public or one that the attacker has amassed. The alternative model is taught to closely resemble the target model's behavior
- Querying the Target Model: Input queries are sent to the target model by the attacker, who then gathers the output predictions. The attacker can produce a variety of queries by picking inputs at random or by employing an active learning strategy.
- Training the Substitute Model: Using the input-output pairs gathered from the target model, the attacker trains the substitute model.
- Refinement of the Substitute Model: To increase the accuracy of the substitute model, the attacker might repeat the earlier processes or employ additional strategies.
- Application of the Stolen Model: The attacker can then employ the stolen model for their own gain, such as in a competitive commercial setting, or for evil deeds like fraud or avoiding malware detection.

B. Data Poisoning Attack

This attack aims to manipulate the training data of a machine learning model to cause it to make incorrect predictions. The attacker can inject malicious data into the training dataset.

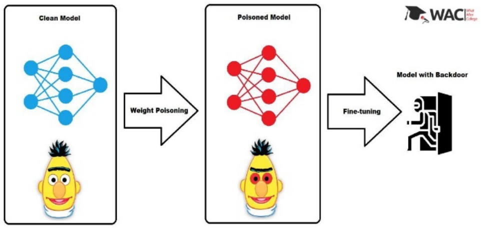


Figure 4: Data Poisoning Attack[4]

Protection schemes against data poisoning attacks include data sanitization and secure multi-party computation.

- Obtaining Training Data: The assailant next acquires or produces a collection of training data that is comparable to the original training data used to train the target model. With the intention of replacing the current model, this data will be utilized to train a new one.
- Introducing Poisoned Data: The attacker inserts poisoned data into the training set. When a model is being trained, poisoned data is purposefully created to deceive it and cause it to notice incorrect patterns or relationships.
- Retraining the Model: Using the tainted training data, the attacker retrains the model. This new model is built to consider the wrong correlations and patterns discovered from the tainted data, which might lead to inaccurate predictions or judgements when applied.
- Poisoned Model Deployment: The attacker introduces the poisoned model either by replacing the original model or by replacing it with the poisoned model. This stage aims to accomplish the attacker's objectives by using the poisoned model to make bad predictions or judgements in a crucial application.

C. *Identification Attack*

Identification attacks are a type of privacy attack that aims to identify the individuals or attributes represented in a dataset. Identification attacks are a significant threat to privacy, as they can reveal sensitive information about individuals, such as their medical history, financial information, and social connections. Machine learning models and algorithms can be used to carry out identification attacks, making them a significant concern in the era of big data and artificial intelligence.[6]

Re-identification attacks aim to identify individuals in a dataset that has been anonymized to protect their privacy. This type of attack is carried out by linking records in the anonymized dataset to records in other datasets that contain identifying information. For example, a re-identification attack may link an anonymized medical dataset to a publicly available voter registration dataset to identify the individuals in the medical dataset.

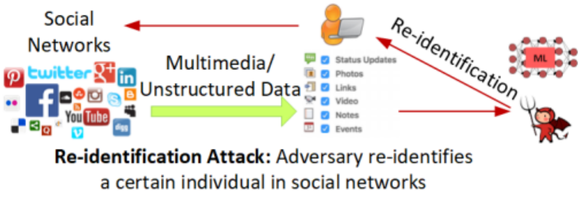


Figure 5: Identification Attack [1]

Impact of Identification Attacks on Privacy: Identification attacks can have a severe impact on privacy, as they can reveal sensitive information about individuals. For example, a re-identification attack on a medical dataset can reveal sensitive medical information about individuals, such as their diagnosis and treatment. Similarly, a linkage attack on a financial dataset can reveal sensitive financial information about individuals, such as their income and spending habits. Inference attacks can also reveal sensitive information about individuals, such as their political beliefs and sexual orientation. [6][7]

The impact of identification attacks on privacy extends beyond individuals to groups and communities. For example, a linkage attack on a social media dataset can reveal sensitive information about a particular group or community, such as their political beliefs or religious affiliations. This can lead to discrimination and stigmatization of that group or community.

D. *Inference Attack*

Inference attack is a type of privacy attack that aims to infer sensitive information about individuals from data that does not explicitly reveal that information. This type of attack is carried out by exploiting patterns in the data to draw conclusions about individuals.

Inference attacks are particularly concerning in the era of big data and artificial intelligence, as machine learning algorithms can be used to uncover patterns in data that are not immediately apparent to humans. These algorithms can use a variety of techniques to infer sensitive information about individuals, such as classification, clustering, and regression.

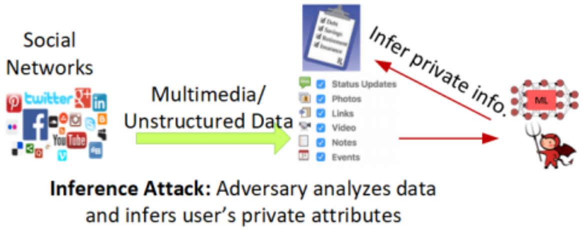


Figure 6: Inference Attack [1]

The impact of inference attacks on privacy can be severe, as they can reveal sensitive information about individuals that they may not want to disclose. Inference attacks can also lead to discrimination and stigmatization of certain groups or communities.

PROTECTION SCHEMES

Adversarial perturbation-based methods are a type of attack that targets machine learning models by manipulating the input data to cause misclassification. These attacks involve adding small, carefully crafted perturbations to input data that are often imperceptible to humans but can cause significant changes in the model's output.[9]

Adversarial perturbations can be generated using various methods, including fast gradient sign method (FGSM), iterative gradient sign method (IGSM), and projected gradient descent (PGD). These methods differ in how they generate perturbations and their effectiveness in fooling the model.

Adversarial perturbations can have significant implications, especially in safety-critical applications such as autonomous vehicles and medical diagnosis. An attacker can manipulate input data to cause the model to make incorrect decisions that can have severe consequences. Additionally, adversarial perturbations can be used to extract sensitive information from models, leading to privacy breaches.

To defend against adversarial perturbations, several defense methods can be used, including adversarial training, defensive distillation, randomization. Adversarial perturbation-based methods are a significant threat to machine learning models and can have severe consequences in safety-critical applications. To defend against these attacks, several defense methods can be used, including adversarial training, defensive distillation, and randomization. However, the arms race between attackers and defenders continues, and more research is needed to develop more robust defense mechanisms against adversarial perturbations.[9]

To Protect against Identification Attacks, there are several techniques used to protect against identification attacks, including data anonymization, differential privacy, and secure multi-party computation.

- Data Anonymization: Data anonymization is a technique used to remove identifying information from a dataset to protect privacy. This technique involves removing or masking identifying attributes, such as name, address, and date of birth, from the dataset. However, data anonymization is not foolproof, as re-identification attacks can still be carried out on anonymized datasets.
- Differential Privacy: Differential privacy is a technique used to protect privacy by adding noise to the data. This technique ensures that the output of a machine learning model is not influenced by any individual in the dataset. Differential privacy is particularly effective against inference attacks, as it ensures that individual attributes cannot be inferred from the data.
- Secure Multi-Party Computation: Secure multi-party computation is a technique used to protect privacy by allowing multiple parties to jointly compute a function on their data without revealing their data to each other. This technique ensures that no single party has access to the complete dataset, protecting against linkage attacks and re-identification attacks.

To protect against inference attacks, several techniques can be used, including differential privacy, k-anonymity, l-diversity, inference attacks are a significant threat to privacy in the era of big data and artificial intelligence. These attacks can reveal sensitive information about individuals that they may not want to disclose and can lead to discrimination and stigmatization of certain groups or communities. To protect against inference attacks, several techniques can be used, including differential privacy, k-anonymity, and l-diversity.

III. Private Machine Learning

Private machine learning is a new field that deals with the development of machine learning techniques that can work on sensitive data without compromising the data's privacy. Private machine-learning methods are now more necessary than ever due to the rising frequency of data breaches and privacy intrusions. Large datasets are typically needed for machine learning algorithms' testing and training. This data is frequently gathered and kept in a single location, which puts the data's privacy in danger. Additionally, gathering and centralizing the data can be a costly and time-consuming procedure.

Private machine-learning algorithms address these issues by maintaining data privacy while allowing the data to remain dispersed. Differential privacy, federated learning, and secure multi-party computation are a few of the methods that are frequently applied in private machine learning.

Differential privacy is a strategy that includes saturating the data with noise before the machine learning system processes it. Even if an attacker can view the algorithm's output, this technique guarantees that the data will stay private. Differential privacy is frequently utilized in the creation of private machine-learning approaches because it offers a high level of privacy protection while maintaining the algorithm's accuracy.

Another method that is frequently applied in private machine learning is federated learning. A machine learning algorithm is trained via federated learning on various devices, each of which holds a different portion of the data. Each device does the training locally, and the results are pooled centrally. By using this method, the data's decentralization and privacy are both maintained.[3]

One other method utilized in private machine learning is secure multi-party computation. With this method, several participants work together to calculate a function while keeping their individual contributions a secret. When data is dispersed among several parties and data centralization is not practical, this strategy is especially helpful.

Private machine learning has several uses, notably in the fields of governance, finance, and healthcare. Private machine learning can be applied to healthcare to create predictive models that can aid in disease diagnosis and offer specialized treatment alternatives. Private machine learning can be used in finance to create fraud detection algorithms that can spot questionable transactions without jeopardizing the data's privacy. Government agencies can employ private machine learning to create algorithms that can detect potential security issues while protecting the confidentiality of the data.[3]

Now, differential privacy algorithms and their many improvement strategies are mostly used to secure the privacy of machine learning. Researchers primarily focus on three components of differential privacy improvement, which are based on gradient, function, and label, correspondingly. Whatever the differential privacy algorithm, its fundamental goal is to introduce noise to the machine learning process to disturb the neural network's recollection of actual training data.

To manage the changes in the gradient descent process and achieve privacy protection, Abadi et al. introduced the DP-SGD technique, which is not suited for the convergence of complicated models. It calculates the reliance of neural network weight parameters on training data. The DP-GAN approach was developed by Xie et al. It uses sensitive data like noise mitigation data to contribute to the gradient, which is then derived using the Wasserstein distance. Because this method depends on generators to produce high-quality training data points, it struggles with complicated datasets.

To offer the best perturbation parameters, Phan et al. coupled a deep encoder with several privacy strategies, including a global sensitivity processing layer to the encoder built on a gradient descent approach. The model parameters were then fine-tuned using a back-propagation algorithm. This approach adds a second network layer without integrating training data, and the detrimental impact is particularly pronounced in large datasets. They presented an improved approach called ADLM, which introduces neural network characteristics to the optimization problem during training to dynamically change the noise distribution by raising the noise on neurons that are weakly linked with the output. The performance of the model is increased to 84.8 percent when this method is implemented in the CIFAR-10 data, which is 14 percent higher than the DP-SGD algorithm, although the efficiency is still not perfect. [4]

A deep learning method that incorporates the information transfer technique in semi-supervised was developed by Papernot et al. This method aggregates the prediction outcomes of numerous instructors by polling and adding noise further to monitor the training of models. It employs disparate datasets to train numerous models. This training strategy can produce relatively accurate models and is effective at protecting privacy. However, because the accuracy of the teacher model's prediction parameters determines the student model's training, to make numerous teacher models exceedingly accurate, the latter takes a substantial quantity of time for data training, which really is unquestionably bad for difficult jobs. It can be difficult to design noise during the aggregation process for various datasets. [4]

The two key elements of machine learning (ML), the model and the data, relate to two distinct kinds of privacy attack targets.

A. *Training Data Privacy*

Datasets used to train machine learning algorithms contain sensitive information that must be protected. This is referred to as training data privacy. Large volumes of data, including sensitive information like personally identifying information, financial information, or medical records, are necessary for machine learning algorithms to learn and make predictions. To keep this information from being abused or ending up in the wrong hands, privacy protection is essential.

When utilizing an ML service, a user frequently wishes to maintain the training data private. For instance, a hospital might create a model for a scientific study using the private medical information of a few individuals. The hospital may wish to use the model to forecast the likelihood of readmission, while a patient may need to use it to determine whether they will develop a specific disease. These instances involve sensitive medical characteristics in the training data, which should not be shared. In other fields, including financial records, there are similar situations. More precisely, training data privacy covers membership, specific traits, statistical aspects, and exact data value. [2]

B. *Model Privacy*

Model privacy relates to the safeguarding of private data kept inside machine learning models. Machine learning models could include sensitive data, including confidential information, trade secrets, and proprietary algorithms. To prevent these models from being taken, duplicated, or modified without permission, their privacy must be protected.

Concerns over privacy also exist regarding the ML model's training procedures and model parameters. A financial organization, for instance, might possess a sensitive model that can precisely forecast stock values or insurance rates. The concept is a significant piece of intellectual and commercial property. Another illustration is the existing commercial Machine Learning API service provided by businesses like Google, Amazon, IBM, and others. Customers are charged for each API access. Revenue will be lost if its models or algorithms are disclosed. In conclusion, the developed model or variables may be the assault target. Depending on how they get the information, the opponents have varying degrees of expertise. The figure 7 depicts the attacks and protection schemes of private machine learning. [2][1]

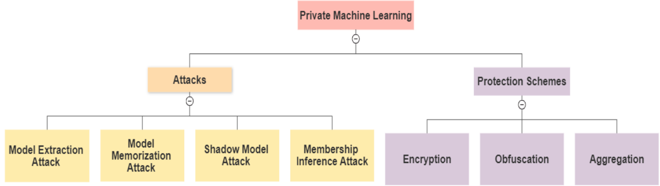


Figure 7: Attacks and Protection Schemes of Private Machine Learning

A. *Model Extraction Attack*

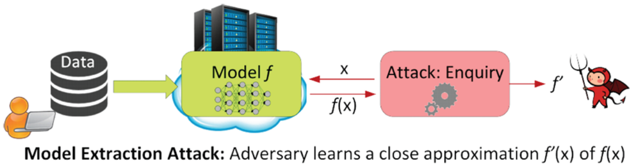

Data
Model *f*
x
f(x)
Attack: Enquiry
f'
Model Extraction Attack: Adversary learns a close approximation *f'*(x) of *f*(x)

Figure 8: Model Extraction Attack Illustration [1]

Figure 8 illustrates a model extraction attack where an adversary can get information from a victim model through query access and use it to train a different model to mimic the functionality of the target model.

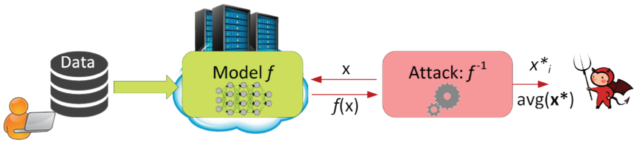
Model extraction attack, which seeks to steal a machine learning model, has drawn a lot of attention from the academic community because of the accessibility of machine learning as a service (MLaaS). The two classes of attacks that fall under this category are accuracy model extraction and fidelity model extraction. Tramèr et al. were the first to suggest accuracy model extraction, with the goal of the attack being to achieve comparable or higher performance on the test dataset for the extracted model. Since then, numerous techniques have been created to decrease the volume of queries, such as model extraction utilizing active learning or semi-supervised learning.[5]

fidelity model extraction calls for the attack model to accurately recreate target model predictions, including target model errors. Typical studies include cryptanalytic extraction, functionally equivalent extraction, and model reconstruction using model explanation. There are various works on model extraction for natural language processing in addition to model extraction attacks on images. Model extraction attacks are conducted by Krishna et al. against BERT-based models, and the extracted model performs marginally worse than the target model

B. *Feature Estimation Attack*

A feature estimation attack seeks to estimate certain features, such as the avg(x*) of the training dataset, or statistical qualities like them. It can be carried out using a model inversion attack, a shadow model assault, or a power side-channel attack in practice. Although it can use a black-box assault with less effectiveness, Model Inversion Attack primarily operates in a white-box model. A white-box assault by Fredrikson et al. can "learn crucial genomic information about individuals." The fundamental concept is to fill in the target feature vector "with each of the number of values and then calculate a weighted probability score that this is the target value," given the information of the linear regression model f. Following that, they expanded the attack to include facial recognition models to achieve two distinct goals: the reconstruction attack, which creates "an image of a person associated with a particular label," and the deblurring attack, which creates the deblurred picture of a specific person given "an image containing a blurred-out face.[4]

The purpose of these assaults is to minimize a cost function using f' using gradient descent (GD). Thus, the model inversion attack operates under a straightforward guiding principle: we can alter the weights and acquire the characteristics for all categories in the network by reverse-engineering by observing the gradients in a trained network. We can nevertheless replicate the prototype example for classes for which we have no prior knowledge. This kind of attack raises the possibility that any reliable deep learning system, irrespective of training techniques, may leak data on the distinct classes. Numerous studies have demonstrated that sample data generated by generative adversarial networks (GANs) are like training data. So, in comparison to average samples, the findings of the model inversion attack can potentially "reveal more personal data about the training data". [4]

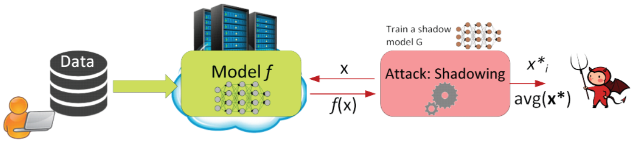


Model Inversion Attack: Adversary learns certain features $x^*_i \in \mathbf{x}^*$ or statistical properties such as avg($\mathbf{x}^*$) of the training dataset

Figure 9: Model Inversion Attack Illustration [1]

Shadow model assault refers to the training of additional ML models by the attacker to reach the target. Either a black-box or a white-box scenario is possible. To extract accent information from trained voice recognition systems, to extract patterns or individual data within the training set, Ateniese et al. developed a meta-classifier that could be trained to penetrate other ML classifiers.

An attack was created by Hitaj et al. from the perspective of group learning. They view the adversary as a participant in the process of collaborative learning who seeks to extrapolate private knowledge from their peers. It is a white-box assault since the attacker can see using the model's internal parameters. "This approach is comparable to facial composites imaging used by police to identify criminals, where the composite artist makes drawings based on eyewitness of the suspect's face, where the adversary utilizes GANs to recover and rebuild information of the victim. The final image is based on eyewitness feedback, even though the composite artist (GAN) had never even seen a real face. [3]

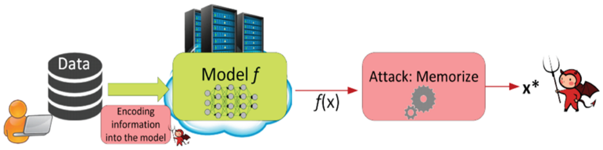


Shadow Model Attack: Adversary learns certain features or statistical properties of the training dataset, with the help of the shadow model G

Figure 10: Shadow Model Attack Illustration [1]

C. *Model Memorization Attack*

The model memory attack, which aims to retrieve the precise attribute values on individual samples, was initially put forth by Song et al. They identify themselves as a "malicious ML supplier" who specializes in building models for clients. The provider has knowledge of the final model but is not present during the training in a business model. He can take sensitive samples and encrypt the data into the model outputs or parameters. During model serving, a malicious third party may be able to extract private data from the model. Figure 5 depicts such an attack.



Model Memorization Attack: Adversary recovers exact feature values $\mathbf{x}^*$

Figure 11: Model Memorization Attack [1]

Both white-box and black-box instances are susceptible to model memorization attacks. In the white-box instance, Song et al. suggested several methods for the attacker to encrypt private information in the models. [1]

(1) Lower-significant bits (LSB) encoding: The adversary may encode the training dataset in the model parameters' least significant (lower) bits.

(2) Found to correlate value encoding: The adversary "may gradually encode data while training model parameters." As an illustration, "the attacker can introduce a malicious term into the loss function that maximizes the association between the variables and the data he intends to encode."

(3) Sign encoding: Like correlation value encoding, the attacker can read model parameters as bit strings by using their sign. For example, positive parameters represent 1 and negative variables represent 0.

The opponent is presumptively refused entry to the design variables in the black-box scenario. The adversary can "augment the training dataset with inputs whose labels encode the important information," according to the plan they developed. The information is then disclosed through the outputs of these additional inputs.

Attacks on model memorization examine how fraudulent training algorithms purposefully produce models that reveal details regarding their training data. Since this model system is more lenient toward the enemy than any other attack, it can collect more data about the training data.

D. Membership Inference Attack

The term "membership inference attack" refers to the process of learning whether a certain data record (x*, y) is a part of the training dataset D for the model. Figure provides an example of such an assault.

A "black-box membership inference" was developed by Shokri et al. using a shadow training method to mimic the target model's behaviors. The target model's forecasts on training and non-training input are compared using the trained inference model to "recognize differences." They also discovered that the primary elements that make a model susceptible to a membership inference attack include overfitting, the model's structure, and its type. Investigating "the link between overfitting and privacy leakage" were Long et al. and Yeom et al.[2]

A membership inference attack technique utilizing unsupervised binary classification was developed by Salem et al. and does not require the training of any shadow models or the assumption of model or data distribution knowledge. To attack GANs and Variational Autoencoders (VAEs), for instance, Liu et al. trained an attacker network to use membership assaults. The goal of Hayes et al. was "generative models in ML-as-a-service applications to train GANs to recognize training inputs."

Melis et al. researched membership inference in group learning where the attack is carried out by examining periodic modifications to the shared database as it is being trained. This approach has worked because neural network gradients are dependent on attributes that may be inferred from participant gradient updates, which in turn are based on the participants' own training data. A membership inference attack by a malicious server against the privacy protection of the distributed learning framework was taken into consideration by Wang et al. The suggested attack framework exploits GAN's multi-task classifier that simultaneously separates input samples' categories, reality, and client identities to recover user-specific private information. [3]

PROTECTION SCHEMES

Private Machine Learning (PML) schemes are necessary to protect sensitive data, abide by data protection laws, promote cooperation, enhance security, and preserve company privacy. These schemes enable businesses to use sensitive data for machine learning while maintaining company confidentiality, protecting privacy, adhering to data protection laws, facilitating collaboration, and safeguarding machine learning models from threats. Additionally, they enable enterprises to work together and develop machine learning models without disclosing private information. They can aid in defending machine learning models from adversarial, model inference, and model memorization attacks. [5]

A. Encryption

The process of transforming plain, understandable information, and data into ciphertext, a coded, unintelligible form, is known as encryption. It serves to guard against unauthorized access to, interception of, modification of, or theft of private data. Machine learning (ML) applications employ encryption as a mechanism to safeguard the confidentiality and integrity of the data used in those applications. The training data, the model parameters, and the inference outcomes can all be protected via encryption as part of the machine learning process

Machine learning models can be encrypted using a variety of encryption techniques, such as symmetric-key encryption, public-key encryption, and homomorphic encryption. [6]

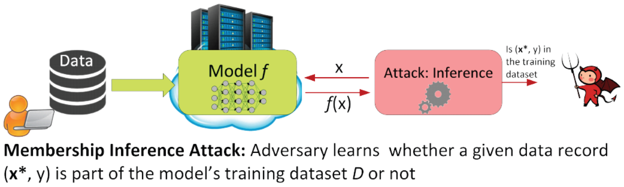

Data
Model f
x
f(x)
Attack: Inference
Is (x*, y) in the training dataset
Membership Inference Attack: Adversary learns whether a given data record (x*, y) is part of the model's training dataset D or not

Figure 12: Membership Inference Attack [1]

- In symmetric-key encryption, the model parameters are encrypted and decrypted using the same key. To prevent unwanted access, the key must be kept a secret. Symmetric-key encryption has the drawback of requiring secure key distribution to all authorized parties.
- Public-key encryption, also known as asymmetric encryption, employs both a public key and a private key. The public key is used to encrypt the model parameters, while the private key is used to decode them. The public key can be freely distributed, but the private key must be kept a secret. One advantage of public-key encryption is that it does not need a safe key transmission.
- Homomorphic encryption is a more advanced encryption technique that permits calculations on encrypted material without first needing to decode it. This suggests that utilizing the machine learning model on the encrypted data does not need first decoding it.

Several elements of the machine learning process, including the training data, the model parameters, and the inference outcomes, can be protected via encryption [3]

1. Training data encryption: The security of training data used in machine learning (ML) applications is protected using a technique called training data encryption. Training data is frequently sensitive and may include financial information, personally identifiable information (PII), or other private data. Sensitive data can be protected from leakage during training by being encrypted to help prevent unwanted access.
2. Machine learning model encryption: Machine learning model encryption is a method used in machine learning applications to secure the secrecy of model

parameters. Machine Learning algorithms are frequently private or include confidential material, and encryption could be used to protect attackers from obtaining this information

3. Inference Encryption: Inference encryption is a method of protecting data privacy during the inference step of a machine-learning model. Inference is the process of making predictions using a trained machine-learning model and new data. Inference encryption entails encrypting input data before it is employed in the machine learning model, as well as an encrypting model output before it is provided to the user.

Homomorphic encryption is the most often used approach for encrypting training data, and it is used for training data for relatively basic classifiers such as hyperplane decisions, Naive Bayes, and decision trees. The work is then extended to deep neural networks by the researchers (DNN). Dowlin et al. introduced CryptoNets, which show how to effectively convert trained neural networks to function with encrypted input data. Hesamifard and colleagues provided a methodology for training the neural network using encrypted input. Li et al. addressed the problem of collaborative learning with diverse keys and offered a solution based on multi-homomorphic encryption.

Phong et al. proposed applying homomorphic encryption on the gradient to protect the model's privacy. Training neural networks, particularly DNNs, over encrypted input remains difficult. Secure multi-party computation (SMC) is a multiparty extension of encryption that employs a mix of encryption and oblivious transfer to secretly conclude the computation without viewing the individual components. SMC has been applied to a wide range of standard machine learning models, including decision trees, k-means clustering, logistic regression, linear regression, and Naive Bayes classifiers. However, SMC approaches incur significant computing overheads, making its use of privacy-preserving neural networks, particularly deep learning, a difficult undertaking.[2]

SecureML a recent SMC example employs two-party computations to secretly train neural networks and logistic regression models. SMC can address both data/model privacy problems at the expense of communication overhead.

B. Obfuscation/Perturbation (Differential Private Learning)

Obfuscation or perturbation is a private machine learning approach used to secure the privacy of training data, model parameters, and inference outcomes. Obfuscation is the process of adding noise to data or model parameters so that the final model stays accurate but cannot be reverse engineered to recover sensitive information.[6]

It is widely used because the DP technique is typically applied in real applications through obfuscation. Obfuscation can be implemented to either the model or the data. When the obfuscation strategy is for the model, it is known as differentially private machine learning in the community. Some early work on regular machine learning with differential privacy exists. Rubinstein et al. explained differentially private support vector machine (SVM) learning processes that add noise to the output classifier and provide near approximations to non-private SVM. Chaudhuri et al. created a deferentially private empirical risk minimization (ERM) classifier using model objective perturbation. [6]

Song et al. developed privacy-preserving SGD for generic convex objectives and confirmed the approach's performance with logistic regression classification. They trained the model by updating the specified local gradients and introducing noise inside the confidentiality limit of each parameter. Based on these findings, Abadi et al. developed a simplified differential private SGD (DPSGD) method that achieves DP by reducing gradients to a maximum l2 norm for each layer. The noise constrained by the l2 norm-clipping-bound is then added. The DPSGD technique was used to train high-quality models while maintaining a moderate privacy expenditure.

The DP noise is applied to the gradients in DPSGD, and the entire training procedure takes several iterations. As a result, it is critical to computing the total confidentiality loss of training, often known as privacy accounting. Although the composition theory may be used to calculate overall privacy loss, it is extremely ambiguous. Abadi et al. proposed a moments accountant approach for tracking privacy loss across numerous training rounds and generating a tighter constraint. Another concept that is closely related is Rényi differential privacy, which provides a quantitatively precise means of measuring cumulative privacy loss" through several rounds of DP mechanisms.[1]

Users have various data sources in several real-world work contexts. They may be pertinent and should be safeguarded. As a result, in some circumstances, the DPSGD technique results in a higher amount of privacy loss. McMahan et al. developed the DP-FedAvg algorithm, a client differential private technique, to safeguard all user's data. The DP-FedAvg approach restricts the input of the entire data set to the learning model rather than restricting the input of a single record. The DPSGD technique was merged with the Distributed Averaging approach, which employs a model averaging server.[6]

Obfuscation of training data has still not been actively analyzed in the context of ML since it is regarded as typical big data privacy. Zhang et al. introduced an obscure function and applied this to the training data prior to giving it to the model training phase. This function adds random noise to current samples or adds new samples to the dataset. This conceals sensitive information about individual sample attributes or statistical qualities of a set of samples. However, the model trained on the obfuscated dataset is still capable of achieving high accuracy. [1]

C. Aggregation.

Aggregation is a technique used in private machine learning schemes to protect the privacy of the training data by aggregating information across multiple sources. The goal of aggregation is to combine information from multiple sources in a way that preserves the overall accuracy of the model while preventing an adversary from inferring the presence or absence of a specific individual in the dataset.

Aggregation can be used both during and after training. When utilized during the training phase, it frequently collaborates with the encryption scheme. Pathak et al., for example, suggested an aggregation technique for separately trained classifiers. They use DP and SMC to average the parameters. However, they do not explicitly assess the correctness of their technique.[4]

The federated learning method presents strategies for collecting gradients fast and reliably. This technique focuses on improving the aggregation process's communication efficiency and making the protocol resistant to attackers. It does not, however, guarantee the quantity of user information leaks during training. However, following the training process, employing aggregation strategies for privacy protection in ML, namely, using ensembles of models, is equally feasible. If an ensemble has enough models, and each model is trained using distinct sets of the training data in a dispersed way, then most of the models' predictions should not be reliant on any single component of the training data. This concept underpins the private aggregation of teacher ensembles (PATE) . For a new student model, the ensemble is viewed in more depth as a group of teachers. The only thing that connects the student to the professors is their ability to forecast. In addition, the learner is instructed through questioning professors about unlabeled instances. This procedure separates the prediction output from the training data. As a result, data privacy can be preserved. PATE has a substantially smaller privacy budget than typical DP ML techniques. However, because it is based on an unlabeled public dataset, it may not be applicable in many actual applications. [4]

Dwork et al. presented a method for aggregating prediction output instead of model output. More specifically, they divide the dataset D into numerous subsamples D1,..., Dr and perform a nonprivate learning process on each of those subsamples to get predictors f1,..., fr, then apply a differentially private aggregation strategy on values f1 (x),..., fr (x) and output the result. This subsample-and-aggregate strategy is simple to deploy since it does not need the development of a new learning algorithm. It focuses on data privacy in training through private prediction.[3]

IV. Machine Learning Aided Privacy

An emerging field called machine learning assisted privacy protection seeks to safeguard people's privacy when processing and analyzing data using machine learning algorithms. Protecting privacy has become a top priority due to the spread of data-driven technologies and the rise in the amount of personal data being gathered and analyzed. Machine learning techniques are frequently employed for extracting insights and patterns from huge datasets. However, if private information is drawn out of the data or inferred from it, this method might compromise people's privacy. By creating algorithms and methodologies that can safeguard people's privacy while yet enabling the extraction of practical insights from the data, machine learning-aided privacy protection aims to address this.

This section analyses various uses of machine learning (ML) to support privacy protection. In the initial part, traditional data privacy issues and threats are discussed and how recently developing ML methods can be used to combat them. The figure 13 depicts the attacks and protection schemes of private machine learning.

Utilizing social networks, mobile applications, and web-based and web-based applications, people are spending an increasing amount of time online. These are all threats to privacy. Online photo sharing has grown tremendously in popularity. On numerous social networking platforms, like Facebook, Google+, and Flickr, users are posting their photos more often. Images that are shared online might expose private information about individuals and their surroundings. Imagine someone posting a picture of a family reunion. The persons in this image may or may not have chosen to be there, but it might also disclose private details about the family's customs, religious views, and eating habits. Consequently, posting images online might grossly breach one's right to privacy and reveal crucial information. Major traditional privacy attacks include Linkage attack, model inversion Attack and Adversarial Attack.

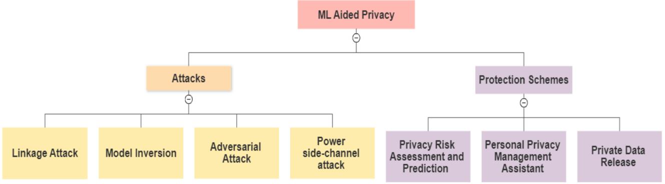


Figure 13: Attacks and Protection Schemes of ML Aided Privacy

A. Linkage Attack

A privacy attack known as a "linkage attack" aggregates data from many sources to link and identify a specific person, even though the data itself does not include such information. An attacker may, for instance, combine an individual's zip code, gender, and date of birth from several data sources to identify them. This poses a severe privacy risk since it can divulge private information about a person, even if they did not choose to disclose it.

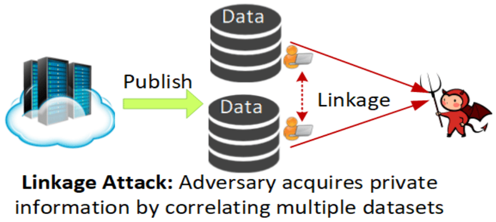

Data
Publish
Data
Linkage
Linkage Attack: Adversary acquires private information by correlating multiple datasets

Figure 14: Linkage Attack [1]

The typical steps in a linkage attack are as follows:

1. Information gathering: The assailant collects information from a variety of sources, including public records, social media, and commercial data brokers. Name, birth date, address, phone number, and email address are a few examples of the data that might be contained in this file.
2. Data preparation: The attacker defines the data's format and extracts pertinent information from the data by preprocessing it. To do this, data may need to be transformed into a standard format, cleaned up of mistakes or duplication, and made suitable for machine learning techniques.
3. Data matching: To link individuals across several datasets, the attacker uses matching algorithms on the preprocessed data. Numerous methods, including probabilistic matching, rule-based matching, and

machine learning-based matching, can be used by matching algorithms.

4. Link validation: The attacker verifies the link by contrasting the associated data with sources outside of the network, such as public records or social media profiles. This process assists in making sure the identified people are accurate and the linkage is accurate.
5. Attacker conducts a privacy assault, such as identity theft, profiling, or discrimination, using the associated data. Individuals may suffer severe injury because of this action, including loss of privacy, monetary loss, or reputational damage.
6. The difficulty of spotting linking crimes is one of the main obstacles to prevention. Since linkage attacks sometimes involve merging data from many sources, it might be challenging to pinpoint the specific source of the attack. Furthermore, the assault can be undetectable to the person whose privacy is being violated since it is subtle.

It is crucial to take precautions to guard against the linking of personal data from various sources to prevent linkage attacks. This could involve using data anonymization techniques, restricting access to personal information, and encrypting data. People should also be cautious about the information they disclose online and the people they share it with. Additionally, they should routinely check their online profiles and accounts for any suspicious activities.

B. Model Inversion Attack

A privacy attack known as a "model inversion attack" is the use of a machine learning model that was built on a person's data to infer sensitive information about that person. Machine learning models frequently divulge details about the data they were trained on, even if that data is not explicitly visible, making this kind of attack viable. By examining a person's online activity or browser history, an attacker may, for instance, use a machine learning model to deduce that person's political beliefs or current state of health.

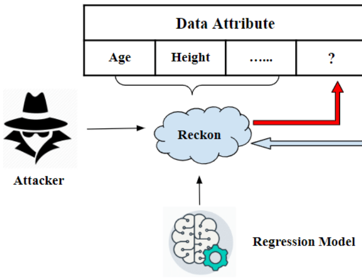


Figure 15: Model Inversion Attack [3]

A model inversion attack involves the following steps:

1. Model exploitation: The attacker utilizes the machine learning model to anticipate details about a person's personal preferences, financial situation, or other sensitive information.
2. Inference: The attacker utilizes the machine learning model's predictions to draw conclusions about the target's personal characteristics, such as their political beliefs, income level, and state of health.
3. Privacy Attack: Blackmail, discrimination, or identity theft are examples of privacy attacks when the attacker exploits the inferred sensitive information. People who do this action risk suffering major consequences including loss of privacy, monetary loss, or reputational damage.

The fact that the attacker does not require direct access to the victim's data makes stopping model inversion assaults difficult. Instead, they might use the data provided by the machine learning model to extrapolate private information about the person. This implies that even if the data is securely stored, the machine learning model may still be able to access private data. Strong privacy-enhancing approaches, such as differential privacy, federated learning, and secure multiparty computing, must be used to stop model inversion attacks. By restricting the amount of data that machine learning models may divulge, these strategies can aid in protecting personal data. Additionally, by including privacy-enhancing methods throughout the training process, machine learning models may be created to be more privacy-preserving.

C. Adversarial Attack

In an adversarial attack, the attacker intentionally modifies the input data to deceive the model's output, posing a security risk to machine learning systems. To dramatically impact the output of the machine learning model, the attacker aims to add minor modifications to the input data that are often invisible to humans. An attacker may, for instance, alter an image such that a machine learning model incorrectly classifies it. Adversarial assaults can have major repercussions because they can be used to force machine learning algorithms to make bad judgements, which can have far-reaching effects.

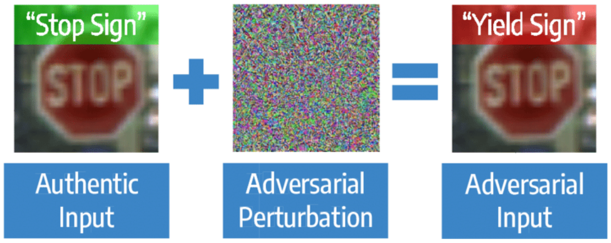


Figure 16: Adversarial Attack of Image classifier [4]

An adversarial attack involves the following steps:

1. Data collection: The attacker gathers samples of data that the machine learning model may use to train itself or to generate predictions.
2. Generation of adversarial samples: The attacker creates adversarial instances by introducing little alterations to the original data. In most cases, this is accomplished by increasing the loss function of the machine learning model while minimizing the changes to the input data.
3. Model evasion: The attacker exploits the adversarial instances to avoid detection or categorization by the machine learning model.
4. Privacy Attack: Attacker conducts a privacy assault, such as identity theft, fraud, or impersonation, using the model's weakness. Individuals may suffer

severe injury because of this action, including loss of privacy, monetary loss, or reputational damage.

The fact that adversarial attacks might be challenging to identify is one of the difficulties in protecting against them. Adversarial assaults are intended to be covert, and frequently the changes made to the input data are not visible to people. It can also be difficult to create machine learning models that are accurate without being vulnerable to adversarial assaults. It is crucial to have strong security mechanisms in place, such as data sanitization, model validation, and model hardening, to fight against adversarial assaults. Data sanitization entails removing any possible security holes that an attacker would try to exploit. Testing the machine learning model to see if it is resistant to hostile assaults is known as model validation.

PROTECTION SCHEMES

There are several privacy protections measures in operation. The main technologies include obfuscation, anonymization, and cryptography techniques, which limit information exchange. The traditional privacy protection strategies, on the other hand, concentrate on structured data, such a database entry. Both the amount and the complexity of the data are growing because of the development of new applications like the Internet of Things (IoT) and vehicular networks. Traditional security measures can't cover all eventualities, and it gets harder for average users and even data curators to comprehend the danger, choose the right measures, and control their privacy.

ML has been developed during the past several years to improve privacy protection. Several techniques like Privacy risk assessment and prediction, Personal privacy management assistant and Private data release are developed to address these issues.

A. *Privacy Risk Assessment and Prediction.*

The privacy risk exists either when the user is just accessing the application (passively collected information by malicious attackers) or sharing on social networks. In both cases, ML can help to prevent the loss of sensitive information. Fig 17 shows the diagram for this purpose.

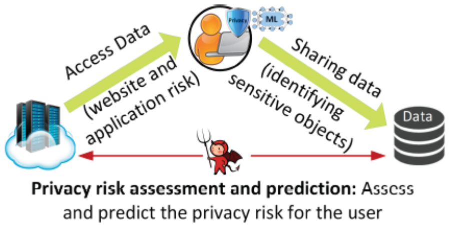

Access Data
(website and application risk)
Privacy
ML
Sharing data
(identifying sensitive objects)
Data
Privacy risk assessment and prediction: Assess and predict the privacy risk for the user

Figure 17: Privacy risk assessment [1]

1. Website and application privacy risk prediction: Website viewing may be made safer because to ML. The suggested browser extension gathers data about websites that users visit and uses machine learning to provide users feedback about the websites' privacy practices. To detecting and identifying websites that may be fraudulent or otherwise endanger users' privacy, a Bayesian classifier-based strategy is presented. To determine if internet reviews for websites are trustworthy or not, the suggested method analyses the reviews. This rate the privacy hazards of apps using an SVM classifier. The findings show that over 90% accuracy may be achieved in identifying privacy threats.
2. Identifying sensitive information when sharing: Sensitive material in multi-media data has historically proven challenging to identify. Users may avoid losing their personal information while posting photographs on social media by using cutting-edge ML algorithms. The ML models can categorize the images and assess the degree of sensitivity to enable decision-making based on previous user actions. A program dubbed "iPrivacy (image privacy)" is suggested to lessen the load on users to define privacy settings when they share photographs online. iPrivacy automates the procedure using ML. It extracts things that are privacy-sensitive from photos and categorizes them accordingly. According to the classification, iPrivacy alerts users to any items that, before sharing, should be repressed or conceal owing to privacy concerns.

B. *Personal Privacy Management Assistant*

The responsibility of managing privacy shifts more and more on individuals as user connectivity rises and web apps grow pervasive. Unfortunately, it is unlikely that users can effectively maintain and fine-tune their privacy choices given the complexity of the surroundings and the lack of understanding about privacy assaults by adversaries. Therefore, the creation of automated privacy management systems is urgently needed to assist users in safeguarding their privacy. Fig. (b) provides an example of such a defensive system.

Users regularly alter both their privacy choices and criteria to get the amount of privacy they anticipate. Furthermore, to give users a personalized experience, mobile and online applications are working to adapt their services to specific tastes. A customized service like this might put consumers at danger. This data demonstrates the necessity of creating assistants to aid users in managing their privacy preferences. ML can be a tremendous help in this situation. For instance, it can assist users in managing their privacy settings and lessen the time and staffing demands placed on maintaining privacy. ML for privacy management is divided into two broad categories: (i) privacy policy evaluation, and (ii) user preference prediction and management.

1. Privacy policy evaluation: Before using practically any program or web application, users are often asked to confirm their agreement to the provider's privacy rules. Complete details on the gathering, storing, and sharing of personal data are provided by privacy policies. They are therefore essential to consumers' privacy. Unfortunately, a lot of this material is written in difficult-to-read technical jargon. Therefore, most readers opt to accept the policy without question without fully understanding its implications. Based on user preferences, a mechanism is built to assess how comprehensive privacy rules are. The system analyses and confirms the existence of the privacy measures

that users designate while also evaluating their level of completeness using natural language processing.

The use of machine learning (ML) is suggested to "summarize the long privacy policies" into a concise paragraph that is legible and intelligible for consumers. improved privacy notifications for users in IoT networks have been given some thought. The authors employ machine learning (ML) to extract notice and choice statements from the privacy policies for IoT devices to aid consumers in comprehending the consequences of privacy notices

2. User Privacy Preference Prediction and Management: Each user has a unique level of privacy sensitivity and choice, which makes protecting user privacy challenging. Applications today frequently offer a wide range of features with varying degrees of privacy protections. Users are frequently asked for rights to access resources that have an influence on their privacy while installing the programs. The consumers' ability to effectively match their personal privacy preferences with the real privacy risk is crucial.

   To forecast user privacy preferences and aid in decision-making, ML approaches are used. It was first demonstrated to be practicable since early research revealed a connection between user privacy preferences and certain statistical and environmental factors. To forecast user privacy preferences and make privacy management decisions, machine learning models may be created. Openness, conscientiousness, neuroticism, extroversion, and agreeableness are characteristics that are fed into machine learning (ML) models to anticipate desirable consumers' preferences.

3. Private Data Release: Currently, database release is a crucial step in data analytical applications. Different entities produce various forms of data, such as health data from hospitals. Then, these data will be sent to data custodians, such governmental organizations. The data custodian then keeps up a platform that allows data consumers, such as other government agencies, people, analysts, etc., to access, organize, and store data. When the data custodians disclose the data, privacy preservation processing is very necessary. Fig. 7(c) provides an example of such a defensive system. Obfuscation with the addition of noise to the source dataset is a regularly employed traditional private data release strategy. Whereas ML approaches offer a novel approach to this issue, employing generative adversarial networks (GAN) or generative neural networks (GNN) to create artificial datasets.

   In general, the most recent deep learning approaches demonstrate the capacity to create a synthetic dataset that is statistically equivalent to the real one. Private data release can be accomplished using this method. The generative model framework for rich semantic data privacy preservation is shown in Fig. 8. The best way to describe the procedure is through an example. Consider a scenario for clinical data sharing in which the data curator trains a deep generative model with the original data in a differentially private manner rather than immediately releasing it, and then publishes a synthetic dataset produced by the model. A deeper generative model from which "an unlimited amount of synthetic data for arbitrary analysis tasks" can be generated may be published by the data curator in a more generic scenario. The use of generative models can considerably increase user privacy because the training of the models can be done using artificial data rather than actual user data. The statistical closeness of models trained using synthetic data and realist data has been frequently demonstrated in the literature, therefore the value of the dataset may be ensured. As an illustration, study by Park et al. demonstrated the statistical resemblance between generated synthetic tabular data and actual data. Researchers Xu and colleagues created training deep neural networks to create synthetic data that closely resembles patient medical records

Though the study of GNNs for privacy protection is still in its early phases, the outlook for the strategy is positive. Synthetic data generation is especially important since semantically rich data cannot be protected from privacy using conventional approaches like anonymity and obfuscation. Additionally, this method does not have the disadvantages of other conventional anonymization methods, such as requiring previous knowledge or connecting the data to other sources.

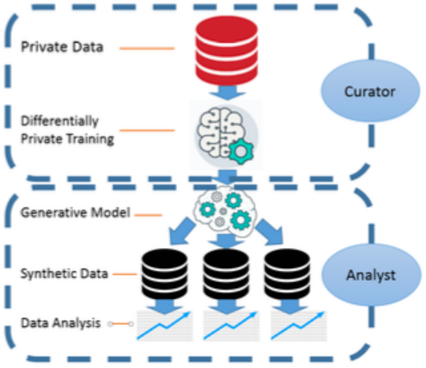


Figure 18: Privacy preserving framework based on generative model approach [7]

V. FUTURE WORK

Some of the future work on the privacy concerns and solutions on the Machine Learning models include:

- The issue with private machine learning has received the most attention lately. Many research projects in this area attempt to employ the differential privacy criteria during the analysis. However, due to the complexity of the data and privacy protection aim, DP notation cannot give an exhaustive privacy evaluation. Therefore, it is currently unclear how to create new privacy measures and notations.
- Research on machine learning-assisted privacy protection is picking up speed. For instance, employing GNN to create artificial datasets offers up a new area of study for privacy protection,

particularly for unstructured data like picture and video.

- The study of defense mechanisms against ML-based privacy attacks is still in its early stages. But given the widespread use of AI techniques in all facets of future networks, it is anticipated to fly in the future. Currently, the adversarial example/perturbation approach is the most widely used technology in this area.

VI. Individual Section

My contribution to the project is to analyze the privacy and security concerns in the field of Machine Learning Aided Privacy. From the available three fields 'Machine Learning Based Privacy, Private Machine Learning and Machine Learning Aided Privacy', I have chosen Machine Learning Aided Privacy protection since it aims to safeguard people's privacy when processing and analyzing data using machine learning algorithms. To progress, I've referred various research papers, publications, and blogs in the process of developing the report. I have identified the key privacy challenges for Machine Learning Aided Privacy and the possible security solutions that can be developed using ML. I identified the concepts like Linkage Attack, Model inversion Attack, Adversarial Attack, Power side-channel attack and the possible solutions like Privacy Risk Assessment and Prediction, Personal Privacy Management Assistant, Private Data Release to provide solutions for the mentioned privacy attacks in the Machine Learning. I also discussed the future research challenges in implementing the solutions. The following are the research papers that I have referred for my topic:

[1] When Machine Learning Meets Privacy: A Survey and Outlook.
[10] Membership inference attacks against machine learning models
[11] Model-based information leakage detection and classification
[12] Privacy-preserving Machine learning: A survey
[13] Model inversion attacks that exploit confidence information and basic countermeasures
[14] Machine Learning Based Network Attacks Classification

We are a group of three people worked on the project. Right from the beginning me and my teammates were interested in the topic of the privacy concerns and solutions in Machine Learning. Through the class we were able to identify the challenges the Machine Learning algorithms were facing and by integrating them we wanted to develop the privacy attacks and security solutions for the Machine Learning. We focused on three main fields 'Machine Learning Based Privacy, Private Machine Learning and Machine Learning Aided Privacy' of the Machine Learning. we shared the work equally by each teammate working possible attacks and solutions on each field. We have frequent team meetings and support each other from the gathering of the contents to the development final report and presentation. As the work is equally shared, I hope all my teammates will get the same grade

VII. Conclusion

In this survey, we provided an overview of existing privacy concerns and solutions in ML algorithms, to improve the privacy protection of the users. We explored various attacks in the Machine learning techniques in the fields of Machine Learning Based Privacy, Private Machine Learning and Machine Learning Aided Privacy and provided some privacy protection schemes which can be used to protect the users from the mentioned attacks. Finally, future research directions were also pointed which can improve the privacy in machine Learning.

References

[1] Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. 2021. When Machine Learning Meets Privacy: A Survey and Outlook. ACM Comput. Surv. 54, 2, Article 31 (March 2022), 36 pages. https://doi.org/10.1145/3436755

[2] M. Al-Rubaie and J. M. Chang, "Privacy-Preserving Machine Learning: Threats and Solutions," in IEEE Security & Privacy, vol. 17, no. 2, pp. 49-58, March-April 2019, doi: 10.1109/MSEC.2018.2888775.

[3] H. Xie, L. Wei and F. Fang, "Research on Privacy Protection Based on Machine Learning," 2021 International Wireless Communications and Mobile Computing (IWCMC), Harbin City, China, 2021, pp. 1003-1006, doi: 10.1109/IWCMC51323.2021.9498632.

[4] M. Xue, C. Yuan, H. Wu, Y. Zhang and W. Liu, "Machine Learning Security: Threats, Countermeasures, and Evaluations," in IEEE Access, vol. 8, pp. 74720-74742, 2020, doi: 10.1109/ACCESS.2020.2987435.

[5] A. Pandey, A. S. Genale, V. Janga, B. B. Sundaram, D. Awoke and P. Karthika, "Analysis of Efficient Network Security using Machine Learning in Convolutional Neural Network Methods," 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2022, pp. 170-173, doi: 10.1109/ICAAIC53929.2022.9793293.

[6] Qiang liu and Muhammad Arif. 2022. Privacy Protection Technology Based on Machine Learning and Intelligent Data Recognition. Sec. and Commun. Netw. 2022 (2022). https://doi.org/10.1155/2022/1598826

[7] Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In Security and Privacy, 2008. SP 2008. IEEE Symposium on (pp. 111-125). IEEE.

[8] Hitaj, B., Ateniese, G., & Perez-Cruz, F. (2017). Deep models under the GAN: information leakage from collaborative deep learning. arXiv preprint arXiv:1702.07464.

[9] Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (pp. 1322-1333).

[10] Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In 2017 IEEE Symposium on Security and Privacy (SP) (pp. 3-18). IEEE.

[11] Salem, A., Zhang, Y., Humbert, M., & Jin, X. (2019). MLeaker: Model-based information leakage detection and classification for deep learning. IEEE Transactions on Dependable and Secure Computing.

[12] Zhang, C., Liu, C., Liu, J., & Chen, Y. (2020). Privacy-preserving deep learning: A survey. IEEE Access, 8, 36700-36719.

[13] Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In Proceedings of the IEEE international conference on computer vision (pp. 3730-3738).

[14] Zhao, S., Gallo, O., Frosio, I., & Kautz, J. (2017). Loss functions for neural networks for image processing. IEEE Transactions on Image Processing, 26(9), 4285-4298.