Toward Zero-Trust Security for the Metaverse

Ruizhi Cheng, Songqing Chen, and Bo Han

The authors demonstrate that conventional federated learning algorithms are not well suited for biometrics-based authentication of virtual reality users, leading to an accuracy of less than 10%. They discuss the root cause of this problem, the associated open challenges, and several future directions for realizing their research vision.

Abstract

By focusing on immersive interaction among users, the burgeoning Metaverse can be viewed as a natural extension of existing social media. Similar to traditional online social networks, there are numerous security and privacy issues in the Metaverse (e.g., attacks on user authentication and impersonation). In this article, we develop a holistic research agenda for zero-trust user authentication in social virtual reality (VR), an early prototype of the Metaverse. Our proposed research includes four concrete steps: investigating biometrics-based authentication that is suitable for continuously authenticating VR users, leveraging federated learning (FL) for protecting user privacy in biometric data, improving the accuracy of continuous VR authentication with multimodal data, and boosting the usability of zero-trust security with adaptive VR authentication. Our preliminary study demonstrates that conventional FL algorithms are not well suited for biometrics-based authentication of VR users, leading to an accuracy of less than 10%. We discuss the root cause of this problem, the associated open challenges, and several future directions for realizing our research vision.

Introduction

Metaverse, with the combination of the prefix "meta" (meaning transcending) and the word "universe," has been deemed as a hypothetical next-generation Internet [1]. While there is no consensus on the definition, a narrow depiction of the Metaverse is a universal virtual world for social interaction, by connecting multiple 3D virtual environments via the Internet. In the Metaverse, users are represented by avatars, allowing them to interact and collaborate with others as if they are in the physical world. To facilitate a fully immersive and seamless experience for users, the development of the Metaverse relies on multiple enabling technologies, such as augmented reality (AR), virtual reality (VR), mixed reality (MR), blockchain, and 5G [1]. Undoubtedly, there will be numerous security and privacy issues in the Metaverse, for example, known attacks on user authentication and impersonation. Also, there would be new types of challenges, for instance, defending against immersive attacks that adjust the location and orientation of users, which may cause collisions with real objects and motion sickness. Hence, we should take security and privacy into consideration from day one when designing the Metaverse instead of treating them as an afterthought.

In this article, we focus on user authentication in the Metaverse and argue that it should satisfy the following four key design principles: *zero trust*, *non-intrusive interaction*, *high reliability*, and *privacy preservation*. For example, zero trust [2] assumes that no entities (e.g., users, transactions, or network traffic) should be always trusted unless verified, as trust itself is a vulnerability. Thus, it requires the pervasive and continuous deployment of security mechanisms, such as authentication and access control [3]. We will discuss the other three design principles later.

While one-shot authentication on head-mounted displays (HMDs) [4] and continuous authentication for mobile devices [3] have been extensively investigated, it is still unclear how to systematically leverage continuous authentication in an immersive environment for effectively supporting zero-trust security. Existing studies bear the following limitations. First, they often utilize a single data source (e.g., head movement [5]), which may not always be available for continuous authentication and usually fails to guarantee a high accuracy (i.e., only around 90 % [5]). Second, they all benefit from collecting biometric information from users, which is highly sensitive and may potentially lead to severe security and privacy issues. Third, they rely on intrusive methods (e.g., throwing a ball [4]) to authenticate users, making them not suitable for continuous authentication. Lastly, they do not consider the practical challenges of continuous authentication on HMDs (e.g., usability and resource utilization).

To address these issues, we explore the emerging security threats and privacy concerns in social VR, an early prototype of the Metaverse [6], and develop a holistic research agenda to protect it with zero-trust authentication mechanisms. In particular, we propose a first-of-its-kind, privacy-preserving continuous and adaptive authentication framework, dubbed MetaGuard (Fig. 1), by employing a federated-learning (FL) based scheme that authenticates users with multimodal biometric data. We build a proof-of-concept of MetaGuard, focusing on preserving user privacy with FL. Our preliminary study reveals that existing FL models such as FedAvg [7] lead to extremely poor performance since each client holds only positive-label data. We discuss the root cause of the issues related to this setup and suggest several potential directions for future improvements. We make the following contributions in this article:
- A comprehensive research agenda for privacy-preserving zero-trust security for the Metaverse.

The authors are with George Mason University, USA.

- Detailed design and proof-of-concept implementation of the proposed framework.
- Potential technical challenges and thorough discussion of future directions for realizing zero-trust security in the Metaverse.

BACKGROUND

ZERO-TRUST SECURITY

Zero trust [2] is a fine-grained security approach that shifts defenses from static perimeters of a protected system to its users, resources, and assets. It is rooted in the principle of "never trust, always verify" [2] and eliminates vulnerable permissions and unnecessary access to help service providers better manage and protect identities, applications, and machines across their networks. Zero trust requires the system to ensure that all requests are continuously verified before accessing any system asset (e.g., continuous user authentication [3]).

METAVERSE

Metaverse could be considered as a large-scale virtual world connected via the Internet. In recent years, with the rapid development of mobile immersive computing [8], providing immersive experiences for users via VR headsets has become one of the most crucial goals of the Metaverse [1]. Therefore, emerging social VR is regarded as a key component of the Metaverse [6]. In social VR, users are free to explore the virtual scene and socialize with others, such as trading virtual content via the non-fungible token (NFT) [6]. Thus, a continuous authentication method is required to protect the account and asset security of social VR users.

VR AUTHENTICATION

There are two main approaches for user authentication in VR: knowledge-based and biometric-based. Since users wear HMDs in VR, knowledge-based methods, such as PINs, become inconvenient and vulnerable to attacks [4]. Therefore, biometric-based authentication has become a more practical option for authenticating VR users [9]. Biometric-based authentication in VR often involves training a classification model using machine learning (ML) or deep learning (DL), as shown in Fig. 2. However, traditional ML/DL models require clients to upload raw data to a server, which raises privacy concerns. To address this issue, utilizing FL for VR user authentication is a promising solution.

THREAT MODEL

In this section, we describe the threat models in VR authentication.

Human Mimicry Attack is a form of security threat in which an attacker records a legitimate authentication and subsequently replays it to gain unauthorized access to a device or account. In the context of VR authentication, an attacker could record a user's interactions with the device, such as body movements, and then mimic the recorded movements to impersonate the user and gain access to the device or account. This type of attack is difficult to detect and prevent, especially in the one-shot authentication, as the authentication information used in the replay may appear to be coming from the actual user. Con-
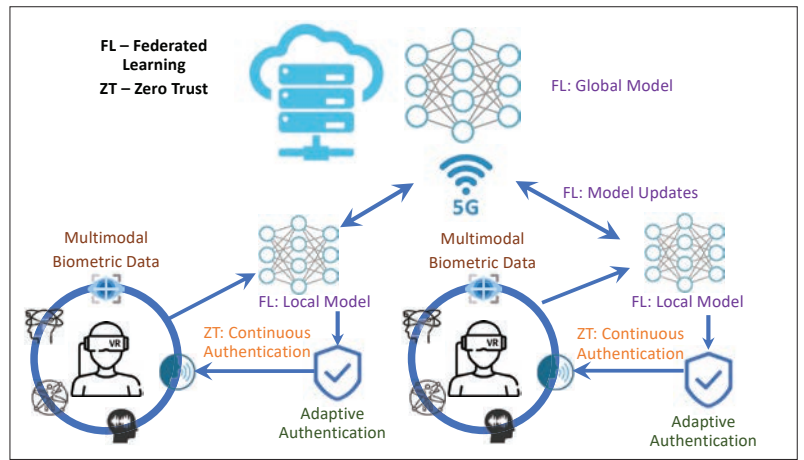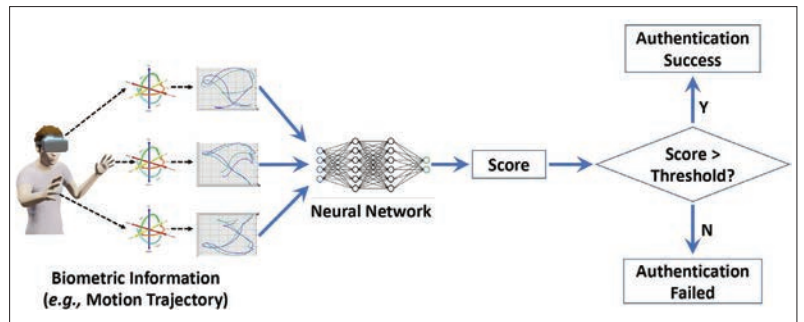

FIGURE 1. System architecture of MetaGuard.

FIGURE 2. Biometric-based authentication process in VR.

tinuous authentication, with its capability to continuously monitor the user's behavior and detect anomalies, can mitigate such attacks by providing an additional layer of security to authentication.

Man-in-the-Middle (MitM) Attack occurs when an attacker intercepts and manipulates the communication between the user and the authentication system. When training the FL-based authentication model, the attackers can intercept the model sent by the server, train the model using their own biometric data, and then upload the updated model to the server. This tampering can result in a model that fails to accurately recognize the legitimate user, which enables the attacker to impersonate the user and gain unauthorized access to the device or account. Thus, one-shot authentication faces a huge risk under this attack. Even continuous authentication alone does not guarantee security since the attackers can continuously use their data to train the model. Hence, it is imperative to design an additional layer of protection in the continuous authentication scheme to defend against this attack.

Data Poisoning Attack is a malicious tactic used by attackers to upload counterfeit data to an ML model with the aim of compromising its performance. In ML-based classification tasks, such as user authentication, this attack involves the injection of incorrect data for certain labels (users), making them unable to be successfully recognized and classified. In the context of FL-based authentication systems, if the counterfeit data is irrelevant to other labels, this attack has a limited impact on the authentication performance of others. However, if the counterfeit data resembles that of other users, such an
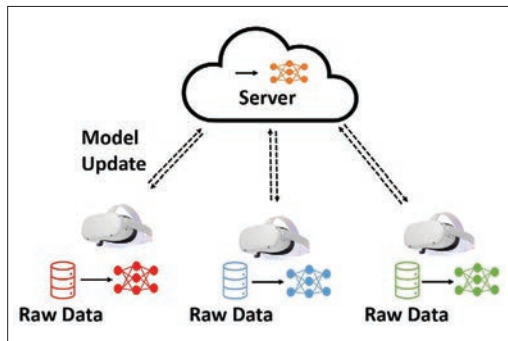
FIGURE 3. A schematic diagram of training FL-based user authentication model in VR.

attack can be considered a variant of the human mimicry attack.

RESEARCH AGENDA

In this section, we propose a holistic research agenda to empower zero-trust authentication for the Metaverse with four concrete steps by presenting the design of MetaGuard.

BENCHMARKING BIOMETRICS-BASED AUTHENTICATION FOR METAVERSE

Early research on authenticating mobile devices [3] and HMDs [5] suggested that collecting physiological (e.g., iris, face, fingerprint, voice, and brain wave) or behavioral (e.g., typing, touch, hand gesture, and gait) biometrics from users to build the ML model would be an effective way to authenticate them. While biometrics-based authentication has been extensively explored, it is still unclear how to effectively adapt it for MetaGuard to offer privacy-preserving zero-trust security mechanisms.

To bridge this gap, we plan to systematically investigate different biometric modalities in terms of their *usability*, *reliability*, *vulnerability*, *collectability*, *sensitivity*, and *adaptability*. For example, continuous authentication in MetaGuard should be non-intrusive, which excludes fingerprint as a candidate, although it is more reliable and accurate than behavioral biometrics, such as gait and hand gestures. Also, behavioral biometrics, such as head and body motion may be more vulnerable to human mimicry attack than others [4], whereas physiological biometrics such as face and voice are more sensitive to the environment [3] and the collection of brain waves is still a challenging task. Moreover, continuously collecting and processing biometrics, such as body and gaze movement, may consume more system resources than others.

To this end, we plan to conduct a large-scale user study by designing a VR environment, where participants will engage in various common activities such as walking, watching videos, and playing games. Meanwhile, we will collect various biometric data using different devices. Hand and head movements can be directly collected from the headset and controllers of VR devices, while collecting gaze movement may necessitate the gaze-tracking device attached to the headset, such as HTC Vive Pro Eye (https://www.vive.com/sea/product/vive-pro-eye/overview/ accessed on Apr. 18, 2022). Moreover, obtaining other physical biometric data, such as heart rate, body temperature, and electrocardiogram, requires the user to wear additional sensory devices. By conducting this study, we aim to gain a deep understanding of the efficacy of different biometric modalities for authentication and to create an open, diverse, and comprehensive dataset for the design and evaluation of MetaGuard. Although in this article we focus on the VR environment as a case study for MetaGuard, our proposed approach can be extended to AR/MR scenarios, as the same type of biometric data such as head and hand motion and eye gaze movement can be collected in AR/MR as well.

LEVERAGING FEDERATED LEARNING TO PROTECT USER PRIVACY

Biometrics-based authentication poses severe privacy concerns to users if such a scheme is blindly employed in the Metaverse, as the collected data often carries sensitive and private information. If biometric data is misused, leaked, or stolen, users could be endangered. To address this concern, we will resort to the recent advances of FL. As shown in Fig. 3, when training the FL model, the sensitive user data will not be uploaded to a central place (e.g., a cloud or edge server). Only the updated gradient of local models will be uploaded to the server for updating the global model, which will, in turn, be used to improve local models. This would effectively mitigate the potential risks and protect user privacy.

To train a user authentication model with FL, an open challenge is that each client holds only positive-label data, which may lead to scalability issues of the model. ML-based user authentication is essentially a binary decision problem. During the training process, the loss function tends to increase the similarity between class embeddings with the same label (L_{pos}) and minimize the similarity between those with different labels (L_{neg}). While in the FL-based authentication scenario, since each client has only its own data (i.e., positive label), the local model cannot optimize L_{neg}. Optimizing only L_{pos} will lead to a trivial solution where all class embeddings collapse into a single point in the feature space [10]. Hence, the model cannot separate feature embeddings among different users, leading the authentication accuracy to decrease as the number of users increases. To address this problem, we propose to utilize the time-series feature of biometric data and recurrent neural network (RNN), as well as contextual information to design the FL-based authentication model.

IMPROVING ZERO-TRUST AUTHENTICATION ACCURACY WITH PERSONALIZED MULTIMODAL BIOMETRIC DATA

The Metaverse necessitates a reliable zero-trust security framework with highly accurate authentication schemes. Given the increasing stake in the virtual world, conducting traditional authentication methods (e.g., password or two-factor authentication) only once at the beginning of the session is no longer sufficient for the Metaverse. With more users joining the Metaverse, the uniqueness of the authentication features for each user is reducing, making it crucial to continuously collect and train the authentication model. Moreover, as we previously discussed, some biometric information, such as voice, has a dynamic nature that can change

over time. Thus, one-shot authentication may not be robust under different environments.

While prior research has proposed continuous authentication [3], a major caveat of these strategies is that, as we mentioned above, they often rely on a single trait of users. However, depending solely on a single biometric modality for authentication in VR may lead to sub-optimal performance [4] and be vulnerable to human mimicry attacks [9]. Moreover, the free-interaction nature of the Metaverse presents challenges to the effectiveness of single-modality authentication. For example, a user may intermittently walk and sit, leaving a scheme that solely relies on users' gait powerless when they do not walk. Similarly, an authentication scheme monitoring users' typing would leave a loophole when they do not behave as expected.

To this end, we will design a multimodal FL model where multiple modalities of biometric data (e.g., pupil size, head motion, body motion, voice, etc.) are collected and fused together for authentication. Moreover, we plan to propose a personalized FL model for each user with a unique combination of modalities that reflects their unique biometric features. This approach will not only improve authentication accuracy but also effectively defend against the MitM attack. This is because users' multimodal biometric data will present unique features. Even if MitM attackers also perform the human mimicry attack, our proposed model is still capable of defending against it since precisely mimicking multiple biometric information is challenging. Furthermore, some biometric data, such as gaze movement, is hard to be mimicked since the user's eyes are covered by the VR headset. However, integrating multiple biometric modalities, particularly when utilizing body movement data, into authentication systems may lead to noise and a decrease in authentication accuracy. To address this challenge, we propose a density-based within-client modality selection method that chooses the best modality combination for each client.

Boosting Usability of Zero-Trust Security with Adaptive Authentication

The Metaverse requires a non-intrusive and easy-to-use authentication that will not impact users' experience while protecting their privacy. Our proposed FL-based framework with multimodal biometric data as input is expected to provide strong protection to end-users. However, HMDs are not well designed to handle computation-intensive tasks, such as FL training and continuous data collection, processing, fusion, and inference, which could inadvertently impact user experience, for example, quickly depleting the battery.

To this end, we will optimize MetaGuard with adaptive and dynamic authentication. For example, a coarse-grain data collection and authentication frequency can be used when users' behavior does not change. Otherwise, intensive authentications will be triggered. Also, we plan to investigate contextual information that MetaGuard can utilize to effectively reduce authentication costs. An inherent and practical challenge of adaptive authentication is to balance the trade-off between reliability, usability, and resource consumption.

For example, in traditional FL algorithms, such as FedAvg [7], the selection weights of all clients are the same by default. In each communication round, the server randomly selects clients to participate in the training. However, clients may present heterogeneous resource and data issues, affecting the training efficiency and model performance. To address this issue, we need to design an adaptive client-selection algorithm that chooses the appropriate clients to jointly train the global model in each round. A unique challenge of MetaGuard is that we cannot simply drop a client for any reason, as everyone should be continuously authenticated in the Metaverse.

Preliminary Results

To better understand the technical challenges of MetaGuard, we build a proof-of-concept that partially implements our proposed research above. In this section, we present our preliminary results of FL-based VR authentication that shed light on realizing the fully-fledged MetaGuard.

Dataset: We use the dataset released by Miller *et al.* [4] for our initial experiments. This dataset was collected with an application of throwing a ball in VR. It contains the trajectories of 41 users throwing the ball using the Oculus Quest VR headset and has six modalities, including the position and orientation of the headset and both controllers. Each user threw the ball ten times per day for two days. We use the first-day trajectories for enrollment (training set) and the second-day trajectories for authentication (test set). Note that leveraging ball-throwing for authentication is intrusive and may not be suitable for continuous authentication. However, given that this is the largest publicly available biometrics-based VR authentication dataset, in terms of the number of users, we believe it can help us gain some initial insights into designing privacy-preserving continuous authentication for the Metaverse.

Implementation Details: We implement three well-known DL models that have been demonstrated to achieve high accuracy in VR authentication tasks [4]: FCN, ResNet, and siamese neural network. FCN and ResNet have a similar architecture, consisting of three convolutional layers and ending with a dense layer using softmax as the activation function. The output of these two models is a set of N scores, where N is the number of users, representing the probability of each candidate being the authenticated user. The candidate with the highest score is then identified as the authenticated user. The siamese network has the same limb as the FCN but does not have a dense layer. It receives two pieces of data as input (e.g., enrollment and authentication motion trajectories), calculates their Euclidean distance, and outputs the distance. The user ID of the enrollment trajectory with the closest distance to the authentication trajectory is the identified ID. The server trains the above three models with all users' data (i.e., trajectories of headset and controllers), compromising their privacy.

Next, we set up and train the FL model that ensures each client utilizes only its own data and does not share it with others, including the server. We leverage a widely used FL algorithm — FedAvg [7] — to enable the server and clients to jointly train the model. The process of training the FL

The preliminary results validate our identified challenges and demonstrate that the FL-based authentication model may not scale due to the fact that each user holds only positive label data.

Model	Privacy Preservation	Accuracy
Siamese Network	✗	90.2 %
FCN	✗	89.3 %
ResNet	✗	87.2 %
FedAvg + FCN	✓	6.34

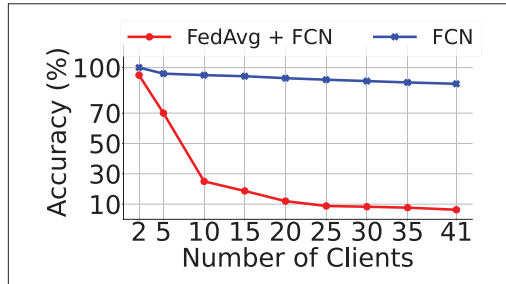TABLE 1. Accuracy of different authentication models.

FIGURE 4. Impact of the number of clients on the FedAvg + FCN and FCN models.

model via the FedAvg algorithm is as follows:
1 In the kth communication round, the server sends the class embedding W_k and model parameters θ_k of the global model to all clients.
2 The tth client updates (W_k, θ_k) to (W_k^t, θ_k^t) based on its local data and the loss function.
3 The server receives the updated (θ_k^t, θ_k^t) from all clients and updates the (W_{k+1}, θ_{k+1}) of the global model by taking a weighted average of (W_k^t, θ_k^t).
4 The server transmits the (W_{k+1}, θ_{k+1}) to all clients and keeps repeating steps 2–4 until the global model is converged.

Based on the above algorithm, the authentication model would not leak user privacy. We use the FCN as the network architecture of the global model for FL, namely FedAvg + FCN. We use the Adam optimizer with a learning rate of 0.001 for training. For the three non-privacy-preserving models, each training runs 2,000 epochs. For the FedAvg + FCN model, we train it with 100 communication rounds and let all clients participate in the training during each round. For each training, we save the model with the best authentication accuracy in the test set. We train each model five times and report the average accuracy.

Results: Table 1 shows the accuracy of different models when using all 41 users' trajectories with all six modalities, which leads to the following two observations. First, on this intrusive authentication dataset, the accuracy of the best non-privacy-preserving model, the siamese network, is only ~ 90%. However, as we discussed earlier, to perform zero-trust continuous authentication, MetaGuard must utilize non-intrusive approaches, whose accuracy may be even lower than the intrusive solutions. For example, Pfeuffer *et al.* [11] leveraged users' gait movement to perform authentication, and the accuracy is only ~ 50%. One possible reason is that it is more difficult to differentiate the features of user data in non-intrusive methods, which introduces a key challenge: *How to improve the accuracy of continuous authentication based on non-intrusive methods to ensure the effectiveness of MetaGuard?*

Second, the FedAvg + FCN method can achieve an accuracy of only 6.34%, while all non-privacy-preserving models have higher than 87% accuracy. The reason may be that FedAvg + FCN cannot distinguish the features of most users, resulting in a low authentication accuracy when the number of users is large. To understand the impact of the number of users on the authentication performance of FedAvg + FCN, we train the FCN and FedAvg + FCN models with different numbers of users using all six modalities and show the result in Fig. 4. As we can see, the accuracy of the FCN model only slightly drops when the number of users increases. However, FedAvg + FCN is significantly affected by the number of users. When there are only two users, it can achieve 95% accuracy. However, when the number of users increases to 10, the accuracy reduces to < 30%; when the number of users further increases to 25, the accuracy is under 10%. Given that the Metaverse will accommodate tens of thousands of users from all over the world [1], the user authentication model must be scalable and robust. Thus, it leads to another key challenge for MetaGuard: *How to ensure the accuracy of MetaGuard remains stable when the number of users increases?*

Next, in order to understand the impact of the number of modalities on authentication performance, we train the FCN and FedAvg + FCN models with different modality combinations. Among the six modalities, there are a total of

$$\binom{6}{1} + \cdots + \binom{6}{6} = 63$$

modality combinations. The result is shown in Fig. 5. We observe that the accuracy of FCN is stable with different numbers of modalities. However, using fewer modalities may result in higher accuracy for the FedAvg + FCN model. For example, when using all six modalities, the accuracy is merely 6.34%. However, when using only three modalities, the average accuracy is 19.32%. Moreover, the best modality combination may be different for different users. During the training phase, we record the modality combinations with the highest local accuracy for each of the 41 users and observe that there are 30 unique modality combinations among the recordings. These findings demonstrate the importance of personalized modality selection for improving the performance of FL-based authentication models in VR.

DISCUSSION AND FUTURE DIRECTIONS

DISCUSSION

The preliminary results validate our identified challenges and demonstrate that the FL-based authentication model may not scale due to the fact that each user holds only positive label data. Next, we discuss the solutions to this problem proposed in existing works and why they do not apply to VR authentication. For example, recent work proposed to address this issue by improving the model update process. FedAwS [12] is a method where the server, in addition to averaging the uploaded gradients, performs a geometric regularization to ensure users' embeddings are separated by a pre-defined margin. FedUV [10] uses an

error-correction code to generate a unique secret vector for each user to update the model. However, these models do not consider the quality of local data (i.e., each client simply uses all of its local data for training). This can be problematic, as low-quality local data, such as noise, can negatively affect the accuracy of the trained model. For instance, as demonstrated in Fig. 5, blindly incorporating all modalities for training can result in suboptimal performance.

Other recent work on face recognition reported an accuracy comparable to the state-of-the-art non-privacy-preserving deep learning models. For example, FedFR [13] proposes to first deploy a feature extractor on the server that has been pre-trained on a large-scale dataset for initiating the class embedding for each client. It then introduces extra data created from the public datasets that have low similarity to the data held by the client for computing its L_{neg}. However, these solutions have the prerequisite of large-scale datasets. This is feasible for face recognition because there are numerous public large-scale datasets available, such as the DigiFace1M (https://github.com/microsoft/DigiFace1M accessed Apr. 18, 2022), which has 1.2M images obtained from 110K individuals. On the other hand, biometric datasets, especially for VR headsets, are limited in their scale. For example, as we introduced before, the dataset released by Miller *et al.* [4], which is the largest among the public biometrics-based VR authentication datasets, has only 41 users.

The root cause for this disparity in size between the face recognition and biometrics-based authentication datasets is that face images are relatively easy to collect, requiring only devices with cameras, and many people voluntarily share their photos on the Internet. Whereas collecting users' biometric data on HMDs requires additional setup for sensors, and users may be reluctant to share biometric data because it could contain sensitive information. Therefore, it is probably unrealistic to expect the publicly available biometric dataset for VR authentication to be as large as the face recognition datasets.

FUTURE DIRECTIONS

Figure 5 highlights a unique challenge associated with employing FL for authenticating VR users. Contrary to traditional authentication scenarios, such as facial recognition or handwriting analysis, which rely on a single modality, applying FL for VR authentication entails the incorporation of multiple modalities to enhance authentication accuracy. Nevertheless, indiscriminate utilization of all modalities may lead to suboptimal performance. Thus, to improve the performance of the authentication model while preserving user privacy, we first propose a personalized density-based within-client modality selection algorithm. Our approach is motivated by the observation that the optimal modality combinations may differ among users and that modalities beneficial for authentication are typically well-clustered. To select the optimal modality combination for each user, we plan to leverage the density-based clustering algorithm for reducing noise and calculating the density for each modality combination. First, after applying the clustering algorithm on each modality combination, the clustered group with only a single point will
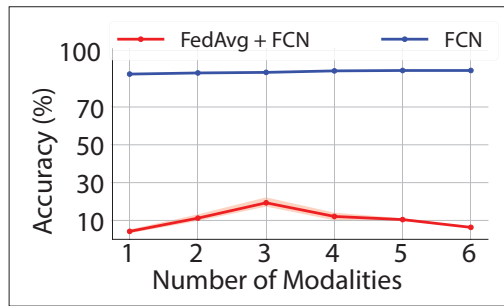
FIGURE 5. The average accuracy of FedAvg + FCN and FCN models using different numbers of modalities for 41 users. The bands represent 95% confidence intervals. For each number of modalities x, it has $\binom{6}{x}$ modality combinations.

be considered as noise and removed. Second, by calculating the distance between the centroids of each clustered group and the center of all data, we will obtain the density of each modality combination. In this manner, each client can select the modality combination with the highest density for training the authentication model.

Next, we exploit the unique features of the biometric data collected on VR headsets to improve the scalability of MetaGuard. We notice an important fact that since all data are captured from the continuous activities of the user, they may present time-series features, which are not available in other authentication methods, for example, those that benefit from face recognition. To leverage the time-series features of biometric data on VR headsets, we can perform time-series analysis to extract data with salient features for authentication. For example, by considering the excellent performance of the RNN, such as long short-term memory (LSTM), in time-series classification tasks, we can train the RNN-based model to improve the accuracy of non-intrusive authentication.

To further improve the scalability of our system, we plan to utilize contextual information in VR to conduct authentication. For example, there is a large number of user-generated content (UGC) in the Metaverse. Thus, we can design a fine-grained data processing method, which categorizes time-series data based on recent user interactions with the UGC. During authentication, the user will be verified using the historical information of their interactions with the UGC. Given that users may exhibit distinct patterns in their interactions with different UGC, this approach can help the model learn more useful representations of the user's behavior by considering mainly the most relevant historical data (i.e., the interaction with the UGC), enhancing the authentication performance.

We finally discuss the potential directions and open research problems for further improving the practicality of MetaGuard. First, to accommodate new users and adjust to the dynamic nature of certain biometric information (e.g., motion trajectories) [9], the authentication model requires frequent training, which may entail significant resource and bandwidth consumption. As we discussed above, clients with insufficient available resources and/or poor data quality may affect the convergence speed of the global model, which motivates us to design a client-selection algorithm to accelerate the convergence of the model, as well as reduce the bandwidth requirement and

> The root cause for this disparity in size between the face recognition and biometrics-based authentication datasets is that face images are relatively easy to collect, requiring only devices with cameras, and many people voluntarily share their photos on the Internet. Whereas collecting users' biometric data on HMDs requires additional setup for sensors, and users may be reluctant to share biometric data because it could contain sensitive information.

> Lastly, despite MetaGuard's ability to protect users' biometric information during authentication, emerging security and privacy concerns remain in the Metaverse.

computation overhead, without sacrificing the authentication accuracy of MetaGuard.

Second, continuous authentication demands non-intrusive data collection. However, given that the data collected in this way may exhibit random patterns, especially for behavioral-based biometrics, training the authentication model using raw data may result in poor authentication performance [11]. This encourages further investigation into encoding strategies for raw data. For instance, calculating movement velocity and acceleration or employing fast Fourier transform (FFT) for feature extraction could potentially enhance authentication performance. Additionally, determining the optimal approach to fuse various types of biometric data for enhancing authentication accuracy, such as at the feature-extraction level, matching-score level, and decision level, remains a topic deserving of in-depth study.

Third, authentication decisions must be made in real-time to safeguard user security. Consequently, we should meticulously design the authentication model. Blindly increasing the complexity of the model may not only prolong the decision-making time but also potentially result in overfitting issues, reducing authentication accuracy. Moreover, offloading decision-making processes to edge devices could potentially reduce the time required to make decisions [14].

Fourth, traditional FL training relies on a centralized approach that necessitates a server for exchanging model parameters with clients, potentially raising privacy concerns [15]. Utilizing edge computing [14] and blockchain techniques presents a promising avenue for realizing decentralized FL training without a central server, enhancing the security and scalability of MetaGuard. During FL training, clients transmit local models to associated edge servers via blockchain transactions. Edge servers collect these transactions and create a block for storage. They then participate in the mining process to verify the new block, attaining consensus among all edge servers. Upon successful mining, the verified block is added to the blockchain and broadcast to local devices, ensuring secure storage of local model updates.

Lastly, despite MetaGuard's ability to protect users' biometric information during authentication, emerging security and privacy concerns remain in the Metaverse. For example, by analyzing headset and controller movements, servers or attackers could potentially deduce users' anthropometric data, such as height and wingspan, to infer demographic information (e.g., gender). Moreover, when users participate in highly immersive applications, such as gaming, servers or attackers might access their fitness, reaction time, and color vision information, enabling them to infer users' age and disabilities. Consequently, developing comprehensive strategies to protect users' private data from being leaked remains a subject warranting further investigation.

Conclusion

In this position article, we presented a holistic research agenda for securing the Metaverse through a zero-trust continuous authentication framework. We first analyzed the challenges of conducting privacy-preserving continuous authentication in the context of social VR (i.e., a prototype of the Metaverse). We then proposed MetaGuard, a first-of-its-kind, FL-based adaptive and continuous authentication framework for VR users by leveraging multimodal biometric data. Our preliminary study through a proof-of-concept of MetaGuard revealed that blindly applying FL to VR authentication will lead to an accuracy of lower than 10%. We finally discussed the root cause of these problems, which shed light on future improvements. We hope our study can inspire more research to realize the grand vision of the Metaverse by tackling its security and privacy challenges.

References

[1] R. Cheng et al., "Will Metaverse be NextG Internet? Vision, Hype, and Reality," *IEEE Network*, vol. 36, no. 5, 2022, pp. 197–204.
[2] S. Rose et al., "Zero Trust Architecture (NIST Special Publication 800-207), https://doi.org/10.6028/NIST.SP.800-207, 2020; [accessed Apr. 18, 2022.]
[3] H. Feng, K. Fawaz, and K. G. Shin, "Continuous Authentication for Voice Assistants," *Proc. ACM MobiCom*, 2017.
[4] R. Miller, N. K. Banerjee, and S. Banerjee, "Using Siamese Neural Networks to Perform Cross-System Behavioral Authentication in Virtual Reality," *Proc. IEEE VR*, 2021.
[5] A. Bhalla et al., "MoveAR: Continuous Biometric Authentication for Augmented Reality Headsets, *Proc. ACM Cyber-Physical System Security Workshop*, 2021.
[6] R. Cheng et al., "Are We Ready for Metaverse? A Measurement Study of Social Virtual Reality Platforms," *Proc. ACM IMC*, 2022.
[7] B. McMahan et al., "Communication-Efficient Learning of Deep Networks from Decentralized Data," *Proc. Int'l. Conf. Artificial Intelligence and Statistics*, 2017.
[8] B. Han, "Mobile Immersive Computing: Research Challenges and the Road Ahead," *IEEE Commun. Mag.*, vol. 57, no. 10, 2019, pp. 112–18.
[9] S. Stephenson et al., "Sok: Authentication in Augmented and Virtual Reality," Proc. IEEE Symposium on Security and Privacy (S&P), 2022.
[10] H. Hosseini et al., "Federated Learning of User Verification Models Without Sharing Embeddings," *Proc. ICML*, 2021.
[11] K. Pfeuffer et al., "Behavioural Biometrics in VR: Identifying People from Body Motion and Relations in Virtual Reality," *Proc. ACM CHI*, 2019.
[12] F. Yu et al., "Federated Learning with Only Positive Labels," *Proc. ICML*, 2020.
[13] C.-T. Liu et al., "FedFR: Joint Optimization Federated Framework for Generic and Personalized Face Recognition," *Proc. AAAI*, 2022.
[14] R. Singh and S. S. Gill, "Edge AI: A Survey," *Internet of Things and Cyber-Physical Systems*, 2023.
[15] P. Kairouz et al., "Advances and Open Problems in Federated Learning," *Foundations and Trends® in Machine Learning*, vol. 14, nos. 1–2, 2021, pp. 1–210.

Biographies

RUIZHI CHENG is currently working toward a Ph.D. degree in the Department of Computer Science, George Mason University, USA. His research interests include federated learning, networked systems, and virtual reality.

SONGQING CHEN is a Professor of Computer Science at George Mason University. His research interests mainly focus on design, analysis, and implementation of algorithms and experimental systems in the distributed and networking environment.

BO HAN is an Associate Professor of Computer Science at George Mason University. His research interests are in the areas of networked systems, mobile computing, and wireless networking, with a recent focus on immersive video streaming and XR.