ISSN: 2710-2599

SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Data Science and Big Data Analytics

Publisher's Home Page: https://www.svedbergopen.com/

Review Article **Open Access**

Data Privacy Preservation with Federated Learning: A Systematic Review

Akinul Islam Jony[1*] and Mubashir Mohsin[2]

[1]Department of Computer Science, American International University-Bangladesh (AIUB), Dhaka 1229, Bangladesh. E-mail: akinul@aiub.edu

[2]Department of Computer Science, American International University-Bangladesh (AIUB), Dhaka 1229, Bangladesh. E-mail: mubashir.mohsin.42884@gmail.com

Article Info

Volume 4, Issue 1, May 2024
Received : 30 January 2024
Accepted : 21 April 2024
Published : 05 May 2024
doi: 10.51483/IJDSBDA.4.1.2024.1-16

Abstract

Federated learning (FL) has emerged as a viable paradigm for decentralized machine learning (DML) across multiple platforms while safeguarding data privacy. This study covers a thorough analysis of FL strategies intended to protect the privacy of data. It investigates the techniques and tactics FL uses to secure data privacy and explores the benefits and constraints of FL privacy protection. Using a methodical approach to the literature review, the study distinguishes FL approaches, explores the nuances of the FL transfer process, assesses current techniques, and identifies inherent vulnerabilities and shortcomings. These outcomes emphasize the vitality FL has for alleviating concerns about privacy while fostering collaborative learning. A variety of FL techniques are identified in the review, each of which contributes a distinct mechanism for maintaining privacy. These include differential privacy, homomorphic encryption, pruning, secure aggregation, secure multiparty computation, and zero-knowledge proofs, among others. This study provides scholars and practitioners with significant perspectives on existing procedures and prospective areas for advancement by integrating ideas from multiple sources to provide an overview of the current FL landscape concerning data privacy protection. The findings are more credible and reliable because of the systematic study, which also provides a strong basis for further research on FL and data privacy protection. At the end of the study, the implications of FL approaches for improving data privacy are covered. The significance of continuing research endeavors to tackle new problems and refine FL techniques for resilient and expandable privacy protection in the distributed machine learning age is underlined.

Keywords: *Federated learning, Privacy preservation, Data privacy, Decentralized machine learning, Systematic review*

* *Corresponding author: Akinul Islam Jony, Department of Computer Science, American International University-Bangladesh (AIUB), Dhaka 1229, Bangladesh. E-mail: akinul@aiub.edu*

1. Introduction

Considering user's privacy protection, collaborative machine learning (Zhang *et al.*, 2020) models are paving the way of a new generational advancement by approaching the likes of decentralized learning and federated learning (Pasquini *et al.*, 2022). To solve the problems associated with large-scale system control, including high dimensions, constraints on the information structure, uncertainties, and data protection, decentralized control systems are proposed (Bakule, 2008). DML describes the practice of distributing machine learning tasks and models over a network of devices, as opposed to storing and processing data on a single server or data center. By dispersing data among several network nodes, DML eliminates the need for centralized data storage (Pasquini *et al.*, 2022). Providing a more safe, effective, and scalable method for developing and implementing machine learning models is the aim of DML, where sensitive data is safeguarded together with the processing load. For privacy solutions, DML has been applied in Healthcare (Gillis, 2022; Rieke *et al.*, 2020; Truhn *et al.*, 2024; Zhou *et al.*, 2024), Finance (Gillis, 2022; Palaiokrassas *et al.*, 2023; Rizinski *et al.*, 2022; Yang *et al.*, 2020a; Zhang and Zhu, 2020), IoT (Gillis, 2022; Khan *et al.*, 2023; Yang *et al.*, 2020a), Advanced Computing Technology (Lim, 2019; Liu *et al.*, 2020; Wang *et al.*, 2020; Zhao, 2020), and many other important sectors. Furthermore, by utilizing the variety of data sources and viewpoints present in a decentralized network, DML can raise the overall accuracy of models (Vergne, 2020). DML uses a variety of techniques, such as peer-to-peer learning (Bellet, 2017), federated learning (Konecny, 2016), and Blockchain-based techniques (Konecny, 2016; Li *et al.*, 2020). More secure, scalable, and democratic access to machine learning models could be made possible by DML, which has the potential to completely change how these models are created and implemented (Wahab *et al.*, 2021). Compared to conventional centralized machine learning (e.g., Jony and Arnob (2024), Lisun-Ul-Islam *et al.* (2023), Tanvir *et al.* (2023)), DML offers several benefits, including improved privacy, efficiency, robustness, and democratization (Vergne, 2020). Data is stored locally on users' devices, which leads to shorter training times and fewer resource requirements. Furthermore, it might be more resistant to mistakes and assaults (Elgabli *et al.*, 2020).

FL is a cutting-edge paradigm in machine learning that will radically transform conventional model training techniques by dispersing the learning process among multiple edge devices and protecting the data inside of them (Asad *et al.*, 2020). As opposed to traditional centralized methods, FL decentralizes the training process, guaranteeing that private information stays localized on individual devices instead of being combined into a single repository (Li *et al.*, 2020). In addition to maintaining user data confidentiality and privacy, this decentralized design also allays worries about computing overhead and data transport costs (Li *et al.*, 2020). FL reduces the requirement for bulk data transfer to a central server, maximizing resource usage and promoting more effective learning procedures by enabling algorithms to be trained locally on the devices where the data is stored. Consequently, FL emerges as a groundbreaking methodology with the potential to revolutionize various domains, ranging from healthcare and finance to IoT and autonomous systems, by harmonizing the imperatives of data privacy with the exigencies of advanced machine learning techniques. The purpose of this study is to thoroughly examine and assess FL's methods for protecting and preserving data privacy. To achieve the goals, this study is opted to appropriately answer the following questions:

- What methods and strategies are currently in use in federated learning to safeguard data privacy?
- What are the benefits and limitations of the privacy protection it offers?

2. Methodology

This study does a thorough and methodical literature evaluation using a combination of conceptual and contextual methodologies (Davaei and Gunkel, 2023). A systematic review is a process of synthesizing information on a topic based on existing literature, for example, the paper by Jony and Serradell-López (2019). The goal of this study is to examine and gather an array of literature on existing methods and frameworks for FL to determine their suitability for preserving data privacy. This will pave the way for future research endeavors that seek to develop a useful FL framework for data security. The chosen method, as outlined by (Davaei and Gunkel, 2023) is used in a systematic review article (Jony *et al.*, 2024), stands out for its effectiveness, consistency, and openness when it comes to research, evaluating the quality of the literature, and compiling findings (Kraus *et al.*, 2022). The concept-context hybrid review is employed, which is useful as it offers comprehensive elucidations of a particular

concept (e.g., federated learning) inside a designated context (e.g., data security) (Davaei and Gunkel, 2023). Planning, conducting, and reporting are the three primary procedures that are covered in this study (Kraus *et al.*, 2022). The study's Section 3 addresses the reporting part of this method. The adoption of qualitative content analysis in the planning and conduct of the review is pertinent due to how it facilitates a heterogeneous corpus of articles to be subjectively categorized into concepts (Kraus *et al.*, 2022). Qualitative analysis is crucial for establishing connections between the pertinent findings, subjects, and concepts of the study to boost comprehension and the range of the literature on FL in data security and preservation. Analyzing and understanding the important elements that have been found concerning data security and privacy protection is made simpler using this technique. PRISMA SLR strategies and guidelines were utilized to specify record identification, screening, and inclusion in some of the most recent systematic literature review (SLR) studies related to FL (Prayitno *et al.*, 2021; Sohan and Basalamah, 2023). The PRISMA SLR standards (Page *et al.*, 2021) are thus being adhered to in the data selection process for this study to conduct the review.

2.1. Planning the Review

Using the resources from Google Scholar, MDPI, ResearchGate, IEEE Xplore, SpringerLink, ArXiv, and PubMed, the literature search was carried out between January 1, 2020, and January 31, 2024. A total of 103 manuscripts were examined. The general search keywords are presented in Table 1:

Table 1: List of Keywords with Number of Papers

Keywords	Number of Papers
"*federated learning*"	31
"*benefits and drawbacks of federated learning*"	7
"*decentralized machine learning*"	23
"*federated learning frameworks*"	11
"*federated learning and data privacy*"	14
"*differential privacy*"	8
"*data privacy in machine learning*"	9
Total Papers Identified Initially (n = 103)	

This study is limited to the domains of data privacy, preservation, and security within the broad context of federated learning. We followed the procedure of selecting the best aligned literature for this investigation through inductive reasoning to perform this systematic review. The criteria for selecting the data to be included in the review's conduct are outlined in the latter subsection.

2.2. Conducting the Review

The focus of our studies is to analyze how FL can protect data privacy with its current methods, and for this, we chose an extensive number of papers relevant to our research topic. The literature found through keyword searches may or may not be related to the topic under investigation. Thus, exclusion criteria have been used for additional screening; the publication is removed if there is no relationship between the study's conclusions and our research topics. Initially, focused keyword searches for terms such as "Federated Learning & Data Privacy", "Decentralized Machine Learning & Data Privacy", "Federated Learning & Differential Privacy", "Data Privacy & Machine Learning", and "Federated Learning & Frameworks" yielded 103 sources in total—journal articles, books, websites, and blog posts. To facilitate the process of verifying the dependability and quality of the evaluated material, only journals with a Scopus index and robust editorial articles, books, websites, and blog posts were taken into account. Iterative refining was used in the selection procedure to filter out duplicate research and extract 84 studies. By concentrating on the titles and abstracts of the literature, 27 pieces of it were eliminated, and 57 works were selected for the collection. After focusing the scope further on research that was directly related to "FL", "Decentralized ML", and "Data Privacy" 46 papers made up the list. Twenty-four papers were chosen after a thematic focus was applied, keeping only those that dealt with

"FL Methods", or "Data Privacy in FL." The 24 papers were chosen based on strict standards for quality, relevance, and significant addition to the theme of federated learning techniques in data privacy protection. The complete procedure for collecting data for analysis is shown in a PRISMA flow selection process diagram in Figure 1. Section 3 of this study provides a detailed review and advancement of federated learning methods for protecting data privacy.

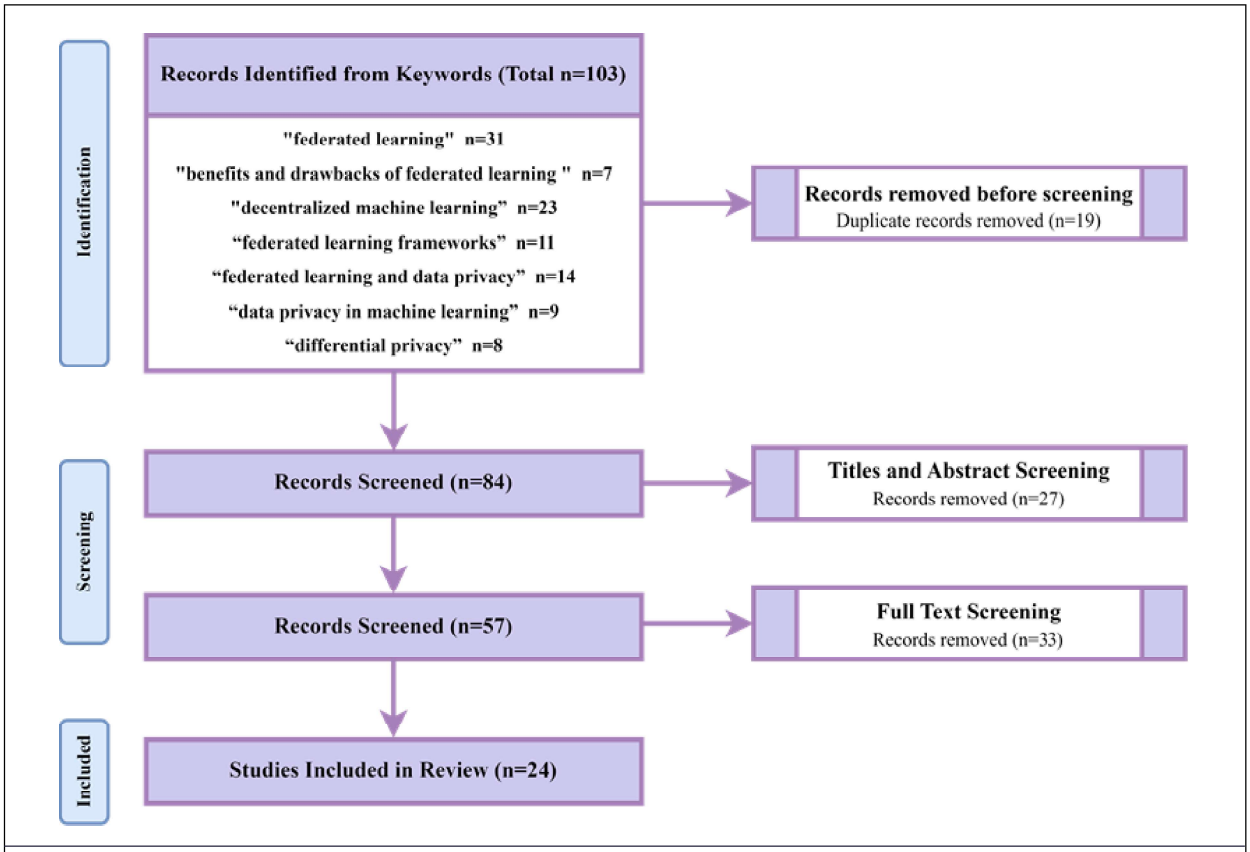

Figure 1: Article Selection Process According to PRISMA Flow Diagram

2.3. Justification of the Selected Literature

This research conducts an in-depth literature review of multiple studies on data privacy in federated learning, looking at different classification systems and approaches used in the FL framework for handling concerns regarding the privacy of data (Gosselin *et al.*, 2022). Diverse techniques are included in FL, such as differential privacy, homomorphic encryption, pruning, secure multiparty computation (SMC), zero-knowledge proof (ZKP), secure aggregation protocols, encrypted aggregation, secure hardware implementations, concealing iterations, a verification framework, gradient noise addition and compression, enlargement of batches, high-resolution data, defense against a malicious server, and FL privacy through blockchain technology. However, our attention is limited to six primary strategies: differential privacy, homomorphic encryption, pruning, SMC, ZKP, and secure aggregation protocols (Gosselin *et al.*, 2022). We investigate the drawbacks, cross-domain applicability, core objectives, and particular strategies used by each approach to protect data privacy guidelines. Interestingly, studies point out that differential privacy is the most widely used strategy, despite its weakness in providing privacy (Rao *et al.*, 2021). Although there are continuous efforts to improve these methods, FL enables calculations to be carried out while maintaining data privacy, which unquestionably provides significant benefits to data privacy (Rao *et al.*, 2021). Our explanation emphasizes the comprehensive analysis and proven effectiveness of the six ways described, even though we do not fully cover other FL approaches since some of the approaches accommodate significant computations while maintaining data privacy. The six strategies outlined here are still some of the most popular ways to protect data privacy in the FL environment; however, other techniques like federated averaging and secure protocol design might be useful based on specific data and computation needs. The works selected for data privacy protection based on various FL approaches are shown in Table 2.

Table 2: List of Federated Learning Approaches

FL Approaches	List of Papers
1. Differential Privacy	(Nguyen and Thai, 2022; Stojkovic, 2022; Wang, 2023; Zhang *et al.*, 2020; Zhou *et al.*, 2024)
2. Homomorphic Encryption	(Iee, 2021; Choi, 2023; G ü nther, 2023; Kishiyama, 2023)
3. Pruning	(Chu *et al.*, 2023; Lin, 2022; Long, 2023; Yu *et al.*, 2021)
4. Secure Multi-party Computation	(Byrd, 2020; Hassani, 2022; Maltitz and Carle, 2018; Yu and Cui, 2022)
5. Zero Knowledge Proofs	(Ahmadi and Nourmohammadi, 2023; Gvili *et al.*, 2021; Liu *et al.*, 2021; Wang *et al.*, 2023b; Xing *et al.*, 2023)
6. Secure Aggregation	(Bonawitz *et al.*, 2016; Feng *et al.*, 2023; So *et al.*, 2021)

3. Federated Learning: Approaches for Data Privacy

3.1. Classification of Federated Learning

FL offers a wide range of classification methods that characterize its applications and approach. One of the classification approaches is to distinguish between vertical, horizontal, hybrid, and federated transfer learning based on how data is distributed among devices or servers. While horizontal federated learning entails many entities with similar data types working together to build models without jeopardizing individual privacy, vertical federated learning permits diverse entities to interact without revealing sensitive information. Transferring information across associated operations or contexts while maintaining data privacy is made possible through federated transfer learning, which extends classic transfer learning paradigms to federated settings. These classification systems preserve data privacy while enabling researchers and practitioners to gain insights into the optimization and implementation of federated learning for practical applications in protecting data privacy. The latter subsection gives a proper review of FL classifications in data privacy preservation mentioned in the selected literature and extends the perspective of dealing with data privacy in various ways.

Horizontal Federated Learning (HFL): HFL is a federated learning method where multiple parties with the same local dataset but different samples train a model, especially useful when data is distributed across different locations or devices but all data points have the same features (Mori *et al.*, 2022). When data sets are similar or share properties among nodes, it is often more appropriate (Nguyen and Thai, 2022; Yang *et al.*, 2020b). The privacy issues related to centralized databases are removed by using this technique, which retains data on the original device (Zhu, 2021). While HFL's clients are presumed to be trustworthy, the HFL system makes the assumption that attacks typically originate from uninvited and untrusted cloud servers (Asad *et al.*, 2020).

Vertical Federated Learning (VFL): In VFL, multiple entities with divergent information about an identical user base collaborate to train machine learning models without revealing their unprocessed data or model parameters. Vertical partitioning of the datasets in VFL results in distinct feature sets being held by several parties for the same collection of entities (Liu *et al.*, 2022; Zhu, 2021). This is typical in situations where several organizations gather different kinds of information about a similar population (Asad *et al.*, 2020). Even though this method is more complex, there are situations in which it can be useful since it permits multiple parties to collaborate on data modeling without compromising individual privacy (Li *et al.*, 2023).

Federated Transfer Learning (FTL): In the context of federated learning, transfer learning is transferring and exploiting the knowledge from one domain (the source domain) to another (the target domain) (Razavi-Far *et al.*, 2022). Transferring knowledge from the source domain to the target domain can greatly enhance the learning process, especially in situations where the destination domain has insufficient data (Zhu, 2021). The issue of domain shift, which can arise when training models on data from several sources, has been addressed

by proposing this approach. With the protection of data privacy, this technique enables models to leverage insights from one dataset to train more effectively on another (Asad *et al.*, 2020).

Hybrid Federated Learning: Hybrid FL is a recently proposed distributed machine learning paradigm for handling private and decentralized data sets. Partial overlaps between feature space and sample space are addressed in this configuration (Zhang, 2020). To address full and partial featured data, it first establishes a new model-matching-based problem formulation for the process. It next provides an effective method that can jointly train the global and local models (Elbir, 2021). The ability to handle situations where the feature space and the sample space partially overlap across many clients makes this parameter crucial in real-world applications. This is typical in situations where many institutions gather disparate kinds of information about the same people (Zhang, 2020).

3.2. Federated Learning Transfer Process

Data Sampling: In federated learning, only a portion of the dataset is sampled and shared for cooperative training, involving data sharing and model training. Restricting the release of particular data points lowers the possibility of re-identification or reverse engineering, thereby lowering the danger of disclosing sensitive information (Zhu, 2021). Data sampling allows entities to contribute without giving up complete control over their datasets, addressing concerns about data security and privacy. In situations where it would be impracticable or resource-intensive to send the complete dataset, this method maximizes computing efficiency and resource use. In general, data sampling is a useful and successful federated learning strategy that strikes a balance between maintaining data privacy and facilitating collaborative model training (Yu *et al.*, 2021).

Data Perturbation: This method includes adding noise or making changes to the data before distributing it to other people (Zhu, 2021). Sensitive information privacy is protected because data perturbation masks the exact values of individual data points (Ding *et al.*, 2023). This method makes sure that the original data points cannot be reliably identified, even in the event that the data is intercepted or viewed by unauthorized parties. The fundamental patterns and trends in the data are unaffected by the introduction of noise, enabling the model to be trained successfully (Zhang, 2023).

Data Encryption: Cryptographic algorithms are used to encrypt data before sharing it with outside parties. A ciphertext format, which can only be unlocked with the right decryption key, is created when the original data is encrypted (Zhang, 2023). Inaccessibility to unauthorized individuals or entities (such as cyber attackers can invade information security (Jony and Hamim, 2023)) guarantees the privacy and security of the information. Federated learning frameworks minimize the possibility of breach of information or unwelcome intrusion by assuring end-to-end secrecy through data encryption (Zhu, 2021).

Data Masking: This method hides or substitutes dummy values for sensitive data points before disclosing them to other parties. Individual confidentiality is maintained by concealing sensitive information, such as personally identifiable information (Zhu, 2021). In addition, the model can be trained on the residual data in parallel, ensuring that important patterns and insights are recorded without jeopardizing data privacy. Data masking allows federated learning systems to obtain valuable insights while safeguarding personal data by forging a compromise between privacy preservation and model efficacy (Zhu, 2021).

Participating in numerous important strategies aimed at maintaining data security, privacy, and effectiveness in responding to various threats is what all these processes have in common. Furthermore, these methods demonstrate resilience when evaluated on real-life data, proving stability and flexibility in response to different types of incursions. This progresses the Robust Federated Learning (RFL) technique in protecting privacy during training in different environments and architectures that the framework is unfamiliar with. RFL takes internal dangers into account, in contrast to SFL techniques that ensure the accuracy of computation results and shield the system from outside threats (Zhang, 2023). To highlight the significance of data processing for data privacy, Table 3 displays the relationship between data processing techniques and FL approaches.

3.3. Federated Learning Approaches in Data Privacy

We restricted our analysis to six main techniques in the field of federated learning: differential privacy, homomorphic encryption, pruning, SMC, ZKP, and secure aggregation protocols. These methods cover a

Table 3: Relation Between Data Transfer Methods and FL Approaches

Data Transfer Method	Federated Learning Approaches
Data Sampling	Secure Aggregation
Data Perturbation	Differential Privacy, Secure Multiparty Computation
Data Encryption	Homomorphic Encryption
Data Masking	Differential Privacy, Secure Multiparty Computation, Secure Aggregation, Pruning

variety of approaches and methods intended to protect data privacy and confidentiality throughout the collaborative training procedure. The topic at hand includes a breakdown of the limitations of each strategy, how well it works in various contexts, what its key goals are, and the methods used to protect privacy requirements. We hope to clarify the complex workings of these methods and any obstacles they might face in practical application by closely examining them all. To enhance comprehension of the landscape of federated learning techniques and their implications for data privacy and security in distributed computing systems, it is also our aim to offer insights into the specific domains in which each approach performs well.

3.3.1. Differential Privacy

Differential privacy is one technique in which a randomized process is deemed differentially private if changing one input element results in only a minor variation in the output distribution. This means that no inferences can be made regarding whether or not a certain sample was used in the learning process. Due to the fact that increasing noise may erode accuracy, there is an inherent trade-off between achieving a high level of model accuracy and utilizing differential privacy (Yang *et al.*, 2020a). By protecting personal information on user devices, guaranteeing individual privacy, and combining client data in a reliable setting, federated learning and differential privacy improve privacy (Stojkovic, 2022). This strategy improves overall privacy protection, alleviates worries, and minimizes the possibility that models hold memorized personal data. Works that specifically address differential privacy include PrivColl (Zhang *et al.*, 2020), QMGeo (Wang, 2023), PPML-Omics (Zhou *et al.*, 2024), FedRKG (Yao *et al.*, 2024), and lots more, often in conjunction with other methods. PrivColl by (Zhang *et al.*, 2020) addresses privacy issues in collaborative learning by utilizing lightweight additive secret sharing approaches to improve model accuracy while reducing computing overhead. It does, however, criticize these approaches for either significantly increasing computational and communication costs or jeopardizing model correctness. In response, the framework trumps traditional techniques such as differential privacy and safe multiparty computation. On the other hand, QMGeo (Wang, 2023) proposes a stochastic method called stochastic quantization. It utilizes a mixed geometric distribution to bolster privacy preservation within frameworks employing differential privacy (DP). By employing randomization for DP without introducing additive noise, this approach demonstrates that, despite the limitations of current methods, quantization techniques can enhance both privacy preservation and communication efficiency within federated learning (FL) frameworks. FedRKG (Yao *et al.*, 2024) is a federated recommendation system that addresses privacy issues in FL recommendation systems by utilizing a global knowledge graph (KG) built from publicly accessible item data. Utilizing Local Differential Privacy (LDP) and pseudo-labeling, it ensures privacy preservation while out-performing centralized algorithms in terms of performance. The PPML-Omics (Zhou *et al.*, 2024), however, offers a privacy-preserving technique for omics data analysis that employs a decentralized differential private federated learning algorithm to protect patient privacy in healthcare. In comparison to previous approaches, this study delivers strong privacy assurances. By using decentralized frameworks for privacy preservation and competitive performance, both strategies emphasize the crucial role that differential privacy plays in FL. Even if the domains of these studies differ, their foundational objective, which is to assure data privacy preservation using differential privacy strategies, underlines the significance of data privacy.

3.3.2. Homomorphic Encryption

Homomorphic Encryption (HE) is a type of encryption that allows operations on encrypted data without decryption. This cryptosystem converts data into ciphertext for machine learning. It converts data into

ciphertext that can be examined and used as if it were still in its original form (Gillis, 2022). There are three common types: Partial HE (PHE), Some-what HE (SHE), and Fully HE (FHE) (Kishiyama, 2023). PHE, introduced in 1978, allows operations while data remains encrypted, but it's the least secure. SHE, also from 1978, can perform addition and multiplication but introduces errors with each operation. FHE, a newer scheme proposed in 2009, allows unlimited mathematical operations on encrypted data, but it incurs high computational costs and large ciphertexts. Despite its challenges, FHE controls the size of mathematical operations to prevent unmanageable growth. It offers several advantages, particularly in cloud services, where it preserves privacy and allows data sharing without compromising sensitive information. However, encryption in these services can be complex and time-consuming. Fully Homomorphic encryption (FHE) is also criticized for its slow processing speed, making it impractical for certain applications, despite its potential benefits for complex datasets. A recent HE- based framework, HElium (Gunther, 2023), has native support for Proxy Re-Encryption (PRE), increasing its effectiveness for private data analysis in multi-party environments. By automating parameterization and circuit optimizations, it reduces overhead and makes FHE accessible to non-cryptographers. Utilizing homomorphic encryption, Blind-Touch (Choi, 2023) is a machine-learning-driven fingerprint authentication system that addresses privacy issues in cloud and web contexts. It leverages a clustered server design for scalability, optimizes feature vectors for distributed architectures, and compresses authentication results. Transactional anonymity problems related to simple stealth address systems are avoided by the additional effort of HE-DKSAP (Yan, 2023), a privacy-enhancing protocol for blockchain transactions. For programmable blockchains, it provides a workable way to protect transaction privacy.

3.3.3. Pruning

Pruning is a strategy used in neural networks to improve efficiency by reducing size and complexity, eliminating redundancies and over-parameterization, and thereby reducing training time (Yu *et al.*, 2021). Federated pruning maintains accuracy through frequent pruning at various stages, further optimizing computing and communication costs in federated learning systems (Lin, 2022). Pruning can help avoid data breaches and safeguard privacy by deleting components that are susceptible or redundant. It can also strengthen the resistance of the model against adversarial attacks and model poisoning by eliminating weights and neurons that are obsolete (Yang *et al.*, 2020a). Although pruning keeps computing costs down without compromising accuracy, its time-consuming nature makes practical use challenging. Numerous fields, including computer vision, reinforcement learning, neural networks, and natural language processing, use this method extensively (Huang *et al.*, 2023). FedDIP by (Long, 2023), an innovative federated learning system, streamlines parameter exchange while retaining accuracy using error feedback and dynamic model pruning. With incremental regularization, FedDIP achieves severe model sparsity and outperforms other pruning techniques. Another framework (Liu, 2023) uses partial model pruning and personalization to improve learning accuracy on resource-constrained devices. As demonstrated by testing findings demonstrating substantial cost savings compared to typical federated learning approaches, the model's split into global and personalized components minimizes computation and communication latency. PriPrune, as put forth by (Chu *et al.*, 2023), presents privacy-aware pruning in federated learning with a maximum limit on information leakage. PriPrune achieves a better balance between privacy and precision than other approaches by using customized defensive masks and varying the pruning rate. In addition (Ma *et al.*, 2023) recommends using neural network model pruning in conjunction with federated topic modeling to safeguard data privacy and accelerate model training. Notably, the two strategies it suggests for figuring out the pruning rate to strike a compromise between training speed and model inference time allow for notable increases in training speed without sacrificing model performance. Although pruning has a lot of potential in the realm of data privacy with FL and cutting-edge frameworks and methodologies demonstrate this, there are a few possible drawbacks to using pruning. Firstly, pruning could make communication overhead worse since it would demand more delay and bandwidth when trimmed models or update parameters are exchanged across dispersed nodes. Secondly, federated learning systems may become less efficient and scalable because of the complexity added by incorporating pruning algorithms. Thirdly, pruning raises security and privacy problems since it can unintentionally reveal private information and expose itself to privacy assaults. Moreover, the performance of federated learning settings with numerous data sources may be negatively impacted by excessive pruning, as it could lead to the loss of significant model

descriptions. By emphasizing area-based implementation, more recent frameworks and research can help overcome these drawbacks.

3.3.4. Secure Multiparty Computation (SMC)

SMC is a technology that improves privacy and security in federated learning processes by allowing distant entities to collaborate without revealing confidential data (Maltitz and Carle, 2018). It makes aggregate statistics or model updates easier while maintaining the privacy of individual data contributions through the use of cryptographic protocols for secret sharing and secure function evaluation (Yu and Cui, 2022). SMC can improve data privacy, ease regulatory compliance, and reduce data breaches and unwanted access in federated learning. This integration enhances trust among participants and facilitates the broader adoption of federated learning frameworks. The work of (Hassani, 2022) introduces PHY-Fed, a new framework for improving FL algorithms that focuses on fixing problems in existing FL methods, especially those using SMC or differential privacy. PHY-Fed helps protect data privacy by reducing the chances of drawing wrong conclusions and lowering the accuracy of traditional FL methods. The research discussed in (Byrd, 2020) examines the combination of SMC and differential privacy within federated learning to enhance data privacy in financial applications while maintaining model accuracy. In financial firms, this method fosters confidence and collaboration by reducing the coordinating server's capacity to extract sensitive client information. It is applicable to activities such as fraud detection, credit origination, and efficient trade execution. While addressing present shortcomings in data privacy techniques, the implementation of SMC in FL can help strengthen data privacy in critical industries like financial applications, as pointed out in (Hassani, 2022) and (Byrd, 2020). There may still be some obstacles to fully guaranteeing data privacy even with SMC, which bring encouraging opportunities for enhancing privacy protection. Potential information leaks during the aggregation process and difficulties balancing the trade-off between model accuracy and privacy preservation are a couple of these issues. Notwithstanding these challenges, the addition of SMC and differential privacy to FL frameworks holds potential for enhancing data privacy and opening the door for innovative data privacy research.

3.3.5. Zero-Knowledge Proofs (ZKP)

Zero-knowledge proofs are a type of cryptography where an entity can prove to another entity that they know a given value 'x' without giving away any additional details except that they know it. In the context of federated learning, ZKPs serve as a pivotal tool for ensuring data privacy as they allow computing processes to be verified without requiring access to plain-text local data (Xing *et al.*, 2023). This method is particularly important when there is strict privacy legislation or standards that necessitate additional protective measures to reduce the possibility of private data being inadvertently disclosed. Scholars can benefit greatly from ZKPs by using them to verify computations, establish reliability, and safeguard data privacy by demonstrating the accuracy of their findings without disclosing private information or compromising privacy (Nguyen and Thai, 2022). Nowadays, ZKPs are utilized for identity verification, voting and password privacy, and membership authentication since they preserve anonymity for users (Yoaquim, 2022). While its application is expanding quickly, it is also a widely employed cryptographic approach in the Bitcoin sector (Xing *et al.*, 2023). RiseFL, described in the paper (Zhu *et al.*, 2023), is a novel approach to safe and verifiable data collaboration in federated learning environments. It uses secure aggregation with ZKP to handle input privacy and integrity concerns. While tested on real-world data, RiseFL surpasses cutting-edge baselines such as ACORN, RoFL, and EIFFeL, providing faster client computation. RiseFL introduces a probabilistic integrity check method and a hybrid commitment scheme for improved Byzantine robustness. zkDFL, in the work (Ahmadi and Nourmohammadi, 2023), is a unique technique that leverages ZKP to allow clients to communicate model parameters with a trustworthy server, guaranteeing the integrity and transparency of the aggregation process, hence improving privacy preservation and scalability in decentralized federated learning systems. In addition to saving gas expenses, zkDFL improves the aggregation process's scalability, verifiability, and privacy enforcement. To determine its limitations and suitability for use in practical situations, further study was recommended. Using the Fiat-Shamir heuristic and the MPC-in-the-head paradigm, the work in (Gvili *et al.*, 2021) of TurboIKOS provides a zero-knowledge argument for general arithmetic circuits, reducing communication per multiplication gate. It works well with Picnic-style post-quantum digital signatures;

however, there are additional research requirements specified. In the work of zkCNN (Liu *et al.*, 2021), the protocol utilizes a sum- check for fast Fourier transforms, ensuring the integrity of machine learning predictions without leaking model information, resulting in significant efficiency improvements while further study is needed. The NuLink platform by (Pawn *et al.*, 2024), facilitates safe data storage, computation outsourcing, and data exchange by integrating ZKP technologies for privacy and security in decentralized applications. ZKPs ensure that all nodes operate decently by promoting trust and discouraging deception. In contrast, the BOOMERANG protocol (Ankele and Haddadi, 2024), which has proven effective in managing high user numbers, suggests a decentralized incentive system that preserves privacy by utilizing cryptographic accumulators and ZKPs to assure verifiability and privacy preservation in incentive systems. Nevertheless, additional investigation is necessary to evaluate their scalability, compatibility with current systems, and any drawbacks in practical use. Aside from these, some noteworthy works are in (Wang *et al.*, 2023a) regarding ZKP mixers in the context of blockchain privacy solutions, in (Wang *et al.*, 2023b) regarding the use of ZKPs to mitigate the vulnerability of FL systems to centralized aggregators, and in (Berke *et al.*, 2023) regarding a ZKP-based tax disclosure system that enables stakeholders to share specific information while keeping virtually nothing. Despite their advantages, latency costs and translation are significant challenges in ZKP transactions, which rely on mathematical formulas and require time-consuming processing. These constraints hinder scalability, efficiency, and data management, especially for geographically distributed parties, posing significant challenges in transaction processing. To address these constraints and potential mitigating strategies, comprehensive and long-term investigations are therefore required for ZKP.

3.3.6. Secure Aggregation Protocols

Protecting privacy and promoting cooperative learning are two important functions of Secure Aggregation Protocols in Federated Learning. They make it possible for several parties to add up their values without sharing them, each managing private data (Bonawitz *et al.*, 2016). In situations where stringent privacy protection is necessary to avoid the inference of personal information, this ensures that the server can obtain the collective model of users without gaining access to their individual models (Nguyen and Thai, 2022). In a single training round, traditional secure aggregation methods give individual user privacy top priority. However, this technique may result in noteworthy privacy weaknesses spanning numerous training rounds. To address this issue, emerging frameworks like Multi-RoundSecAgg provide multi-round privacy assurances to improve overall data protection (So *et al.*, 2021). A secure aggregation protocol for FL called FSSA (Luo *et al.*, 2023) tackles faults in current techniques while increasing efficiency and lowering dropout rates. It ensures data privacy and keeps security preserved even in the event of client dropouts. By covertly aggregating user-trained models, EPPDA (Song *et al.*, 2023) is a privacy-preserving data aggregation approach for FL that ensures privacy. Because of its fault tolerance, which minimizes the amount of computation and communication resources needed, it can continue to function even when certain participants disconnect. Using FL in conjunction with the BERT model for sequence classification in VANETs, FL-BERT (Ahsan *et al.*, 2024) is a unique intrusion detection technique. When using individual devices for model aggregation, it safely stores sensitive data to protect data privacy. Sentinel (Feng *et al.*, 2023), on the other hand, offers a three-step aggregation mechanism as a defense against poisoning attempts in decentralized FL. Both strategies concentrate on data security and privacy in FL and provide complimentary answers to various data security issues with secure aggregation methods in federated learning. Prioritizing the usage of public randomness to reduce communication rounds and per-user expenses is the work in (Van Kempen *et al.*, 2023), LiSA, yet it may encounter issues with scalability and committee selection. On the other hand, AHSecAgg (Zhang *et al.*, 2023) uses additive homomorphic masks to mitigate privacy issues in FL while preserving model accuracy and minimizing computation complexity. Furthermore, the TSKG technique improves secure aggregation in cross-silo situations, doing away with the need for secret sharing during aggregations and demonstrating higher computational effectiveness than existing protocols. Still, scalability and robustness problems exist, especially in highly dynamic or hostile contexts, leading to the need for more practical research.

3.4. Risks and Vulnerabilities of Federated Learning

Federated learning is quickly becoming a ground-breaking idea for protecting valuable and sensitive information in today's data-rich environment. Federated Learning's collaborative model training approach

presents several issues, especially regarding anonymity, but it also shows promise for protecting data privacy. In a review paper by Ding *et al.* (2022) outlined the primary challenges and chances to improve federated learning. Among the noteworthy issues raised are the complexity of memory and time, the risk of catastrophic forgetting, data incompleteness, complicated data connections, and data leakage that compromises privacy. Since both accuracy of learning and raw data security depends on collaborative learning, they are yet unreached requirements in terms of protecting data privacy (Nguyen and Thai, 2022). As worrying as these problems are, there may be a more immediate one to make sure a few others come before. The study carried out (Zhang *et al.*, 2021) tackles several major issues, including trustworthiness, system heterogeneity, insufficient computational end-devices, narrow network bandwidth, and limited on-device resources. Furthermore, none of these problems or difficulties are directly related to the others; therefore, no remedy might lessen the difficulties as a whole. Numerous federated mechanisms have been proposed to address these obstacles to the greatest extent possible, but they continue to raise concerns about the vulnerability of data privacy protection.

In the differential privacy approach, centralizing data collection helps reduce privacy leakage, prevent data imbalance, and enhance fairness. However, it overlooks concerns such as computational overhead, data poisoning, backdoor attacks, and reliance on assumptions. Conversely, homomorphic encryption enhances data security but shares many vulnerabilities with differential privacy, exacerbating key management challenges. Pruning addresses communication overhead, bandwidth constraints, and privacy concerns but introduces challenges with non-IID data and compromises performance and interpretability. Secure multiparty computation offers enhanced privacy protection compared to other methods, but it introduces additional challenges such as protocol complexity, scalability issues, and trust biases. Zero-knowledge proofs are relatively less susceptible to poisoning and backdoor attacks, thereby improving data security. However, they significantly increase computational overhead, bandwidth latency, technical complexity, and the risk of catastrophic forgetting. On the other hand, secure aggregation techniques are prone to issues such as reliance on aggregation type, compromised anonymity, trust biases, and scalability limitations. Most approaches face vulnerabilities due to the computational and communication overheads associated with their complexity and are ill-equipped to handle emerging challenges arising from their heterogeneity. The most important and enduring problems, aside from everything else, are Privacy Leakage, Computational Overhead, Communication Overhead, Non-IID Data, Interpretability, and Scalability. Table 4 provides a summary of the vulnerabilities that the FL methods under discussion are susceptible to.

Table 4: Vulnerabilities of FL Approaches

FL Approach	Privacy Leakage	Computational Overhead	Communication Overhead	Non-IID Data	Interpretability	Scalability
Differential Privacy	✓	X	X	✓	X	X
Homomorphic Encryption	X	✓	✓	X	X	X
Pruning	X	✓	X	X	✓	✓
Secure Multiparty Computation	✓	✓	✓	X	X	✓
Zero Knowledge Proofs	✓	✓	✓	X	X	X
Secure Aggregation	✓	✓	✓	X	X	X

Note: ✓ - The Approach is prone to vulnerability; X - The approach is not prone to vulnerability.

FL techniques have a significant impact on protecting data privacy, yet they encounter numerous challenges such as cost-effectiveness, bandwidth limitations, communication resources, privacy protection, parameter tuning, specialist requirements, and deployment. Differential privacy demands rigorous parameter tuning to establish a balance between privacy and accuracy, though the inclusion of random noise can have an unfavorable effect on model precision. Even though homomorphic encryption provides strong data privacy

without noise, it has a large computational overhead and requires a lot of bandwidth and communication resources. Pruning approaches aim to reduce model complexity and communication costs, but they may compromise model interpretability and efficiency. SMC methods allow for cooperative computing while maintaining data privacy, although they may come with significant computational and communication overhead. On the other hand, ZKPs provide strong privacy guarantees free from noise, but they come with a high computational burden and significant proof transmission band- width cost. Secure aggregation methods combine model updates and protect data privacy; however, they can cause issues with anonymity compromises and scalability constraints. To enable safe and efficient aggregation, it is critical to address scalability issues and make sure resources are allocated effectively, as well as ensuring cost effectiveness and minimizing communication overhead while implementing differential privacy. A summary of the value of each FL approach's constraints is provided in Table 5.

Table 5: Constraints of FL Approaches

Constraints	DP	HE	Pruning	SMC	ZKPs	SA
Cost Effectiveness	H	H	L	L	L	L
Bandwidth	L	L	L	H	H	H
Communication Resources	L	L	L	M	M	M
Privacy	H	H	L	H	H	L
Careful Parameter Tuning	H	L	L	H	H	L
Specialist Requirement	H	H	L	L	L	L
Deployment	L	H	H	M	M	M

Note: H - High; L - Low; M - Moderate.

4. Conclusion

To sum up, this review study accomplishes the primary purpose of exploring federated learning techniques aimed at enhancing data privacy. Through careful consideration of the research objectives presented in the introduction, this study spans different FL techniques, examines the nuances of the FL transfer process, assesses current methodology, and identifies relevant risks and weaknesses. By employing a methodical approach to literature review, the work carefully looks at a large number of pertinent studies, which strengthens the validity and reliability of its findings. This establishes a strong basis for subsequent studies in FL and data privacy preservation, in addition to providing a thorough overview of the context of FL's data privacy protection system at present. This work is significant because it provides scholars and practitioners with useful insights into current methods and identifies future directions for innovation and progress in the field of FL-based data privacy protection. Moreover, the results of this review study highlight FL's potential as a viable strategy for protecting data privacy in distributed learning contexts. FL has the potential to become a key component of privacy-preserving machine learning with improvements in FL techniques and a better comprehension of associated hazards. Future work on novel ways to improve FL's privacy guarantees might look into including safe multiparty computation, homomorphic encryption, or differential privacy methods. Furthermore, more research into resolving FL's scale problems and non-IID data concerns could enhance the technology's effectiveness in protecting user privacy across a range of applications and domains. All things considered, this work advances the field of FL for protecting data privacy and offers insightful information for future studies in this area.

References

Ahmadi, M. and Nourmohammadi, R. (2023). ZKFDL: An Efficient and Privacy-Preserving Decentralized Federated Learning with Zero Knowledge Proof.

Ahsan, S.I., Legg, P. and Alam, S.M.I. (2024). Privacy-Preserving Intrusion Detection in Software-Defined VANET Using Federated Learning with BERT.

Ankele, R. and Haddadi, H. (2024). The Boomerang Protocol: A Decentralised Privacy. Preserving Verifiable Incentive Protocol.

Asad, M., Moustafa, A. and Yu, C. (2020). A Critical Evaluation of Privacy and Security Threats in Federated Learning. Sensors.

Bakule, L. (2008). Decentralized Control: An Overview. *Annual Reviews in Control*, 32: 87-98.

Bellet, A. (2017). Personalized and Private Peer-to-Peer Machine Learning.

Berke, A., South, T., Mahari, R., Larson, K. and Pentland, A. (2023). zkTax: A Pragmatic Way to Support Zero-Knowledge Tax Disclosures.

Bonawitz, K.A., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H.B., Patel, S., Ramage, D., Segal, A. and Seth, K. (2016). Practical Secure Aggregation for Federated Learning on User-Held Data. *CoRR*, abs/1611.04482.

Byrd, D. (2020). Differentially Private Secure Multi-Party Computation for Federated Learning in Financial Applications. arXiv.org

Choi, H. (2023). Blind-Touch: Homomorphic Encryption-Based Distributed Neural Network Inference for Privacy-Preserving Fingerprint Authentication. arXiv.org

Chu, T., Yang, M., Laoutaris, N. and Markopoulou, A. (2023). PriPrune: Quantifying and Preserving Privacy in Pruned Federated Learning. arXiv preprint arXiv:2310.19958.

Davaei, M. and Gunkel, M. (2023). The Role of Intelligences in Teams: A Systematic Literature Review. *Review of Managerial Science*.

Ding, J., Tramel, E.W., Sahu, A.K., Wu, S., Avestimehr, S. and Zhang, T. (2022). Federated Learning Challenges and Opportunities: An Outlook. CoRR, abs/2202.00807.

Ding, S., Zhang, L., Pan, M. and Yuan, X. (2023). Patrol: Privacy-Oriented Pruning for Collaborative Inference Against Model Inversion Attacks.

Elbir, A.M. (2021). A Hybrid Architecture for Federated and Centralized Learning.

Elgabli, A., Park, J., Bedi, A.S., Bennis, M. and Aggarwal, V. (2020). Communication Efficient Framework for Decentralized Machine Learning. Online.

Feng, C., Celdran, A.H., Baltensperger, J., Beltran, E.T.M., Bovet, G. and Stiller, B. (2023). Sentinel: An Aggregation Function to Secure Decentralized Federated Learning.

Gillis, A.S. (2022). Homomorphic Encryption. Accessed: Feb. 10, 2024.

Gosselin, R., Vieu, L., Loukil, F. and Benoit, A. (2022). Privacy and Security in Federated Learning: A Survey. *Applied Sciences*, Vol. 12, Page 9901, 12:9901.

Gvili, Y., Ha, J., Scheffler, S., Varia, M., Yang, Z. and Zhang, X. (2021). Turboikos: Improved Non-Interactive Zero Knowledge and Post-Quantum Signatures. *Cryptology ePrint Archive*, Paper 2021/478.

Gunther, M. (2023). HElium: A Language and Compiler for Fully Homomorphic Encryption with Support for Proxy Re-Encryption. [Accessed on Jan. 21, 2024].

Hassani, M. (2022). PHY-Fed: An Information-Theoretic Secure Aggregation in Federated Learning in Wireless Communications. arXiv.org

Huang, H., Zhang, L., Sun, C., Fang, R., Yuan, X. and Wu, D. (2023). FedTiny: Pruned Federated Learning Towards Specialized Tiny Models. openreview.net. Accessed: Apr. 01, 2023.

IEEE (2021). What is Homomorphic Encryption?. IEEE Digital Privacy.

Jony, A.I. and Serradell-López, E. (2019). Effective Virtual Teamwork Development in Higher Education: A Systematic Literature Review. *Edulearn19 Proceedings*, 873-882.

Jony, A.I. and Arnob, A.K.B. (2024). A Long Short-Term Memory Based Approach for Detecting Cyber Attacks in IoT Using CIC-IoT2023 Dataset. *Journal of Edge Computing*, 3(1), 28-42.

Jony, A.I. and Hamim, S.A. (2023). Navigating the Cyber Threat Landscape: A Comprehensive Analysis of Attacks and Security in the Digital Age. *Journal of Information Technology and Cyber Security*, 1(2), 53-67.

Jony, A.I., Rithin, A.T. and Edrish, S.I. (2024). A Comparative Study and Analysis of Text Summarization Methods. *Malaysian Journal of Science and Advanced Technology*, 4(2), 118-129.

Khan, M., Glavin, F.G. and Nickles, M. (2023). Federated Learning as a Privacy Solution—An Overview. Vol. 217, pp. 316-325. Elsevier B.V.

Kishiyama, B. (2023). A Review on Searchable Encryption Functionality and the Evaluation of Homomorphic Encryption. arXiv.org.

Konecny, J. (2016). Federated Optimization: Distributed Machine Learning for On-Device Intelligence.

Kraus, S. *et al.* (2022). Literature Reviews as Independent Studies: Guidelines for Academic Practice. *Review of Managerial Science*. Accessed on Jan. 04, 2024.

Li, L., Yang, F., Tse, M. and Lin, K.-J. (2020). A Review of Applications in Federated Learning. *Computers & Industrial Engineering*.

Li, Q. *et al.* (2023). A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection. *IEEE Transactions on Knowledge and Data Engineering*.

Lim, W.Y.B. (2019). Federated Learning in Mobile Edge Networks: A Comprehensive Survey. Accessed: Jan 21, 2024.

Lin, R. (2022). Federated Pruning: Improving Neural Network Efficiency with Federated Learning. arXiv.org.

Lisun-Ul-Islam, M., Rahat, M.R.H., Esha, S., Faiyaz, A. and Jony, A.I. (2023). Hourly Air Quality Prediction in Dhaka City Using Time Series Forecasting Techniques: Deep Learning Perspectives. *Tuijin Jishu/Journal of Propulsion Technology*, 44(5), 568-579.

Liu, T., Xie, X. and Zhang, Y. (2021). zkCNN: Zero Knowledge Proofs for Convolutional Neural Network Predictions and Accuracy. In Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, CCS'21, pp. 2968-2985, New York, NY, USA. Association for Computing Machinery.

Liu, X. (2023). Adaptive Model Pruning and Personalization for Federated Learning Over Wireless Networks. arXiv.org.

Liu, Y., Kang, Y., Zou, T., Pu, Y., He, Y., Ye, X., Ouyang, Y., Zhang, Y.-Q. and Yang, Q. (2022). Vertical Federated Learning: Concepts, Advances and Challenges. *IEEE Transactions on Knowledge and Data Engineering*, pp. 1-20.

Liu, Y., Yuan, X., Xiong, Z., Kang, J., Wang, X. and Niyato, D. (2020). Federated Learning for 6g Communications: Challenges, Methods, and Future Directions. *China Communications*.

Long, Q. (2023). FedDIP: Federated Learning with Extreme Dynamic Pruning and Incremental Regularization. arXiv.org.

Luo, F., Al-Kuwari, S., Wang, H. and Yan, X. (2023). FSSA: Efficient 3-Round Secure Aggregation for Privacy-Preserving Federated Learning.

Ma, C., Li, Y., Liang, M. and Li, A. (2023). Federated Topic Model and Model Pruning Based on Variational Autoencoder. *Lecture Notes in Electrical Engineering*.

Maltitz, M.V. and Carle, G. (2018). A Performance and Resource Consumption Assessment of Secret Sharing Based Secure Multiparty Computation. *Lecture Notes in Computer Science*.

Mori, J., Teranishi, I. and Furukawa, R. (2022). Continual Horizontal Federated Learning for Heterogeneous Data. in 2022 International Joint Conference on Neural Networks (IJCNN).

Nguyen, T. and Thai, M.T. (2022). Preserving Privacy and Security in Federated Learning.

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D. and Others (2021). The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ*, 372.

Palaiokrassas, G., Scherrers, S., Ofeidis, I. and Tassiulas, L. (2023). Leveraging Machine Learning for Multichain Defi Fraud Detection. arXiv preprint arXiv:2306.07972.

Pasquini, D., Raynal, M. and Troncoso, C. (2022). On the (in)Security of Peer-to-Peer Decentralized Machine Learning.

Pawn, Rookie, and Cheng, Z. (2024). Zero-Knowledge Proof in NuLink.

Prayitno, P. *et al.* (2021). A Systematic Review of Federated Learning in the Healthcare Area: From the Perspective of Data Properties and Applications. *Applied Sciences*, 11(23), 11191.

Rao, J., Gao,. S., Li, M. and Huang, Q. (2021). A Privacy-Preserving Framework for Location Recommendation Using Decentralized Collaborative Machine Learning.

Razavi-Far, R., Wang, B., Taylor, M.E. and Yang, Q. (2022). An Introduction to Federated and Transfer Learning. *In Adaptation, Learning, and Optimization*.

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H.R., Albarqouni, S., Bakas, S., Galtier, M.N., Landman, B.A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R.M., Trask, A., Xu, D., Baust, M. and Cardoso, M.J. (2020). The Future of Digital Health with Federated Learning. *npj Digit. Med.*, 3, 119. doi: https://doi.org/10.1038/s41746-020-00323-1

Rizinski, M., Peshov, H., Mishev, K., Chitkushev, L., Vodenska, I. and Trajanov, D. (2022). Ethically Responsible Machine Learning in Fintech. *IEEE Access*, 10, 97531-97554.

So, J., Ali, R.E., Guler, B., Jiao, J. and Avestimehr, S. (2021). Securing Secure Aggregation: Mitigating Multi-Round Privacy Leakage in Federated Learning. CoRR, abs/2106.03328.

Sohan, M.F. and Basalamah, A. (2023). A Systematic Review on Federated Learning in Medical Image Analysis. *IEEE Access*, 11, 28628-28644.

Song, J., Wang, W., Gadekallu, T.R., Cao, J. and Liu, Y. (2023). EPPDA: An Efficient Privacy-Preserving Data Aggregation Federated Learning Scheme. *IEEE Transactions on Network Science and Engineering*, 10(5), 3047-3057.

Stojkovic, B. (2022). Applied Federated Learning: Architectural Design for Robust and Efficient Learning in Privacy Aware Settings. arXiv.org

Tanvir, K., Jony, A.I., Haq, M.K., Nazera, F., Dass, M. and Raju, V. (2023). Clinical Insights Through Xception: A Multiclass Classification of Ocular Pathologies. *Tuijin Jishu/Journal of Propulsion Technology*, 44(04), 2023.

Truhn, D., Arasteh, S.T., Saldanha, O.L., Muller-Franzes, G., Khader, F., Quirke, P., West, N.P., Gray, R., Hutchins, G.G., James, J.A., Loughrey, M.B., Salto-Tellez, M., Brenner, H., Brobeil, A., Yuan, T., Chang-Claude, J., Hoffmeister, M., Foersch, S., Han, T., Keil, S., Schulze-Hagen, M., Isfort, P., Bruners, P., Kaissis, G., Kuhl, C., Nebelung, S. and Kather, J.N. (2024). Encrypted Federated Learning for Secure Decentralized Collaboration in Cancer Image Analysis. *Medical Image Analysis*, 92.

Van Kempen, E., Li, Q., Marson, G.A. and Soriente, C. (2023). LISA: LIghtweight Single-Server Secure Aggregation with a Public Source of Randomness. arXiv preprint arXiv:2308.02208.

Vergne, J.-P. (2020). Decentralized vs. Distributed Organization: Blockchain, Machine Learning and the Future of the Digital Platform. *Organization Theory*.

Wahab, O.A., Mourad, A., Otrok, H. and Taleb, T. (2021). Federated Machine Learning: Survey, Multi-Level Classification, Desirable Criteria and Future Directions in Communication and Networking Systems. *IEEE Journals & Magazine*, IEEE Xplore.

Wang, X., Han, Y., Leung, V.C.M., Niyato, D., Yan, X. and Xu, C. (2020). Convergence of Edge Computing and Deep Learning: A Comprehensive Survey. *IEEE Communications Surveys and Tutorials*.

Wang, Z. (2023). QMGeo: Differentially Private Federated Learning via Stochastic Quantization with Mixed Truncated Geometric Distribution.

Wang, Z., Chaliasos, S., Qin, K., Zhou, L., Gao, L., Berrang, P., Livshits, B. and Gervais, A. (2023a). On How Zero-Knowledge Proof Blockchain Mixers Improve and Worsen User Privacy. In Proceedings of the ACM Web Conference 2023, WWW'23, pp. 2022-2032, New York, NY, USA. Association for Computing Machinery.

Wang, Z., Dong, N., Sun, J. and Knottenbelt, W. (2023b). zkFL: Zero-Knowledge Proof-Based Gradient Aggregation for Federated Learning.

Xing, Z., Zhang, Z., Li, M., Liu, J., Zhu, L., Russello, G. and Asghar, M.R. (2023). Zero-Knowledge Proof-Based Practical Federated Learning on Blockchain.

Yan, Y. (2023). HE-DKSAP: Privacy-Preserving Stealth Address Protocol via Additively Homomorphic Encryption. arXiv.org

Yang, K., Shi, Y., Zhou, Y., Yang, Z., Fu, L. and Chen, W. (2020a). Federated Machine Learning for Intelligent IoT via Reconfigurable Intelligent Surface. *IEEE Network*, 34(5), 16-22.

Yang, Q., Liu, Y., Yang, C., Ke, Y., Chen, T. and Han, Y. (2020b). Horizontal Federated Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*.

Yao, D., Liu, T., Cao, Q. and Jin, H. (2024). FedRKG: A Privacy-Preserving Federated Recommendation Framework via Knowledge Graph Enhancement. *Lecture Notes in Computer Science*.

Yoaquim (2022). What is Zero Knowledge Proof Used for?. *All ZKP Use Cases. Medium.*

Yu, S. and Cui, L. (2022). Secure Multi-Party Computation in Federated Learning.

Yu, S., Nguyen, P., Anwar, A. and Jannesari, A. (2021). Heterogeneous Federated Learning Using Dynamic Model Pruning and Adaptive Gradient.

Zhang, S., Liao, Y. and Zhou, P. (2023). AHSecAgg and TSKG: Lightweight Secure Aggregation for Federated Learning Without Compromise.

Zhang, T., Gao, L., He, C., Zhang, M., Krishnamachari, B. and Avestimehr, S. (2021). Federated Learning for Internet of Things: Applications, Challenges, and Opportunities. *CoRR*, abs/2111.07494.

Zhang, X. (2020). Hybrid Federated Learning: Algorithms and Implementation.

Zhang, Y. (2023). A Survey of Trustworthy Federated Learning with Perspectives on Security, Robustness, and Privacy.

Zhang, Y., Bai, G., Li, X., Curtis, C., Chen, C. and Ko, R.K.L. (2020). PrivColl: Practical Privacy-Preserving Collaborative Machine Learning.

Zhang, Y. and Zhu, H. (2020). Additively Homomorphical Encryption based Deep Neural Network for Asymmetrically Collaborative Machine Learning. arXiv preprint arXiv:2007.06849.

Zhao, Y. (2020). A Comprehensive Survey of 6G Wireless Communications. Accessed: Jan 21, 2024.

Zhou, J., Chen, S., Wu, Y., Li, H., Zhang, B., Zhou, L., Hu, Y., Xiang, Z., Li, Z., Chen, N., Han, W., Xu, C., Wang, D. and Gao, X. (2024). PPML-Omics: A Privacy-Preserving Federated Machine Learning Method Protects Patients' Privacy in Omics Data.

Zhu, H. (2021). Federated Learning on Non-IID Data: A Survey.

Zhu, Y., Wu, Y., Luo, Z., Ooi, B.C. and Xiao, X. (2023). Secure and Verifiable Data Collabo-Ration with Low-Cost Zero-Knowledge Proofs.

Cite this article as: Akinul Islam Jony and Mubashir Mohsin (2024). Data Privacy Preservation with Federated Learning: A Systematic Review. *International Journal of Data Science and Big Data Analytics*, 4(1), 1-16. doi: 10.51483/IJDSBDA.4.1.2024.1-16.