

Research Article

Jansi Rani Amalraj* and Robert Lourdusamy

Security and privacy issues in federated healthcare – An overview

<https://doi.org/10.1515/comp-2022-0230>
received August 5, 2020; accepted January 20, 2022

Abstract: Securing medical records is a significant task in Healthcare communication. The major setback during the transfer of medical data in the electronic medium is the inherent difficulty in preserving data confidentiality and patients' privacy. The innovation in technology and improvisation in the medical field has given numerous advancements in transferring the medical data with foolproof security. In today's healthcare industry, federated network operation is gaining significance to deal with distributed network resources due to the efficient handling of privacy issues. The design of a federated security system for healthcare services is one of the intense research topics. This article highlights the importance of federated learning in healthcare. Also, the article discusses the privacy and security issues in communicating the e-health data.

Keywords: federated learning, healthcare, privacy, security, authentication

1 Introduction

Federated learning (FL) is a shared learning model that works without exchanging users' original data. The research on FL is emerging due to the privacy issues pointed out by the data owners across the world. FL allows users to train the shared model through cooperation and allows personal information to be intact on their devices. FL system consists of two components, namely, the data owners and the model

owner. Data owners denote the participants and model owner is the FL server used.

Let $X = \{1, \dots, N\}$ denote the set of N data owners. Each owner has a particular dataset $D_{i \in X}$. Each participant ' i ' uses its dataset D_i to train a local model w_i and transfer only the local model parameters to the FL server. The global model w_G is generated by aggregating all local models. This type of working is completely different from the centralized training model. In a centralized approach, each source is aggregated in the initial stage and then "centralized model training" is initiated.

Traditional machine learning involves a centralized approach which requires the training data to be aggregated on a single machine or in a data center. Such data collected in a specific environment need to be communicated back to the central server. The advancement in Internet of Things (IoT) and mobile network communication has given rise to exponential data storage at device level. Hence, conventional methods are not suitable for the current generation of networks with high volume of data. FL is a decentralized learning approach where training is performed over a federation of distributed learners. The aspect of data security and privacy is the foremost advantage of using this model. The storage of critical data at centralized repository is avoided and valuable real time prediction is made possible without any time lag in federated model. Also, the requirement of minimal hardware makes this system highly compatible. Compared to cloud computing models, FL models offer privacy-preserving mechanism to control those decentralized computing resources inside the mobile devices to train machine learning models. But the reliability of end devices and communication bandwidth needs to be kept intact during FL process.

FL training process has three steps: (1) task initialization, (2) local model training with update, and (3) global model aggregation with an update. The training task is informed by the server. After deciding on the hyperparameters, the server will send the initialized global model to chosen participants. In the second step, each participant updates local model parameters by using the global model. This updation is communicated to the

* **Corresponding author: Jansi Rani Amalraj**, Department of Computer Science, Government Arts College, Coimbatore-641 018, Tamil Nadu, India; Department of Information Technology, Nirmala College for Women (Autonomous), Coimbatore-641018, Tamil Nadu, India, e-mail: jansiramalraj@gmail.com
Robert Lourdusamy: Department of Computer Science, Government Arts College, Coimbatore-641 018, Tamil Nadu, India, e-mail: robertatgac@gmail.com

server. During the last step, the server does the grouping process of local models, and updated global model parameters are sent to data owners. The FL process is shown in Figure 1.

Traditional data processing models have become ineffective to address the security and privacy issues. On May 25, 2018, European Union proposed General Data Protection Regulation (GDPR) for regulating data breach and protecting the privacy concerns [2]. Recently, untoward incident happened in Facebook community, where large amount of data were misused. The proposal draft from GDPR emerged as a control measure to protect users' personal privacy and provide data security. Due to the stringent laws, the traditional data processing methods are affected in certain stages during the processing and they encounter data fragmentation and isolation issues. FL offers better alternative in these situations.

Privacy is a significant factor in FL. Secure multiparty computation (SMC) is one of the privacy method adopted in FL [3]. This security model provides no knowledge to the transacting parties. The parties know only about the input and output. Due to the inherent complication in offering such a knowledge base, partial knowledge sharing is initiated in a certain case. Differential privacy model, along with k-anonymity and diversification features, is an alternate privacy model used. The sensitive information is made unclear to third parties by adding noise [4]. Parameter exchange with the help of the encryption mechanism is carried out under a new model called Homomorphic encryption [5].

The rest of this article is organized as follows. Section II introduces the types and background of FL. Section III reviews the privacy issues of FL model. Section IV

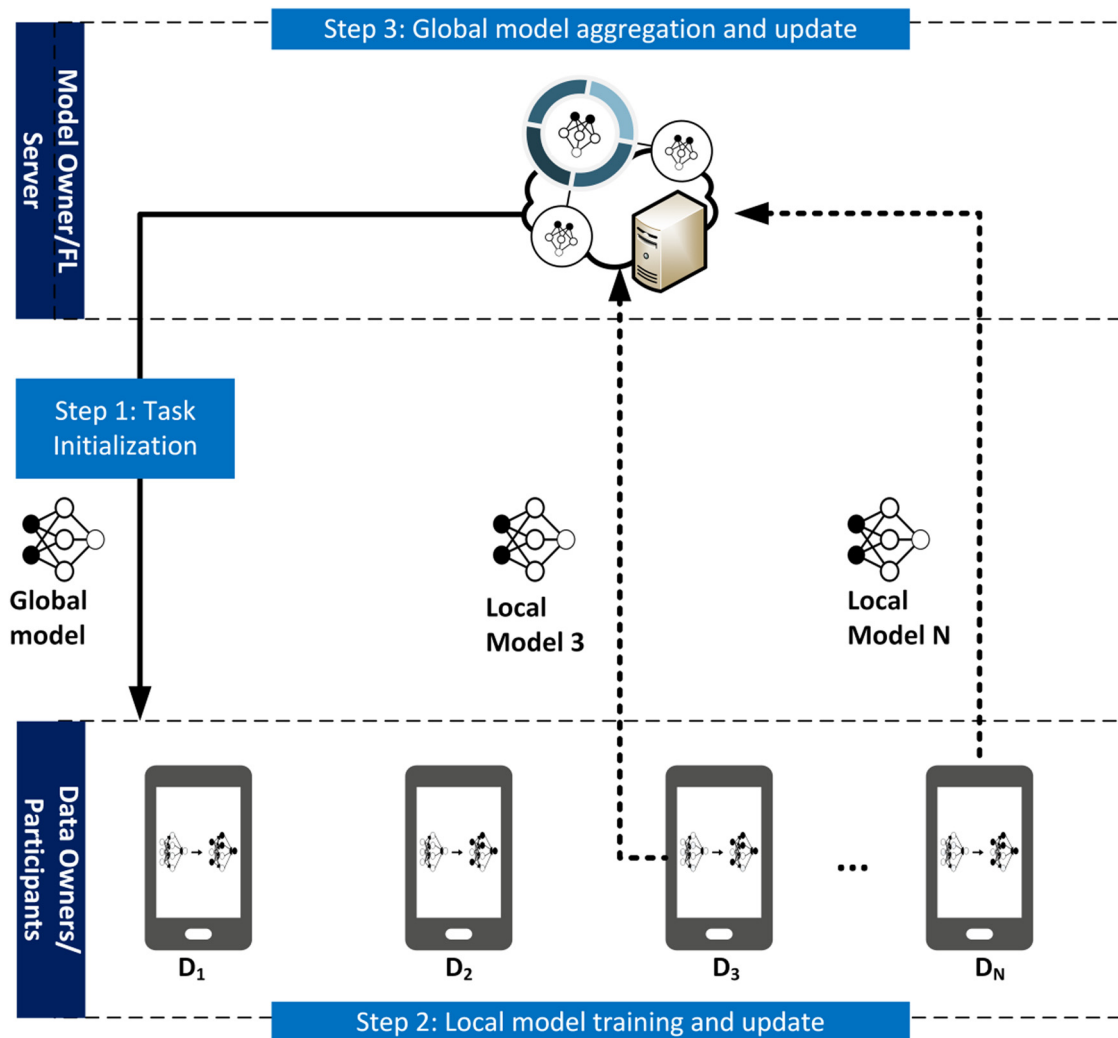


Figure 1: FL process [1].

discusses the issues in federated healthcare. Section V discusses the challenges faced by the FL community.

2 Types of FL

FL frameworks can be classified as horizontal and vertical FL. In the horizontal model, datasets share the same feature space but different space in samples. In the vertical model, two datasets share the same sample ID space but differ in feature space. There is a hybrid model termed as federated transfer learning. Here it is used in situations where two datasets differ not only in samples but also in feature space. The architecture of horizontal FL is shown in Figure 2.

In horizontal FL, k participants with the same data structure collaboratively learn a model with the help of a parameter. The vertical FL relies on trusted third-party collaborator. The model is elucidated in Figure 3.

The mechanism of FL makes it possible for institutions and enterprises to share a united model without data exchange [6].

The broad areas of the distributed system, machine learning, and privacy have given rise to a new model called FL. In FL system, the computation is carried out on the parties and the manager. The communication is initiated between the parties and the manager. Computation is done for the model training, and the communication is performed for exchanging the model parameters. Four

different aspects of FL are data partition, model, privacy level, and communication architecture [7]. The aspects of FL system is shown in Figure 4.

3 Security and privacy issues in the federated system

Privacy concerns of healthcare data is a significant problem in the healthcare industry. Medical experimentation needs foolproof security as they are closely related to national security. Protection of patient information is regulated by the official name, Health Insurance Portability and Accountability Act (HIPAA). Consensus solution and pluralistic solution are recommended to overcome the statistical challenge of FL. Data poisoning and model poisoning are security attacks pertaining to the FL model. In data poisoning model, the attacker can simply train its model on label-flipping or backdoor inputs. Under model poisoning, in inference attack, the learning process is initiated to check if a particular individual participated in the training. These attacks are based on “Honest-but-Curious and Adversary” scenario. Defense (Honest-but-Curious scenario) attempts to capture extra information in numerous ways. Different authors have proposed varying solutions to combat such issues (Table 1).

FL can be applied effectively in the medical domain to access different forms of medical data without

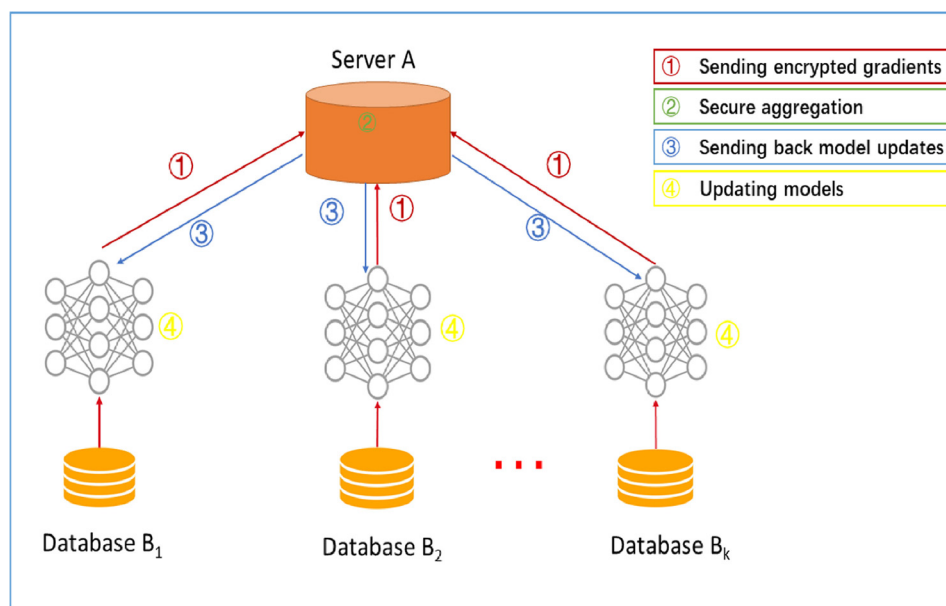


Figure 2: Horizontal FL.

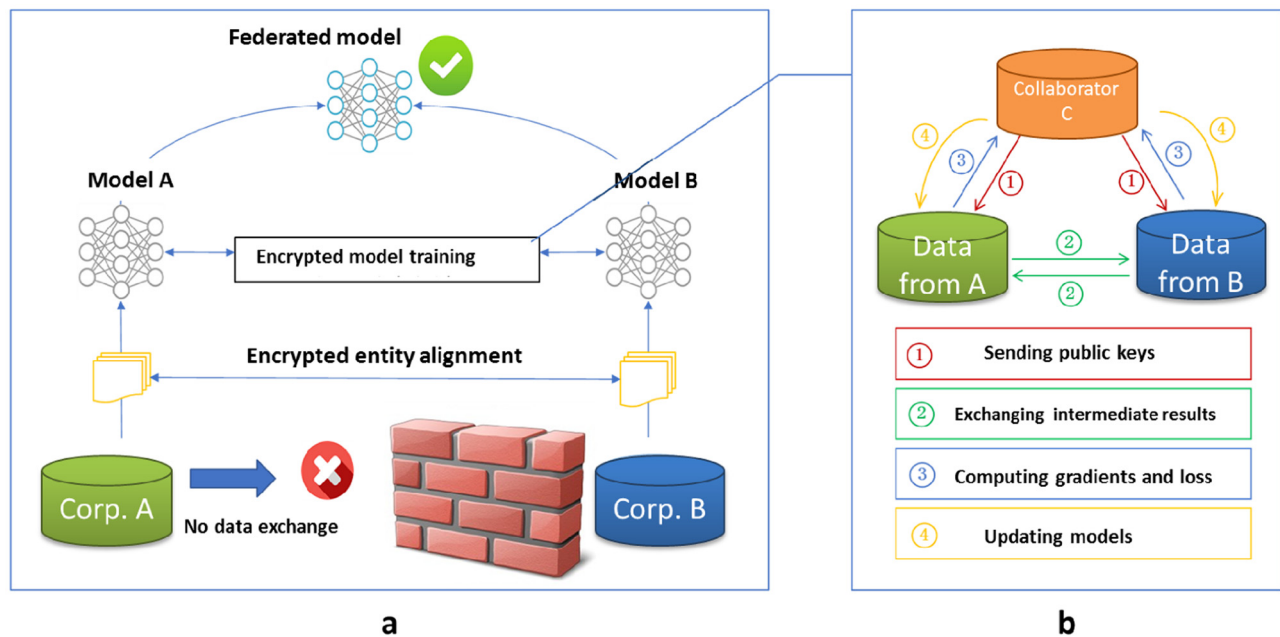


Figure 3: (a and b) Vertical FL.

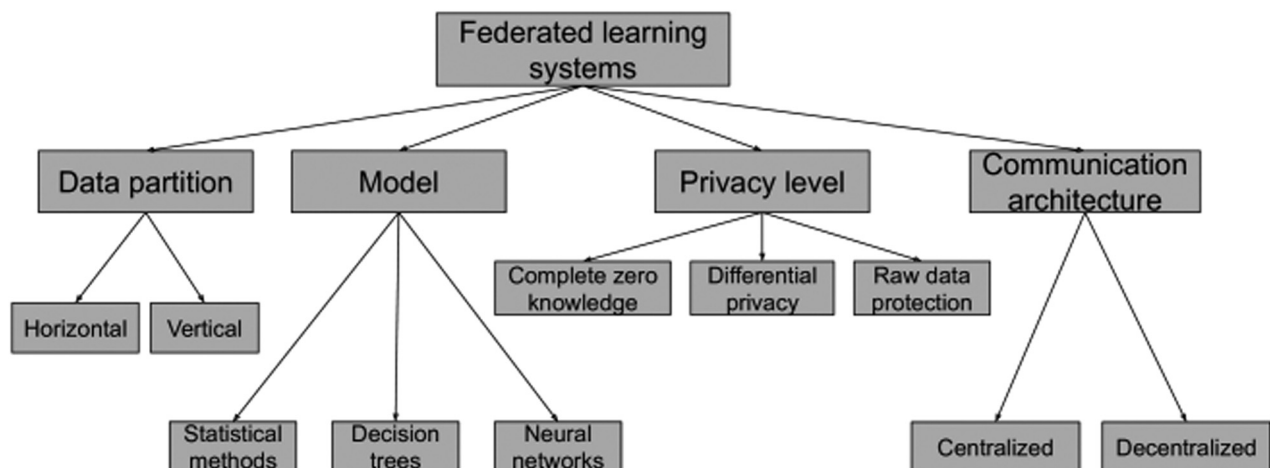


Figure 4: Aspects of the FL system [7].

compromising the security and privacy issues. In this novel system, training happens for a shared global model with a centralized server. But the most sensitive data are kept in the local region in the custody of data owners. Huge medical data need to be processed with utmost care [17].

The recent trend in healthcare allows the implementation of federated architecture that addresses the issues of privacy. Security requirements for health information system include confidentiality, integrity, availability, non-repudiation, and accountability. Qualitative assessment of security measures in the health system is performed. It is proved that

federated health information systems can offer security properties by adopting proper mechanisms to protect the exchanged data [18].

Aggregation of Electronic Health Record (EHR) systems using conventional methods end up in compromising patient privacy and data security. Novel access control policies are required for federated EHR systems. Cryptographic techniques are applied for EHR systems to provide safe transmission using the public key infrastructure. But such methods have no control over the access control mechanisms. Discretionary Access Control (DAC), Mandatory Access Control (MAC), and Role-Based Access

Table 1: Privacy methods for FL

Methodology	Author(s)	Description	Performance
SMC	Bonawitz et al. [8]	Encryption model to make updates of a single device undetectable by the server and the sum is revealed only after receiving a sufficient number of updates.	By complexity analysis and implementation output, the runtime and communication overhead remain low even on large datasets and client pools. For 16-bit input values, the protocol offers $1.73 \times$ communication expansion for 2^{10} users and 2^{20} -dimensional vectors.
Homomorphic encryption	Liu et al. [9]	Federated transfer learning model is used with adapting additively homomorphic encryption to multi-party computation.	Performance of the proposed transfer learning approach compared to self-learning approach. Weighted F1 score of transfer learning with Taylor loss and with logistic loss is observed. Also, comparison is made with self-learning with logistic regression (LR), with support vector machines, and with stacked auto encoders. Highest F1 score of 0.718 ± 0.033 for transfer learning approach is reported and hence it is better than conventional models.
	Chai et al. [10]	Distributed matrix factorization framework is enabled with homomorphic encryption.	The model is secure against an honest-but-curious server. Testing the computation time of FedMF shows better performance.
Three-party end-to-end solution	Hardy et al. [11]	Privacy-preserving entity resolution and federated LR is used. The method shows the learning process of a linear classifier in a privacy-preserving federated fashion when data are vertically partitioned.	Performs exactly on par with LR run on perfectly linked datasets
Differential privacy	Shokri et al. [12]	Perturbed values of the selected gradients under a consistent differentially private framework.	Selective Stochastic Gradient Descent (Selective SGD or SSGD) protocol achieves comparable accuracy compared to conventional SGD.
	McMahan et al. [13]	Client-level privacy protection where the model does not reveal whether a client participated during training.	Achieving differential privacy comes at the cost of increased computation
	Geyer et al. [14]	Dynamically adapting the differentially private (DP)-preserving mechanism is used.	For 100 and 1,000 clients, model accuracy does not converge and stays significantly below the non-DP performance. However, 78% and 92% accuracy for $K \in \{100, 1,000\}$ are still substantially better.
	Chen et al. [15]	DP autoencoder-based generative model (DP-AuGM) is used.	DP-AuGM in FL compared to original setting without DP-AuGM. Accuracy drops only within 5% for all datasets in this model.
	Cheng et al. [16]	Lossless privacy-preserving tree-boosting. The framework is known as the SecureBoost in a FL.	Performance of proposed completely secureBoost model compared to SecureBoost Model. Accuracy of above 93% of both the models are reported.

Control (RBAC) are the prominent access control mechanisms adopted by healthcare systems [19].

DAC is a way of containing access to objects based on the identity of subjects to which they belong. Though this method offers flexibility, it may create complexity due to EHR inherent properties applied to patient data. In MAC,

decisions are made by a central authority, and not by the individual owner of an object. In RBAC, users are granted membership into roles based on their proficiency and responsibilities in the group. The combined model access to a particular EHR utilizes all the three policies. Combined access model is shown in Figure 5.

4 Secure healthcare data sharing in the federated system

Federated Health Information Systems (FHIS) need to maintain security and privacy in preserving critical medical data. To overcome the security threat offered by internal and external attackers, pseudonymization and encryption are combined to form a hybrid model [20]. Using this model, all the medical data are kept without applying encryption. This allows for particular access to data. But the metadata is encrypted with advanced cryptographic techniques. A novel case study of a federation of health insurance data warehouses (HEWAF) is explained [21]. A federated data warehouse model for evidence-based medicine is required to address privacy issues. Depersonalization and pseudonymization are deployed to protect patient privacy and confidentiality. The restriction in re-identification of patient data during the processing is maintained using the properties of depersonalization. Social security number of a patient is made as a secret entity by pseudonymization. Object-oriented reference information model (RIM) introduces a single federation warehouse for the medical community and XML-based CDA messaging concept.

A secure EHR system to protect patient privacy is proposed to improve the quality of the healthcare system using cryptographic methods [22]. Provision is given to patients to efficiently store and retrieve their protected health information (PHI) in a secure and private manner even with a public server, such that only the patient can learn the content of his PHI. Identity-based cryptography, Searchable symmetric encryption, and Searchable public-key encryption are reported as efficient methods for healthcare system for patient privacy system.

Medical Cyber-Physical System (MCPS) provides soaring care in healthcare environments by means of constant monitoring and treatment. MCPS forms a network of medical devices that stores sensitive medical and personal data of patients. Such data require high security and privacy. FL method is used in MCPS to minimize the computation and communication cost and maximize the protection of the data. The existing Intrusion Detection System (IDS) puts forward more false-positive rates and needs manual specification and modification of data. So a new IDS is designed for use in MCPS to overcome the disadvantages. It also explores the feasibility of robust attack detection in FL methods [23]. FL-based IDS is depicted diagrammatically in Figure 6.

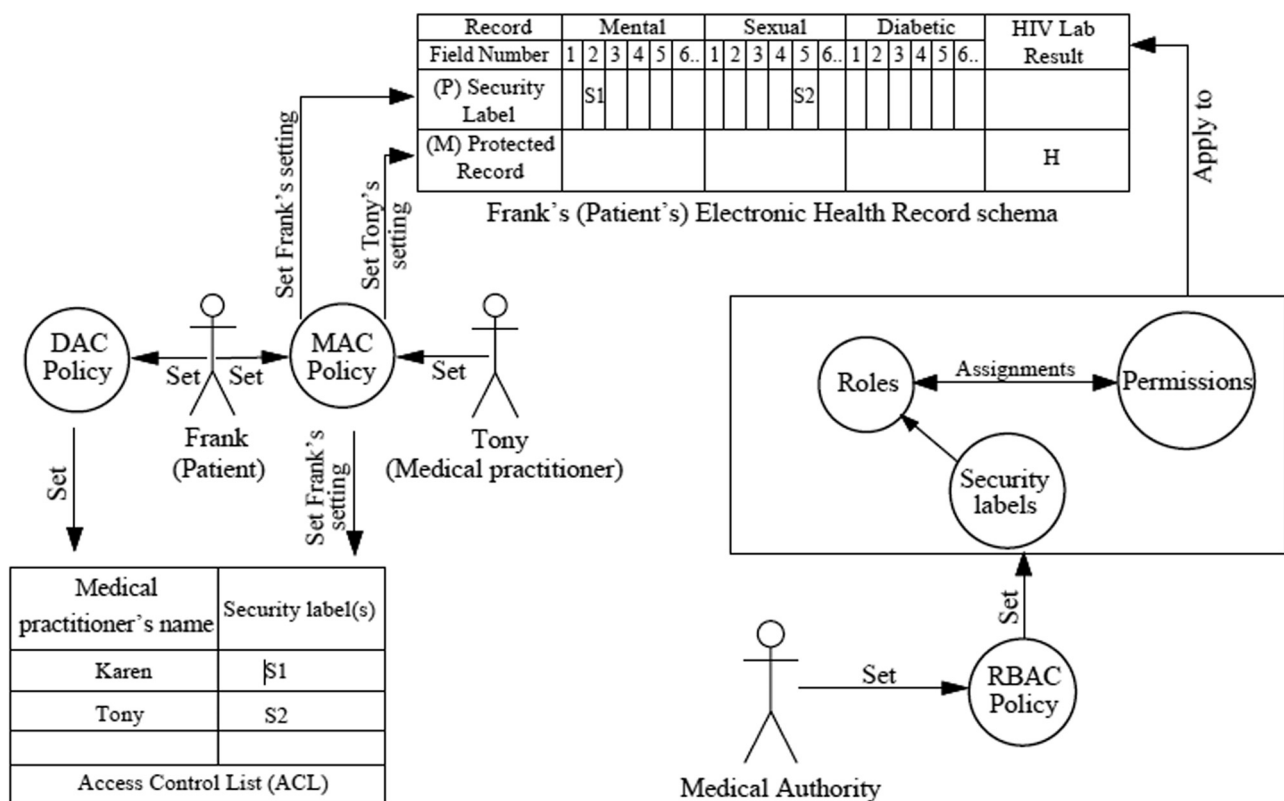


Figure 5: Combined access control protocol [19].

The FL builds a global model by averaging the weights “ w ” to each device over the number of communication round “ t ”. The figure shows the FL method that runs the detection module by utilizing the computational resources of the mobile devices. The server of the federated model becomes the centralized authority that registers the mobile users, calculates the federated model, and stores the model. The design procedure of the federated model in healthcare systems has three components, namely, clustering of patients, training and updating the model, and attack detection.

The security and performance of the model lead to the reduction in false-positive rates and increase in the accuracy of the model. The communication cost and energy consumption have also significantly reduced in the federated model. The detection accuracy of the federated model is more accurate compared to the other non-FL models. The privacy of the healthcare system is increased because only an updated vector is communicated across the patients rather than the personally identifiable information.

Security and privacy in healthcare systems have become a more promising challenge in recent years. FL

includes direct training of statistical models on remote devices, thereby preserving privacy. The challenges that make FL method distinct from other distributed learning are expensive communication, systems heterogeneity, statistical heterogeneity, and privacy concerns [24].

5 Challenges in FL

5.1 Communication efficiency

Communication is a significant tailback in FL as the method uses the outsized number of devices, and hence the commutation becomes slower. To overcome such a disadvantage, it is necessary to reduce the communication rounds and reduce the size of each message transmitted during the round. To provide efficient communication in the FL methods, several general directions are provided which are classified into local updating, compression schemes, and decentralized training.

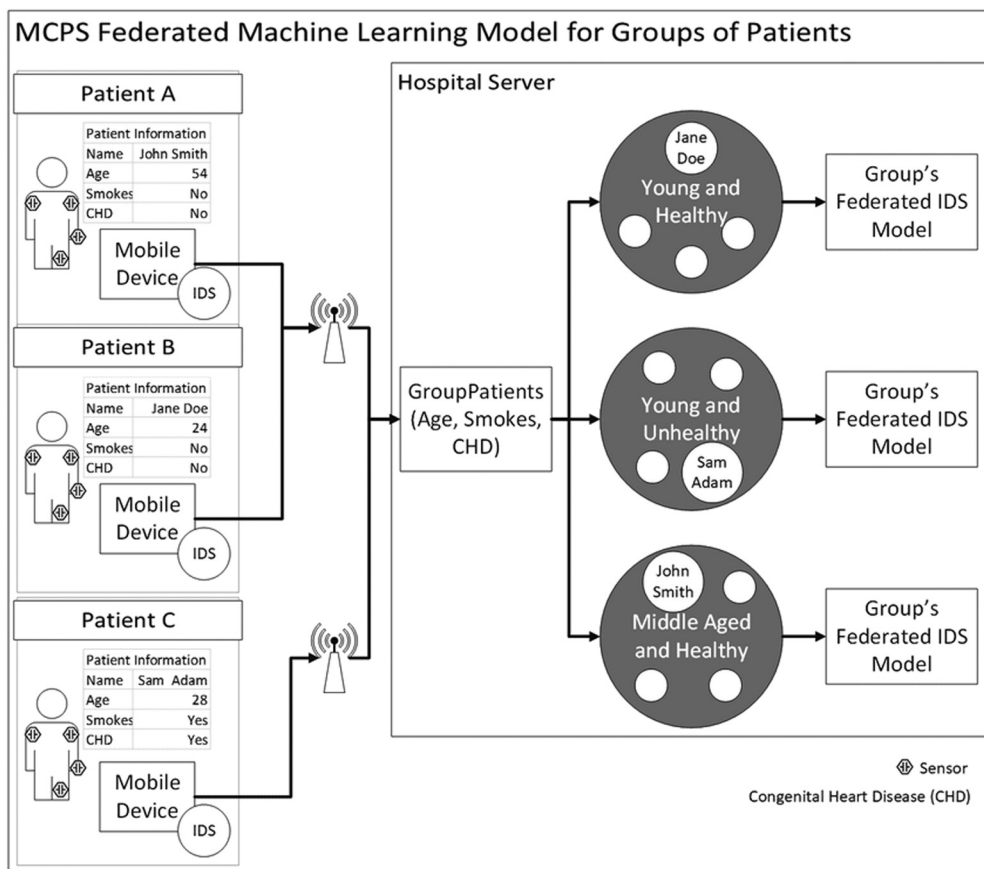


Figure 6: FL-based IDS [23].

5.2 Systems heterogeneity

The FL suffers from communication, computation, and storage capabilities as there is a difference in power, network connectivity, and hardware. To overcome such disadvantages, the federated system should inculcate lesser participation, tolerate heterogeneous hardware, and it should be strong enough to handle the dropped devices in the network. Several methods to handle the heterogeneity of the system are classified into asynchronous communication, active device sampling, and fault tolerance.

5.3 Statistical heterogeneity

To handle the statistical heterogeneity in the federated network, multi-tasking and meta-learning approaches can be implemented. The related works in such approaches are classified into modeling heterogeneous data and convergence guarantees for non-independent and intrusion detection data.

5.4 Privacy

The FL suffers from privacy concerns as it reveals the sensitive information to third party throughout the training process. Differential privacy and SMC can be effectively implemented to maintain the privacy of the learning method [25]. The privacy in FL can be classified into local privacy and global privacy. Local privacy makes the data private to both the server and the third party. Global privacy makes the data generated at each round private to the untrusted third party rather than the central server. SMC protocol can be used to preserve the privacy of individual model updates. The central server receives the aggregate results during each computation round without any local updates, thereby maintaining privacy. SMC is a lossless method that provides original accuracy of the data with high privacy guarantee [26].

FL methods can inculcate differential privacy to maintain global privacy [14]. In differential privacy, the trusted curator combines the parameters of multiple clients in a decentralized manner. The resulting model is then distributed to the clients in a joint representative model without sharing the data explicitly. On the other hand, the protocol is vulnerable to attacks from any party in the federated network. An algorithm has been proposed to maintain the client-side differential privacy in FL models. Random subsampling and distorting methods are used in the algorithm

to hide the client's contribution, thereby hiding from the decentralized learning methods.

Without reliable identification methods, all trust sharing and data encryption become compromised. Identification techniques are used along with trust sharing and access devices to establish the trust levels of users. The methods utilized for this purpose are biometric, digital, and physical methodologies. Fingerprint scanners, iris scanners, and signature recognition are used extensively as biometric models in healthcare data communication. Radio frequency identification (RFID), smart cards, and USB electronic token are prominent digital techniques used by various researchers. Post-quantum cryptography includes hash function-based cryptography, lattice-based cryptography, code-based cryptography, multivariate cryptography, and other post-quantum cryptography algorithms [27].

The medical community must use a secure authentication system to access patient records. Biometric-based access models are used effectively in many applications. The biometric-based system offers more advantages than token-based and knowledge-based systems. Multimodal biometric systems provide foolproof authentication in many scenarios. Voice and signature verification models are used by biomedical authentication system for healthcare applications. The authentication system is in line with HIPAA regulations. The protection and privacy of medical records are maintained carefully without security breach [28]. Deep learning model construction for a semi-supervised classification with feature learning was discussed by authors in ref. [29]. The authors proposed a novel trust-aware routing framework for IoT [30]. Blockchain and business process management in healthcare, especially for COVID-19 cases were suggested in ref. [31] which discusses the importance of artificial intelligence, IoT and blockchain technologies. An efficient, ensemble-based classification framework for big medical data was proposed in ref. [32] where the authors shared the idea of developing an ensemble-based classification framework based on machine learning.

6 Conclusion

Application of federated architecture in the healthcare system is gaining momentum in the research community. The need for novel methods, techniques, and architectures for implementing an advanced medical system using modern cryptographic techniques is in the steady rise. This article explains the overview of FL and the privacy issues that need to be addressed effectively. The challenges of federated system in medical domain

are highlighted. The future research may be extended to bring more efficient schemes in FL. Future research may include extensive research on modern access control and privacy protection schemes for federated healthcare.

Conflict of interest: Authors state no conflict of interest.

References

- W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y. -C. Liang, Q. Yang, et al., "Federated learning in mobile edge networks: A comprehensive survey.", arXiv preprint arXiv:1909.11875, 2019.
- EU, *Regulation (EU) 2016/679 of the European parliament and of the council*. Retrieved december 26, 2016. 2018 from <https://eur-lex.europa.eu/legal-content/en/txt>.
- D. Bogdanov, S. Laur, and J. Willemson, "Sharemind: A framework for fast privacy-preserving computations.", *European Symposium on Research in Computer Security*, Berlin, Heidelberg, Springer, 2008, pp. 192–206.
- R. Agrawal and R. Srikant, "Privacy-preserving data mining.", *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 2000, pp. 439–450.
- R. Rivest, L. Adleman, and M. Dertouzos, "On data banks and privacy homomorphisms." *Found. Secure Comput.* vol. 4, no. 11. pp. 169–180, 1978.
- Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications." *ACM Trans. Intell. Syst. Technol. (TIST.)*, vol. 10, no. 2. pp. 1–19, 2019.
- Q. Li, Z. Wen, and B. He, "Federated learning systems: Vision, hype and reality for data privacy and protection.", arXiv preprint arXiv:1907.09693, 2019, T19.
- K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. Brendan McMahan, S. Patel, et al., "Practical secure aggregation for privacy-preserving machine learning." *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 1175–1191.
- Y. Liu, T. Chen, and Q. Yang, "Secure federated transfer learning.", arXiv preprint arXiv:1812.03337, 2018.
- D. Chai, L. Wang, K. Chen, and Q. Yang, "Secure Federated Matrix Factorization.", arXiv preprint arXiv:1906.0, 2019, 5108.
- S. Hardy, W. Henecka, H. Ivey-Law, R. Nock, G. Patrini, G. Smith, et al., "Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption.", arXiv preprint arXiv:1711.10677, 2017.
- R. Shokri, and V. Shmatikov, "Privacy-preserving deep learning.", *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015, pp. 1310–1321.
- H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang., "Learning differentially private recurrent language models.", arXiv preprint arXiv:1710.06963, 2017.
- R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective.", arXiv preprint arXiv:1712.07557, 2017.
- Q. Chen, C. Xiang, M. Xue, B. Li, N. Borisov, D. Kaarfar, et al., "Differentially private data generative models.", arXiv preprint arXiv:1812.02274, 2018.
- K. Cheng, T. Fan, Y. Jin, Y. Liu, T. Chen, and Q. Yang, "Secureboost: A lossless federated learning framework." arXiv preprint arXiv:1901.08755, 2019.
- J. Xu and F. Wang, "Federated Learning for Healthcare Informatics." arXiv preprint arXiv:1911.06270, 2019.
- M. Sicuranza, M. Ciampi, G. D. Pietro, and C. Esposito, "Secure healthcare data sharing among federated health information systems." *Int. J. Crit. Computer-Based Syst.*, vol. 4, no. 4. pp. 349–373, 2013.
- B. Alhaqbani and C. Fidge, "Access control requirements for processing electronic health records.", *International Conference on Business Process Management*, Berlin, Heidelberg, Springer, 2007, pp. 371–382.
- J. Heurix, M. Karlinger, M. Schrefl, and T. Neubauer, "A hybrid approach integrating encryption and pseudonymization for protecting electronic health records." *Proceedings of the Eighth IASTED International Conference on Biomedical Engineering*, 2011, pp. 117–124.
- N. Stolba, M. Banek, and A. M. Tjoa, "The security issue of federated data warehouses in the area of evidence-based medicine." *First International Conference on Availability, Reliability and Security (ARES'06)*, IEEE, 2006, p. 11.
- J. Sun, X. Zhu, C. Zhang, and Y. Fang, "HCPP: Cryptography based secure EHR system for patient privacy and emergency healthcare." *2011 31st International Conference on Distributed Computing Systems*, IEEE, 2011, pp. 373–382.
- W. Schneble, and G. Thamaras, "Attack Detection Using Federated Learning in Medical Cyber-Physical Systems." *28th International Conference on Computer Communications and Networks (ICCCN)*, 2019.
- T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions.", arXiv preprint arXiv:1908.07873, 2019.
- A. Bhowmick, J. Duchi, J. Freudiger, G. Kapoor, and R. Rogers, "Protection against reconstruction and its applications in private federated learning." arXiv preprint arXiv:1812.00984, 2018.
- K. Bonawitz, V. Ivanov, B. Kreuter, H. Antonio Marcedone, B. McMahan, S. Patel, et al., "Practical secure aggregation for privacy-preserving machine learning." *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 1175–1191.
- Y.-L. Gao, X.-B. Chen, Y.-L. Chen, Y. Sun, X.-X. Niu, and Y.-X. Yang, "A secure cryptocurrency scheme based on post-quantum blockchain." *IEEE Access.*, vol. 6, pp. 27205–27213, 2018.
- S. Krawczyk, and A. Jain, "Securing electronic medical records using biometric authentication." *International Conference on Audio-and Video-Based Biometric Person Authentication*, Berlin, Heidelberg, Springer, 2005, pp. 1110–1119.
- S. Mandapati, S. Kadry, R. L. Kumar, K. Sutham, and O. Thinnukool, "Deep learning model construction for a semi-supervised classification with feature learning." *Complex. & Intell. Syst.*, pp. 1–11, 2022.
- S. Sankar, R. Somula, R. L. Kumar, P. Srinivasan, and M. A. Jayanthi, "Trust-Aware Routing Framework for Internet of Things." *Int. J. Knowl. Syst. Sci. (IJKSS)*, vol. 12, no. 1. pp. 48–59, 2021.
- I. Abunadi, and R. L. Kumar, *Blockchain and Business Process Management in Health Care, Especially for COVID-19 Cases*, Security and Communication Networks, 2021.
- F. Khan, B. V. V. Siva Prasad, S. A. Syed, I. Ashraf, and L. K. Ramasamy, *An Efficient, Ensemble-Based Classification Framework for Big Medical Data*, Big Data, 2021.