Privacy of the Metaverse: Current Issues, AI Attacks, and Possible Solutions

Chamara Sandeepa*, Shen Wang†, Madhusanka Liyanage‡
*†‡School of Computer Science, University College Dublin, Ireland
Email: *abeysinghe.sandeepa@ucdconnect.ie, †shen.wang@ucd.ie, ‡madhusanka@ucd.ie

***Abstract*—Metaverse is a key emerging digital transformation concept for the next generation of cyberspace. It is expected to create a self-sustaining virtual ecosystem of fully immersive, real-time experiences with numerous opportunities for general users and industries to interact with the world. With the introduction of 6G networks and enabling technologies, the metaverse will achieve its success on a large scale. However, with increasing interaction using new technologies and a lot of third-party services, there would be an arena for more possibilities of privacy threats. Hence, privacy requirements are critical for the metaverse. They should be cautiously investigated since we see the commercial adoption of the metaverse is imminent. Therefore, this paper discusses different privacy issues, potential Artificial Intelligence (AI) related privacy attacks and possible solutions to the metaverse. We initiate this by introducing the concepts of the metaverse and privacy. We then discuss potential privacy issues that could occur with the future metaverse. We present a new attack approach utilizing combined membership inference and reconstruction attacks that can be launched against metaverse users. We also propose viable techniques and tools that could act as possible solutions to those issues.**

***Index Terms*—Metaverse, Privacy Attacks, AI, 6G, Virtual Identity, PII**

I. Introduction

Since the beginning of the Internet in the 1980s, cyberspace's emergence has revolutionized people's day-to-day interactions with the world. Also, during a similar period, the rise in wireless networks has completely changed the notion of communication despite the physical barriers. It is visible that the Internet, wireless networks, and associated infrastructure have evolved tremendously within the previous decades, driving people to a facilitated, virtually connected environment. The metaverse is very likely to become mainstream as the industry's attention to the metaverse has increased significantly in recent years. One major example is the re-branding of the Facebook company to Meta [1], which shows the company invests heavily in making the metaverse a reality. According to [2], almost 50% of Europeans have switched to at least partial Work from Home (WFH) option compared to 12% before the outbreak, showing the past COVID-19 pandemic also had a significant impact on migrating from physical to virtual workplaces.

The advent of 6G communication will speed up the metaverse's ongoing efforts. Due to its capabilities, including ultra-high peak data rates in the terabit range, very low latency communication with less than one millisecond, enhanced mobility from information exchange via all mediums, and extremely high reliability beyond 99.99999%, the 6G will serve as a key enabler for the metaverse [3], [4]. All these features are significantly higher than the current 5G networks, and they are crucial for the metaverse because it requires rapid data transfers with minimum interruption for communication. Faster rates will be especially important for multi-sensory remote devices and high-quality 3D environment rendering. In complex virtual environments, the capacity to manage real-time connections between millions or even billions of people is crucial for fostering a seamless social experience. With the development of 6G networks, a variety of new technologies will shortly be available, enabling the metaverse. However, with these features, technologies, and great community interest with possible future engagement, metaverses will undergo numerous privacy threats from internal providers and external parties. Therefore, we bring this discussion of privacy challenges through this paper. Our paper discusses existing privacy issues, including possible attacks on metaverse AI-based services, and a combination of potential existing solutions that can be used to mitigate the identified privacy issues.

Our Contributions: We summarise our contributions from this paper in the following key points:

- We provide an overview and discuss the importance of privacy in the metaverse.
- We identify a set of key privacy issues in the metaverse over multiple communication layers, from the sensing layer to metaverse services.
- The possible solutions that we can apply for the metaverse privacy issues to mitigate and their relative impact are discussed.
- We present a novel privacy attack on AI by combining membership inference and reconstruction against metaverse wearable IoT devices to recover user emotional status and real identity.

II. Background

A. Metaverse

The term "metaverse" refers to a vast, computer-generated virtual space that exists alongside the real world and was first used in 1992 from a book named "Snow Crash" [5]. There are individuals in this setting known as "Avatars," and they have characteristics that are equivalent to or even beyond those perceived in the physical world. As a result of the Internet, web technologies, and Extended Reality (XR), the metaverse is now seen as a blend of the physical and digital worlds [6]. Low latency, reliable communication through 6G networks, and technologies such as AI-based decision-making and edge AI are key enablers of the metaverse. The XR application quality will also be highly dependent on the capacity of the 6G wireless network to provide a fully immersive experience.

There are three steps in the development of the metaverse [6]: 1) *digital twins* - To make the virtual infrastructure/world match with the real world, 2) *digital natives* - create virtual content by people, through methods such as avatars, and 3) *coexistence of physical-virtual reality* - to build a sustainable coexistence between the actual world and the virtual world. It will also have a sustainable coexistence that can work independently. For this, the metaverse would include its own ecosystem having a virtual economy with internal economic governance, metaverse commerce, a trading system, and ownership [6]. The new technologies such as Brain-Computer Interface (BCI) or haptics will enable sharing of details up to emotional and sensation levels and facilitate further content creation [7]. However, with innovations and more ways of exposing personal details, inevitably, privacy concerns are critically arising with the metaverse.

B. Privacy

In general, the concept of privacy assures data owners the ability to control or influence their information on the collection, storage, and by whom and to whom the information may be disclosed [8]. There are many proposed taxonomies of privacy based on different perspectives. One such example is the consideration of different actions done on the information of a data subject [9]: information collection, dissemination, processing, and invasions. The General Data Protection Regulation (GDPR) Art. 4 [10] defines two categories of data to be considered: personal and non-personal data. Personal data is any data that can identify a specific person (data subject). Conversely, non-personal data refers to data that has never been associated with an identified or otherwise identifiable natural person, according to [11]. Privacy aspects are crucial to address before the commercialization of the metaverse, as the adverse outcomes of privacy leakages will affect organizations and millions or even billions of individuals. If privacy is compromised in any aspect, the individuals will lose their controllability in cyberspace, leading to the loss of trust on the metaverse platform. The reflections would impact stock prices of the metaverse platform providers and huge levies on privacy leakages, such as GDPR fines [12]. The survey in [13] shows privacy threats in the metaverse, including pervasive data collection such as facial expressions, privacy leakage in data transmission, processing and storage, and compromised end devices. To show such a practical example of such privacy leakage, we implement a use-case scenario of leakage of facial expressions in Section V.

III. Key Privacy Issues in Metaverse

Prior identification of potential privacy issues would help mitigate the weaknesses beforehand and support a privacy-enhanced metaverse. Therefore, we provide several privacy concerns associated with the metaverse as shown in Figure 1. The detailed discussions are in the following subsections.

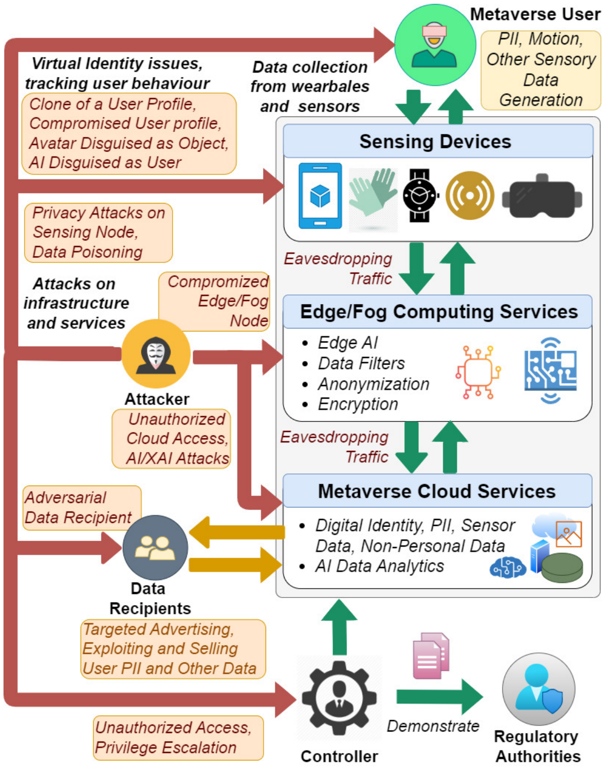

Metaverse User
PII, Motion, Other Sensory Data Generation
Virtual Identity issues, tracking user behaviour
Clone of a User Profile, Compromised User profile, Avatar Disguised as Object, AI Disguised as User
Data collection from wearbales and sensors
Sensing Devices
Privacy Attacks on Sensing Node, Data Poisoning
Eavesdropping Traffic
Attacks on infrastructure and services
Compromized Edge/Fog Node
Edge/Fog Computing Services
Edge AI
Data Filters
Anonymization
Encryption
Attacker
Unauthorized Cloud Access, AI/XAI Attacks
Eavesdropping Traffic
Metaverse Cloud Services
Adversarial Data Recipient
Digital Identity, PII, Sensor Data, Non-Personal Data
AI Data Analytics
Data Recipients
Targeted Advertising, Exploiting and Selling User PII and Other Data
Unauthorized Access, Privilege Escalation
Controller
Demonstrate
Regulatory Authorities

Fig. 1: Metaverse overview and its associated privacy issues

A. Data Collection from Wearable and Sensing Technologies

With numerous technologies in the metaverse, many types of multidimensional data will be generated during user actions. They will be collected by multiple wearable and sensing devices introduced with it. Extended reality/digital twin, haptic engagement, and holographic telepresence are some of the envisioned 6G technologies discussed in [14] that will be available in the metaverse. However, these technologies will have a high chance of exposing user privacy. For example, the work [14] discusses that XR/digital twin could capture biometric data and physical movements. Through these devices, real-world biometric information, such as gait, eye or head movements, physical characteristics, residence details, heart rate, inferred emotions, and more, could be obtained. [15] demonstrates how dwelling data, for example, might include a information of objects in the household to create an individual's psychological profile. Previously, the worst-case when a password is lost is that a person would lose some data and have to make a new one. However, if biometric data is exposed, it will be permanent [6] since they are unique to the person. Therefore, such data can be regarded as highly sensitive Personally Identifiable Information (PII). Hence, if exposed, sensing data and associated technologies in the metaverse pose a significant vulnerability for users.

B. Attacks on Metaverse Edge Infrastructure and Services

In addition to the possibility of privacy violations from the metaverse's authorized data holders and processors, there could be numerous flaws in the metaverse's hardware, software, and network.

1) Vulnerabilities of edge devices: Edge computing is a model that minimizes the overhead of cloud computing by moving computing resources closer to the "edge" or nearer to end users [16]. Metaverse will use many edge devices for its spatial, motion-sensing requirements and wearables such as VR headsets. Since edge devices have limited capacity to implement privacy preservation, issues are highly probable. These edge devices can typically be available in various real-world locations where attackers may have access. A malicious edge computing device deployed or compromised by the attacker may intercept or steal private information, including biometric, motion, and health data. The work in [17] shows issues like data manipulation and privacy leaks might occur in the edge core infrastructure.

2) AI attacks: Metaverse may use a variety of Machine Learning (ML) and AI models to determine user actions and intentions from sensor data. An AI model has created a privacy risk if they expose sensitive information about individuals. The authors present a taxonomy of several threats against ML models in [18]. An adversary can poison the input data during ML model training or testing phases, making the model less accurate or vulnerable to privacy threats. A malicious entity can also use reverse attacks to reverse-analyse the model during the testing phase. Deep Learning (DL) is also vulnerable to adversarial attacks, model inversion, extraction, and poisoning attacks, according to work in [19]. Attackers might predict personal characteristics, such as location, preference in gender, and political opinions, using public data [20]. Another significant attack is model stealing, also known as prediction poisoning, in which an adversary tries to duplicate a target ML model's functionality by taking advantage of its black box queries, like inputs and outputs [21].

Inference attack is another significant privacy concern in the metaverse, where, an adversary is attempting to infer certain information, such as the membership or properties of a target. Metaverse can consist of many associated ML models running and shared among services such as object detection, facial recognition, and motion sensing in wearable devices. If an attacker gains access to these ML models but not directly to data, they can still infer some properties on the data where the ML models are trained on. For example, in the case of membership inference, the attacker queries a trained ML model to predict if a particular example was contained in the model's training dataset [18], [22]. Such an attack is critical in the metaverse since an anonymized individual's real identity in private spaces can be revealed if ML models are trained with activities done by the users in these private spaces. Another possible attack is a reconstruction from gradients or deep leakage, where an attacker attempts to recover the original private dataset from the gradient information of publicly shared ML models [23]. This will be particularly risky for data owners with collaborative learning such as Federated Learning (FL). Because in FL, multiple parties collectively train local ML models based on their private data and later share the local models that will be aggregated to create a unified global model.

C. Tracking User Behaviours in Metaverse

The metaverse is associated with privacy concerns related to user behaviour, as it provides many interactive opportunities for users with numerous sensor tracking, as discussed. According to research describing the metaverse game Second Life, most players (72% of women and 68.8% of men) exhibit the normal behaviour similar to the reality when playing the game [24]. Their results indicate that the organizations behind the metaverse have the possibility of tracking the actual behaviour of the majority of users. Since private information like user preferences and emotions may be readily anticipated through behaviour, there could be a significant possibility of privacy breaches in the metaverse.

D. Privacy Issues of Virtual Identity

The metaverse users interact with the world through a virtual identity. Their avatars could be different based on user preferences. One possible issue with virtual identity is the difficulty in verifying if the person behind the avatar is real as the one that claims to be. We see there are several such scenarios: 1) an avatar disguised as a surrounding object to tracking users [6] 2) a compromised user profile, 3) a copy of the profile that acts as the original owner of the avatar through approaches similar to *deep-fakes*, 4) an AI claiming to be a real person, especially with the rise of large language models like Chat-GPT [25]. If a user profile is compromised, an attacker can gain access to private conversations and any historical data which is available. Further, an attacker may masquerade as a person by using the same avatar and creating a digital copy of the original user. They can be successful in pretending to be the original victim, gaining access to locations or interactions that only the victim is allowed. With the possibility of detecting emotions in the metaverse, the robots may be able to interact with users more realistically [26]. If an adversary uses this, they can extract sensitive private information of individuals. Handling the sensitive data on minors will also be challenging since they resemble a large portion of XR users [6].

IV. Privacy Solutions for Metaverse

To address the aforementioned privacy issues, we propose the following set of solutions achievable with the existing technologies.

A. Privacy Protection for User PII

The PII play an important role in metaverse privacy since exposed PII reveals the identity of individuals, leading to privacy leakage. Especially in future networks with fast and new modes of data transmission, a massive amount of the metaverse PII can be sent to a third party in milliseconds.

Many works mention techniques to ensure the privacy of PII. The authors of [27] created a library that uses homomorphic encryption for privacy-preserving image processing. The survey in [28] offers a detailed de-identification approach for non-biometric and soft-biometric identifiers in multi-media assets. Considering the avatars, the work in [6] discusses using multiple avatars and privacy copies to add noise to create confusion for attackers. Also, they discuss using temporary private portions of the metaverse to interact among users to prevent eavesdropping.

B. Privacy by Design Approaches

We consider privacy by design a crucial consideration that should be made at every stage of the design process of metaverse services. Independent authorities should assess the capacity of these services. Furthermore, even without any prior requirements from consumers, well-designed services should protect privacy needs by default. That is, taking steps to secure itself before a data breach occurs [29], [30]. AI privacy should also be a primary consideration in the design process, as we have shown many privacy attacks on AI in Section III. Some privacy by design strategies discussed in [31], [32] are: 1) reducing the amount of personal information that one collect as much as feasible, 2) concealing personal data from plain view and separating with decentralised manner, 3) processing data at the highest possible level of aggregation, 4) maintaining transparency of the data subjects, 5) enforcing a privacy policy compliant with the law, and, 6) Adhering to the current privacy policy and any applicable legal obligations.

C. Edge and Fog Computing Privacy Preservation

Edge computing increases data processing efficiency via 6G networks in the metaverse with its capability of processing data close to the user. However, it requires innovative data privacy mechanisms due to its heterogeneity and distributed nature [6]. The authors of [17] suggest lightweight data encryption techniques, fine-grained information-sharing platforms, decentralised security controls, and effective privacy-preserving practices for edge computing.

Fog computing is a layer that sits between edge devices and cloud servers, acting as a middleman for functions including data filtering and forwarding to the cloud. With fog, only a small percentage of data will be transferred to the cloud, lowering cloud server overhead and network congestion. As a result, fog computing could help to protect the privacy of IoT and users by reducing the need to transmit sensitive data to the cloud for processing [33]. However, as data from the edge will directly reach the untrusted fog layer, privacy concerns within the fog node must be considered. A compromised fog node poses the possibility of attackers listening in on or directly altering user data. To ensure privacy in Fog, the work in [34] proposes a multi-functional data aggregation methodology based on ML for fog computing with differential privacy. To establish decentralised privacy, prevent poisoning threats, and achieve high efficiency in fog computing, the work in [35] employs blockchain-enabled FL.

D. Explainable AI (XAI) Privacy Measures

XAI can provide reasonable justifications for metaverse AI-based decisions when implementing intelligence-based solutions. Also, the decisions of these actions may be influenced by how transparent and rational the AI judgment is. The authors in [36] divide the explainability space for predictions/data in the context of the security domain into three regions: 1) explanations for the predictions/data themselves, 2) explanations for covering privacy properties, and 3) explanations for covering the threat model. Depending on the nature of the data, privacy requirements, and complexity of the model as a privacy solution, we must evaluate data, privacy attributes, and model explanations. In the last few years, there has been a lot of interest in the topic of XAI. The survey in [37] displays several related XAI works from 2007 to 2020 and categorizes them by scope, technique, and application. They also show that open-source XAI products have vastly improved in recent years.

E. Blockchain-enabled solutions

Blockchain is a peer-to-peer network that uses a decentralised and distributed public ledger technology [38]. It can help many technological innovations associated with metaverse to improve their data privacy. It is also useful for user identity in the metaverse. A person's digital identity is currently shared with several organizations, including entities of government, social media sites, and other private/public organizations. The user consequently has relatively lesser control over their personal information. Digital identities that are self-managed or self-sovereign can be created using blockchain. The person would then be in charge of their own online identity. Users can access various digital services using this identification to confirm their identity [39]. Non-fungible Tokens (NFT) can be used to verify a user's identity. The work in [6] proposes blockchain can be used to enforce a democratic process to implement guidelines and penalty systems for misbehaviours. To maintain the privacy of metaverse user data without directly exposing to blockchain, several techniques can be incorporated together with blockchain, such as Multiparty Computation (MPC), Zero Knowledge Proofs (ZKP), homomorphic hiding, and ring signatures [40].

F. Regulations and Standardisation at a Global Scale

Regulations would help address privacy issues, including lack of awareness of rights and privacy concerns in public because doing so would automatically protect personal data. The regulatory approach is divided into three groups by the authors in [41]: Three types of regulation exist: 1) governmental, 2) industry-driven, and 3) consumer or market-driven. Government regulations and industry trends impact privacy issues on a broad scale. Consumers are frequently aware of their privacy rights, but [42] shows that they often lack the knowledge and tools to use these rights effectively. Therefore, enacting privacy legislation should ultimately protect consumers from invading their privacy. The work [6] discusses the proposal of standardising privacy trading through compensation for selling personal data by the original data owners. However, despite the efforts of standardisation, there can be a potential to exploit user privacy in the metaverse as they may not be immune to all possible leakages. Hence, suitable metrics by businesses and governments should be in place.

Table I below compares the various solutions we presented for resolving the issues in Section III.

TABLE I: Proposed privacy solutions for challenges in metaverse

Privacy Solution	Issues Addressed			
	IA	*IB*	*IC*	*ID*
Privacy protection for user PII	H	H	H	H
Privacy by design approaches	H	H	H	H
Edge fog computing privacy preservation	H	H	M	H
XAI privacy measures	M	M	H	M
Blockchain-enabled solutions	H	H	M	H
Regulations standardisation at a global scale	H	H	L	H

IA - Data collection from wearables and sensing technologies
IB - Attacks on metaverse edge infrastructure and services
IC - Tracking user behaviours in metaverse
ID - Privacy issues of virtual identity

L Low Impact M Medium Impact H High Impact

V. Case Study: Hybrid Membership Inference and Reconstruction Attacks

A. Introduction

In the metaverse, wearable devices such as VR/AR headsets are an essential component in establishing an immersive experience. However, a recent study in [43] has shown that many available devices have potential weaknesses in privacy, such as flaws in privacy policies, no clarity on what data is shared with third parties, less privacy customisability for users, and lack of multi-factor authentication. Therefore, there is a high possibility that privacy attacks may get launched against these devices.

To provide an example of privacy vulnerabilities in the metaverse, we designed an experiment by combining two privacy attacks, membership inference [44] and reconstruction from gradients [45]. None of the associated works in the metaverse discuss specific scenarios of privacy leakage through attacks; thus, to address this gap, we demonstrate it is highly possible to exploit privacy by combining multiple attacks. Furthermore, the combined effect of the two attacks, membership inference, and deep leakage data reconstruction attack, are not assessed in previous works. Therefore, we perform our attack where membership inference supports the attacker in identifying a target and reconstructing the data efficiently. In the metaverse, we expect decentralised ML techniques such as FL to be heavily used to train and deploy privacy-preserved AI models near the user end devices like VR headsets. With FL, metaverse users locally train their wearable device services with data like facial expressions. These may contain captures of relatively less data, which are highly user-specific. It is, therefore, possible that local models in FL get overfitted to these data.

A data reconstruction attack attempts to recover the input dataset by matching an adversarial model with the original gradients of a victim ML model [23]. An attacker launching a data reconstruction attack attempts to imitate the gradients of a target model $f(x;W)$ by lowering the loss L_g between a randomly initialized set of gradients $\Delta\tilde{W}$ with the target model gradients ΔW. This can be represented in the following equation:

$$L_g = ||\Delta\tilde{W} - \Delta W||_f^2 \tag{1}$$

With the improved attacker gradients from equation 1, data $\tilde{x}$ that is close to original inputs x can be reconstructed as:

$$\tilde{x} \leftarrow \tilde{x} - \eta\Delta_{\tilde{x}}L_g \tag{2}$$

Considering metaverse capturing devices at the edge, the data used to train FL models will be relatively low compared with big-data sets. Therefore, it is possible to recover these gradients if an attacker gains access to these models by attacking the devices like VR headsets. Due to resource limitations, weaker privacy mechanisms may be used in these devices, or low security in communication channels makes it easier for an attacker to launch the attack. However, many studies [23], [45] show it is an expensive procedure to recover the gradients with an increasing number of data that was used to train the models. Furthermore, FL may create many ML models that do not contain data from the target user. Suppose the attacker acts as an aggregator in FL from client updates with unknown origin. In that case, the attacker will not get sufficient information on the target unless they reconstruct all the received updates. However, if reconstruction attacks were launched on all ML models without knowing where to look for a target, that would cause high computation costs for the attacker.

To make the attack more directed, the adversary can combine membership inference to initially identify if the target user's data is included in a received ML model. Membership inference can be used to identify if a certain data record was present in the training dataset of a given ML model. For example, if face images of the target user are used to train the ML model, the adversary should be able to predict the presence of user information if a sample face record of the victim is available. For this, the attacker trains an ML

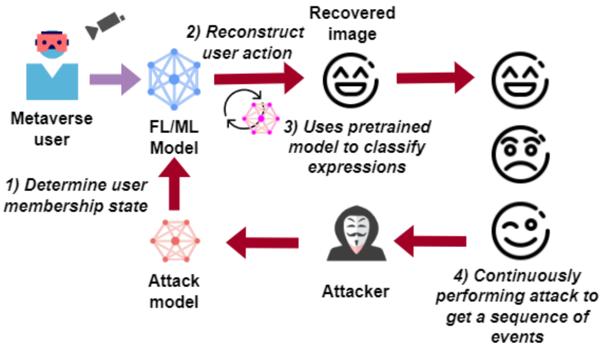

Metaverse user
FL/ML Model
2) Reconstruct user action
Recovered image
3) Uses pretrained model to classify expressions
1) Determine user membership state
Attack model
Attacker
4) Continuously performing attack to get a sequence of events

Fig. 2: An attacker uses an ML model to infer the membership state of a target user. If victim data is present, reconstruction is done to recover user images from ML model gradients. This process is repeated to track the user's emotional behaviour.

model named *attack model*. An attack model Ω is trained to predict the membership state of a given data record x, whether $x \in D_{train}$ where D_{train} is the training dataset of a target model f. This can be denoted in equation 3 as:

$$\Omega(\{f(x), y\}) = \begin{cases} 1, & \text{if } x \in D_{train} \\ 0, & \text{if } x \notin D_{train} \end{cases} \tag{3}$$

where $f(x)$ is the target model prediction and y is the label for the record x. To train the attack model Ω, the attacker generates a dataset named *shadow dataset* $D_{shadow} \leftarrow \{f_s(x'), y', s\}$. Here, f_s is called a *shadow model*, which consists of the same NN architecture of the target model f. We can create multiple copies of shadow models to increase the amount of data in the shadow dataset. $f_s(x')$ is the shadow model prediction of a representative example data record x' that resembles the original data in D_{train} of the target model. These representative data records can be generated using generative models such as Generative Adversarial Networks (GAN) [46]. In the shadow set, y' represents the labels of the example records. The value s represents the membership state of either In/Out.

B. Dataset and Procedure

For experimental simulations, we used the facial expression recognition dataset FER-2013 consisting of 28,709 training and 3,589 testing data in 7 categories of human expressions in 48x48 pixel grayscale images. We used 400 images from each class for the model training in experiments. We assume this dataset resembles a scenario when a face recognition device in a VR box captures or preprocesses the data in a low-resolution setting for improved performance. For running the experiments, we use a compute instance with a Xeon 2.20 GHz CPU, 26 GB RAM, and a GPU of NVIDIA Tesla T4. We used a Neural Network (NN) with a hidden layer of 512 dense units followed by a dropout layer with a 0.2 dropout rate as the basic model in our experiments as the target model. We named it V_0. We also used two different versions of this basic NN by adding two more extra 512 dense layers in each version named V_1 and V_2. In our experiments, we also used two other Convolutional Neural Network (CNN) architectures: LeNet-5 and AlexNet.

C. Inference Attack

As the first step in the attack, we launch membership inference with the aid of a dataset created from the outputs of shadow models as discussed in Subsection V-A. To train the shadow models, we used 1,120 input data records and generated ten copies of shadow models. We simulated the attacker's dataset by augmenting original data at different proportions to train the shadow models. Then, we train the attack model with the data collected from the shadow models. We set this attack model as a random forest classifier with 50 estimators. With the trained attack model, we evaluated the membership inference accuracy of the attack model by getting the membership state predictions for the target model's training dataset. The training dataset of the target model is arranged in small batch sizes from 1 to 52, which resembles small data availability in metaverse IoT sensor devices running FL algorithms. Thus, the training dataset may get overfitted to the local ML model. The impact of overfitting is analysed by running a varying number of epochs of 1, 10, 20, and 30 by the target model. An attack on a particular local batch size with one epoch configuration is run ten times to get an average accuracy figure for the attack round.

We further considered the availability of a victim data record. In a practical scenario, the attacker may not contain an actual image example used in the target model's training dataset. Instead, they may possess a similar image of the target user's face. To simulate this, we modify the source image using augmentation by rotating the target image to a random degree during the attack phase. We evaluate the changes in attack accuracy with 0%, 50%, and 100% augmentation percentages of the test dataset, as shown in Figure 3.

1) Impact of batch size: From the experiments, we observe that the accuracy of the attack is better for small batch sizes, and when increasing the batch size, it reduces. This means the sensor/image capturing devices that capture user data less frequently or perform frequent model updates with smaller batch sizes are at higher risk of privacy leakage. From the observations, having larger batch sizes for training models is better. However, the utility of data for larger batches and their processing time also should be considered since the captured data may be time-critical for highly accurate predictions with low latency to maintain a satisfactory user experience.

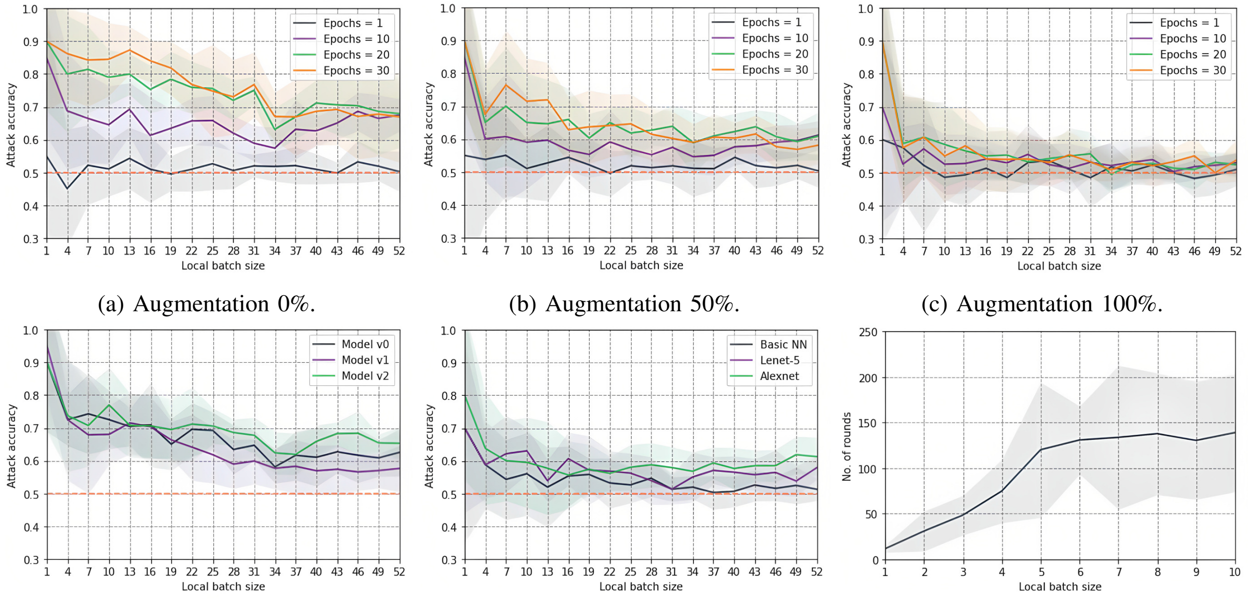
2) Impact of training rounds: The number of model training rounds also can considerably affect the attack accuracy, as observed from Figure 3. Here, the accuracy is high when the target model has a higher number of training epochs. This signifies the nature of data fitting; the more trained the data to the model, the more leakage is possible. Therefore, images captured by metaverse devices using FL/ML can use lesser training epochs. Still, it will degrade the overall model accuracies, and models may not get the total learning capacity from the captured image samples.

3) Impact of data augmentation: The disadvantage is on the side of the attacker when they do not possess exact samples of the target user's data, as observed from the augmentation percentages in the experiments. The attack success rate will drop when more augmented data is present in the attacker. However, there is still a reasonable attack capability for smaller batch sizes, even when augmentation is 100%.

Model complexity can also play a role in attack accuracy. Figure 3d and 3e show deviations in attack accuracy, where more complex models have higher average attack accuracies.

D. Reconstruction Attack

The next step in the attack is to launch the reconstruction of a target model. The reconstruction is done via improved Deep Leakage from Gradients (iDLG) attack [45] for reconstructing images with different batch sizes from 1 to 10. We selected this range since the membership inference is highly likely with a lesser batch size. As the target model, we perform



(a) Augmentation 0%.

(b) Augmentation 50%.

(c) Augmentation 100%.

(d) Basic NN architecture.

(e) Multiple NN architectures.

(f) Reconstruction attack rounds vs batch size

Fig. 3: The accuracy of the attack vs. the local batch sizes with (a), (b), (c) different augmentation percentages of the target test set, (d), (e) different model architectures of NN and variation of required iterations for reconstruction attack with the number of NN models. (f) Required average attack rounds for reconstruction for varying target model training batch sizes.

our experiments with the LeNet-5 model. Figure 4 provides an example reconstruction made with a trained target model having a batch size of 3 with one epoch for the target model. The average number of rounds taken to reconstruct the images via gradients of the target model is shown in Figure 3f. Here, we observe that the reconstruction of larger batch sizes consumes more time than smaller batches. In an FL system, if the attacker gains access to all the models but does not have information on which model the target data is, they will have to reconstruct all the models. It may be seemingly impossible if millions of models are available. However, a lookup for the potential target can be made via the membership inference, which will significantly reduce the time the attacker takes to recover the facial emotion data.

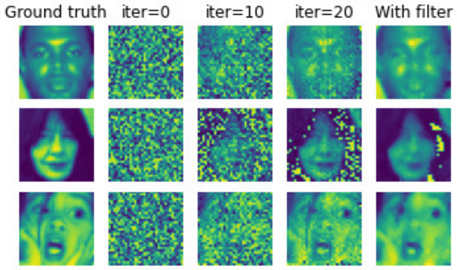


Fig. 4: Example reconstructed dataset of face images for a batch size of 3, with a median filter to reduce noise.

E. Possible Attack Mitigation Techniques

The solutions we proposed in Section IV can be used to mitigate the privacy attack. Perturbation techniques such as differential privacy on the edge and fog [34] can be implemented to reduce the attack accuracy with noisy inputs. To avoid eavesdropping by a malicious entity to track model updates, lightweight encryption [17] can be applied. Privacy should be evaluated in the design phase [29], [30] of these algorithms, and the trade-offs between utility and privacy can be considered when implementing privacy-preserving mechanisms. Existing standardizations such as GDPR, free flow of non-personal data, and acceptable threshold privacy levels should also be considered when designing the algorithms. XAI also has the possibility to detect any abnormalities in the fitted model, as well as provide explanations on the top features used to train the models.

VI. Conclusion

In this paper, we evaluated privacy considerations for the metaverse. The metaverse can combine many technologies which are already existing, yet they should be improved further to achieve privacy expectations with growing threats and vulnerabilities. We show privacy issues can exist and are emerging through new technologies, internal service providers, and external attackers. For example, we introduced a novel metaverse-related potential privacy attack where adversaries can harness users' emotional status by attacking the ML model and recovering the image data from VR sensors. We showed that overfitting, batch size variations and original user data availability can change the attack accuracy in inferring the users' membership, which can be used against the attack. The potential solutions with already existing technologies and tools are also proposed in this paper to address privacy issues.

However, some privacy solutions will be more applicable than others when considering their practicality, maturity, and availability of future related work.

Acknowledgment

This work is partly supported by European Union in SPATIAL (Grant No: 101021808), and Science Foundation Ireland under CONNECT phase 2 (Grant no. 13/RC/2077_P2) projects.

References

[1] J. Kim, "Advertising in the metaverse: Research agenda," pp. 1–4, 2021.

[2] T. Galanti, G. Guidetti, E. Mazzei, S. Zappalà, and F. Toscano, "Work from home during the covid-19 outbreak: The impact on employees' remote work productivity, engagement, and stress," *Journal of occupational and environmental medicine*, vol. 63, no. 7, p. e426, 2021.

[3] G. Liu, Y. Huang, N. Li, J. Dong, J. Jin, Q. Wang, and N. Li, "Vision, requirements and network architecture of 6g mobile network beyond 2030," *China Communications*, vol. 17, no. 9, pp. 92–104, 2020.

[4] Z. Zhang, Y. Xiao, Z. Ma, M. Xiao, Z. Ding, X. Lei, G. K. Karagiannidis, and P. Fan, "6g wireless networks: Vision, requirements, architecture, and key technologies," *IEEE Vehicular Technology Magazine*, vol. 14, no. 3, pp. 28–41, 2019.

[5] N. Stephenson, *Snow Crash: A Novel*. Spectra, 2003.

[6] L.-H. Lee, T. Braud, P. Zhou, L. Wang, D. Xu, Z. Lin, A. Kumar, C. Bermejo, and P. Hui, "All one needs to know about metaverse: A complete survey on technological singularity, virtual ecosystem, and research agenda," *arXiv preprint arXiv:2110.05352*, 2021.

[7] L.-H. Lee, Z. Lin, R. Hu, Z. Gong, A. Kumar, T. Li, S. Li, and P. Hui, "When creators meet the metaverse: A survey on computational arts," *arXiv preprint arXiv:2111.13486*, 2021.

[8] W. Stallings and M. P. Tahiliani, "Cryptography and network security: principles and practice, vol. 6," 2014.

[9] D. J. Solove, "A taxonomy of privacy," *U. Pa. L. Rev.*, vol. 154, p. 477, 2005.

[10] "Art. 4 gdpr – definitions," Mar 2018. [Online]. Available: https://gdpr-info.eu/art-4-gdpr/

[11] M. Finck and F. Pallas, "They who must not be identified—distinguishing personal from non-personal data under the gdpr," *International Data Privacy Law*, 2020.

[12] "Gdpr enforcement tracker." [Online]. Available: https://www.enforcementtracker.com/

[13] Y. Wang, Z. Su, N. Zhang, R. Xing, D. Liu, T. H. Luan, and X. Shen, "A survey on metaverse: Fundamentals, security, and privacy," *IEEE Communications Surveys & Tutorials*, 2022.

[14] V.-L. Nguyen, P.-C. Lin, B.-C. Cheng, R.-H. Hwang, and Y.-D. Lin, "Security and privacy for 6g: A survey on prospective technologies and challenges," *IEEE Communications Surveys & Tutorials*, 2021.

[15] J. Happa, A. Steed, and M. Glencross, "Privacy-certification standards for extended-reality devices and services," in *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 2021, pp. 397–398.

[16] W. Yu, F. Liang, X. He, W. G. Hatcher, C. Lu, J. Lin, and X. Yang, "A survey on the edge computing for the internet of things," *IEEE access*, vol. 6, pp. 6900–6919, 2017.

[17] J. Zhang, B. Chen, Y. Zhao, X. Cheng, and F. Hu, "Data security and privacy-preserving in edge computing paradigm: Survey and open issues," *IEEE access*, vol. 6, pp. 18 209–18 237, 2018.

[18] Y. Sun, J. Liu, J. Wang, Y. Cao, and N. Kato, "When machine learning meets privacy in 6g: A survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 4, pp. 2694–2724, 2020.

[19] X. Liu, L. Xie, Y. Wang, J. Zou, J. Xiong, Z. Ying, and A. V. Vasilakos, "Privacy and security issues in deep learning: a survey," *IEEE Access*, vol. 9, pp. 4566–4593, 2020.

[20] J. Jia and N. Z. Gong, "Attriguard: A practical defense against attribute inference attacks via adversarial machine learning," in *27th {USENIX} Security Symposium ({USENIX} Security 18)*, 2018, pp. 513–529.

[21] T. Orekondy, B. Schiele, and M. Fritz, "Prediction poisoning: Towards defenses against dnn model stealing attacks," *arXiv preprint arXiv:1906.10908*, 2019.

[22] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, "Membership inference attacks from first principles," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 1897–1914.

[23] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," *Advances in neural information processing systems*, vol. 32, 2019.

[24] R. Leenes, "Privacy in the metaverse," in *IFIP International Summer School on the Future of Identity in the Information Society*. Springer, 2007, pp. 95–112.

[25] K. Elkins and J. Chun, "Can gpt-3 pass a writer's turing test?" *Journal of Cultural Analytics*, vol. 5, no. 2, 2020.

[26] A. Paiva, I. Leite, H. Boukricha, and I. Wachsmuth, "Empathy in virtual agents and robots: A survey," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 7, no. 3, pp. 1–40, 2017.

[27] M. T. I. Ziad, A. Alanwar, M. Alzantot, and M. Srivastava, "Cryptoimg: Privacy preserving processing over encrypted images," in *2016 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2016, pp. 570–575.

[28] S. Ribaric, A. Ariyaeeinia, and N. Pavesic, "De-identification for privacy protection in multimedia content: A survey," *Signal Processing: Image Communication*, vol. 47, pp. 131–151, 2016.

[29] A. Cavoukian, "Privacy by design," *Identity in the Information Society*, 2009.

[30] P. Schaar, "Privacy by design," *Identity in the Information Society*, vol. 3, no. 2, pp. 267–274, 2010.

[31] G. D'Acquisto, J. Domingo-Ferrer, P. Kikiras, V. Torra, Y.-A. de Montjoye, and A. Bourka, "Privacy by design in big data: an overview of privacy enhancing technologies in the era of big data analytics," *arXiv preprint arXiv:1512.06000*, 2015.

[32] [Online]. Available: https://iapp.org/resources/article/a-guide-to-privacy-by-design/

[33] A. Alrawais, A. Alhothaily, C. Hu, and X. Cheng, "Fog computing for the internet of things: Security and privacy issues," *IEEE Internet Computing*, vol. 21, no. 2, pp. 34–42, 2017.

[34] M. Yang, T. Zhu, B. Liu, Y. Xiang, and W. Zhou, "Machine learning differential privacy with multifunctional aggregation in a fog computing architecture," *IEEE Access*, vol. 6, pp. 17 119–17 129, 2018.

[35] Y. Qu, L. Gao, T. H. Luan, Y. Xiang, S. Yu, B. Li, and G. Zheng, "Decentralized privacy using blockchain-enabled federated learning in fog computing," *IEEE Internet of Things Journal*, vol. 7, no. 6, pp. 5171–5183, 2020.

[36] A. Kuppa and N.-A. Le-Khac, "Black box attacks on explainable artificial intelligence (xai) methods in cyber security," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[37] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (xai): A survey," *arXiv preprint arXiv:2006.11371*, 2020.

[38] Q. Feng, D. He, S. Zeadally, M. K. Khan, and N. Kumar, "A survey on privacy protection in blockchain system," *Journal of Network and Computer Applications*, vol. 126, pp. 45–58, 2019.

[39] R. Rivera, J. G. Robledo, V. M. Larios, and J. M. Avalos, "How digital identity on blockchain can contribute in a smart city environment," in *2017 International smart cities conference (ISC2)*. IEEE, 2017, pp. 1–4.

[40] J. B. Bernabe, J. L. Canovas, J. L. Hernandez-Ramos, R. T. Moreno, and A. Skarmeta, "Privacy-preserving solutions for blockchain: Review and challenges," *IEEE Access*, vol. 7, pp. 164 908–164 940, 2019.

[41] M. Liyanage, J. Salo, A. Braeken, T. Kumar, S. Seneviratne, and M. Ylianttila, "5g privacy: Scenarios and solutions," in *2018 IEEE 5G World Forum (5GWF)*. IEEE, 2018, pp. 197–203.

[42] L. Zhang-Kennedy and S. Chiasson, ""whether it's moral is a whole other story": Consumer perspectives on privacy regulations and corporate data practices," in *Seventeenth Symposium on Usable Privacy and Security ({SOUPS} 2021)*, 2021, pp. 197–216.

[43] N. Noah, S. Shearer, and S. Das, "Security and privacy evaluation of popular augmented and virtual reality technologies," in *Proceedings of the 2022 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence, and Neural Engineering (IEEE MetroXRAINE 2022)*, 2022.

[44] 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)*. IEEE, 2017, pp. 3–18.

[45] B. Zhao, K. R. Mopuri, and H. Bilen, "idlg: Improved deep leakage from gradients," *arXiv preprint arXiv:2001.02610*, 2020.

[46] J. Zhang, J. Zhang, J. Chen, and S. Yu, "Gan enhanced membership inference: A passive local attack in federated learning," in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*. IEEE, 2020, pp. 1–6.