Tampere University

TIJO.414 DATA AND INFORMATION MANAGEMENT

FEDERATED LEARNING AND ITS APPLICATIONS

Individual Essay Assignment

Fizra Khan (152177548)

ABSTRACT

Introduction: Federated learning (FL) is a growing machine learning technique that utilizes a decentralized approach for training a model. This opens many possibilities in every industry specifically healthcare and medical, to minimize the issue of security and privacy. Hospitals generate heaps of patient data, and an AI model may require access to it to leverage this huge data. Due to strict regulations with data privacy, the training of AI models is a great challenge that can be addressed by FL.

Methods: In this essay, an extensive literature search was carried to understand FL, its medical and healthcare applications, challenges in implementing it and how data governance is required for this new approach

Finding: Our findings suggest that although FL claims to be secure and protects privacy, there is always some data leakage that needs to consider when building these systems. There are multiple areas like genomics and medical diagnostics where FL can be applied to. Moreover, it poses some challenges in final analysis of data and high cost. Furthermore, data governance is still applicable to this new system.

Conclusion: The uptake of FL can increase in the future if further measures are accounted for, to maintain full data privacy.

USE OF ARTIFICIAL INTELLIGENCE TOOLS FOR THE REPORT

I used an AI tool (e.g., ChatGPT) for the report: NO

Note. Unreported use of AI tools can result in a lowered grade or failure of the report.

If yes, please name the tool that you used and describe the prompts that you applied.

Tool:

Prompts:

If yes, please describe which parts of the report of the report were written with the help of AI and/or the ways you applied AI tools for producing the report: (max 0,5 page)

CONTENTS

.

1. INTRODUCTION

Artificial Intelligence (AI) is a growing paradigm that requires the usage of data to produce results. AI has multiple branches or methods, like machine learning, deep learning, neural networks, and others, that assist in leveraging the dataset to help achieve its goals. In these methods, a model is created that is trained on centralized data, i.e., a set of data stored on a server and then utilized from that source. Federated learning (FL) is one of the methods that uses decentralized i.e., data stored in different places to train the model. (Bharati, 2022)

FL is gaining much popularity in different industries including health related areas. Some of the health areas that it is used are healthcare systems, biomedical applications, digital health, medical data analysis and health informatics among others.

The need for adopting FL arises in health-related fields due to the sensitive nature of health data such as patient's personal information, history, diseases, prescriptions etc. which is deemed to be confidential and private. When using data from a single source, there is a chance of including bias in the health AI model due to lack of geographic or demographic diversity (Prayitno, 2022). Moreover, to accomplish the goals of some AI models, a large dataset is required to train the model to produce results as accurate as possible for it to be dependable in a healthcare/clinical setting. However, these healthcare systems are generally complex making them hard to navigate to acquire the dataset to achieve these results (Prayitno, 2022).

As we know, there is a surplus of data in every industry, especially in health-related areas, which enables the innovation of new technologies that deal with security, privacy and other data related concerns. In this essay, we aim to explore federated learning as a technological innovation, its application in health-related areas, implementation, and data governance challenges. An extensive literature search was carried out on the above mentioned aims and the findings are elucidated in the upcoming sections of this essay.

2. FEDERATED LEARNING (FL)

Federated learning (FL) is a machine learning technique that allows a model to be trained on dataset located with different owners. When employing federated learning, an artificial intelligence (AI) model is created and placed on a server. A group of data providers is created and selection of data providers from this group either occurs randomly or some algorithm techniques are utilized. The copies of AI model parameters are shared with the selected providers, local data against these parameters are retrieved and trained locally. In the end, the trained parameters from each provider are returned to where model is located and accumulated (Mammen, 2021).

Even though FL doesn't pose any privacy or security threats, the data transfer from data provider to the server is risky and data leakage or breaches may happen. Techniques such as differential privacy, k-order anonymity, secure multiparty computation (SMC), and homomorphic encryption addresses these issues and ensure that the privacy of the data is protected (Bharati, 2022). FL differs from machine learning only in the training technique, and the basic concept of producing an AI model using classifier among other things remains the same.

2.1 Types of Federated Learning (FL):

There are different types of FL based on feature space and sample size. These categorizations influence the choice of server to be used and algorithms related to privacy, security among others. Following are some of the types discussed:

1. **Federated Transfer Learning (FTL):** When the dataset from each provider differs in feature space and sample size. It suggests that each provider offers different required features or parameters, and the number of data samples also vary among the providers. (Bharati, 2022)

2. **Vertical Federated Learning (VFL):** When dataset from each provider differs only in feature space but the sample size is same. This type of FL is also called feature-based FL. (Bharati, 2022)

3. **Horizontal Federated Learning (HFL):** It is also called sample-based FL, where different dataset has the same feature space but differ in sample size. (Bharati, 2022)

4. **Cross-Device Federated Learning:** If there are many participating data providers, this type of FL is used. (Bharati, 2022)

5. **Cross-Silo Federated Learning:** If there is a small number of participating data providers, this type of FL is used. (Bharati, 2022)

3. APPLICATIONS OF FL IN MEDICAL/HEATLH FIELD

Some of the medical areas where federated learning is used are discussed below:

3.1 Medical Imaging:

Radiology produces heaps of data in a single healthcare organization. These data may include X-ray imaging, CT, MRI, and PET scans. These scans are used as image data to create classification models for disease prediction. For example, a COVID-19 prediction model based on chest X-rays. However, for accuracy of the prediction, it is important to train the model on large dataset, therefore, FL can be utilized to train the model at each healthcare organization locally. (Malik, 2023)

3.2 Oncology:

FL can also be used to for cancer detection. This systematic review shows how FL has been applied in different paper for detecting different cancers like brain, lung, breast cancer etc (Chowdhury, 2022). While comparing machine learning and federated learning, it has been shown that federated learning has improved the performance of the model (Chowdhury, 2022).

3.3 Genomics:

High throughput sequencing analysis is a technique that allows bioinformaticians to detect the biological variation between different groups. The identity of the person can be verified by the samples in these data which is a privacy concern while there is also a lack of omics data due to its high biological variability. Similar concept of federated learning can be utilized in the analysis of omics datasets to minimize these issues. (Zolotareva O, 2021)

3.4 Mobile Health (mHealth) and Health monitoring:

mHealth refers to the use of mobile and wearable devices to collect the data for creating a disease prediction machine learning model. For example, monitoring the health of epileptic or Parkinson's patients using sensors or smartphones. The copies of the model can be sent out to each patient's smartphone or wearable device for training and sent

back to the server for aggregation to preserve the privacy of the patient. This is also a kind of edge computing. (Wang, 2023)

3.5 Data Standardization and Optimization:

In this paper, an FL based framework has been proposed that standardize and optimize the biomedical data by customizing the dataset to avoid significant variations in large dataset produced by healthcare organizations. This shows that variations in FL frameworks can be made to achieve the required results. (Fathima, 2023)

4. CHALLENGES ASSOCIATED WITH FL

There are multiple challenges associated with implementing FL in a real-world setting:

1. In FL, a large network of different devices exist which interact with the server. When there is any issue with the network environment even at any one link in the network, for example, any device goes offline during training, it will impact the functionality of the network. (Bharati, 2022)

2. There are variations in how an organization, or a device generates data. It could be because of multiple reasons, like recording environment or technical issues. This impacts the randomness of dataset from each data provider and can impact statistical analysis. (Bharati, 2022)

3. Machine learning models are only as good as the data is. For a model to be accurate and fair, the data must be reliable. However, when the data is not disclosed during training, it poses a challenge to completely rely on the data. (Bharati, 2022)

4. In order to get the permission to access the data from each data provider, major communication is involved between all stakeholders which can hinder the timely completion of the project and may be costly. (Bharati, 2022)

5. As we know, all new methods and technology are subjected to regulations. New policies and laws may emerge to regulate this new AI system.

5. FL AND DATA GOVERNANCE DOMAINS

As FL uses data for training, even after minimizing the privacy and security risk, it must pay attention to areas related to data governance. Although, there is no formal law or regulation specifically for FL and cloud-based technology, it is subjected to a few regulations indirectly (Chalamala, 2022). The responsibility of complying with the regulations to produce a quality AI model falls on both organizations providing data and building an FL model.

Data Privacy: In FL, the main computation of data occurs at the server, it needs to comply with data privacy and protections law of GDPR. (Chalamala, 2022)

Data Security: Technological measures to ensure that the system is completely secure to process the data is necessary. This includes the server where data is processed, the devices i.e. mobile device or others, that are providing the devices, and the intermediate links where the transmission of trained parameters. (Chalamala, 2022)

Data Accessibility: The responsible personnel for using the data to train the model locally or when there is no server involved in the training, proper authentication, and authorization methods for them are required. (Chalamala, 2022)

Data Quality: There is no direct regulation on maintaining the quality of the used data. However, to build a robust and accurate FL model, which is used for predicting diseases or other health related issues, measures must be taken to ensure the high quality of data.

Data Lifecycle: When building an FL system, the health of the data utilized must also be taken into consideration. It must be shown how the data was dealt i.e. inventoried, archived, removed, during its lifecycle. (Chalamala, 2022)

6. CONCLUSION

In conclusion, as the health data is much vulnerable to breaches due to its sensitive nature, we explored FL as a technical solution in AI paradigm that proceeds to address the issues concerning privacy, security, and confidentiality of health data. It works by sending the model copies to each data owner, training locally and combining all the parameters together to make a model. Many types of FL exist based on server use such as centralized, decentralized learning, and data availability such as horizontal, vertical, federated transfer, cross device, and cross silo learning.

Our findings have shown that although the FL process minimizes the privacy related issues, as the data remains with the actual data owners without transferring, it poses breaches and attacks risk to overall model privacy during the transfer of trained parameters. Further techniques like differential privacy among others have been mentioned to tackle this issue.

There are many medical/health areas where FL can be applied such as medical imaging, oncology, genomics, mHealth and health monitoring etc. Some of the challenges associated with implementing FL are; interruptions in network connection between data providers and the final aggregation in the process, data reliability, statistical analysis due to variation in the data coming from multiple sources, and the costly and time-consuming communication among many stakeholders to carry out the whole task.

When employing an FL system, due to the usage of data for training, data governance is still required. Some of the governance domains that are applicable to FL system are, quality, accessibility, lifecycle, security, and privacy.

FL has surely proven to be a promising solution to tackle health data related issues but there is a lack of direct regulation of these systems. It is the responsibility of all stakeholders to adhere to data governance in order to increase the quality and privacy of the FL system to assist in easing the technological advancements that address health related issues around the world.

REFERENCES

1. Bharati, Subrato et al. 2022. 'Federated Learning: Applications, Challenges and Future Directions'. International Journal of Hybrid Intelligent Systems, vol. 18, no. 1-2, pp. 19-35, doi:10.3233/HIS-220006
2. Prayitno, Shyu C-R, Putra KT, Chen H-C, Tsai Y-Y, Hossain KSMT, Jiang W, Shae Z-Y. 2022. A Systematic Review of Federated Learning in the Healthcare Area: From the Perspective of Data Properties and Applications. Applied Sciences. 11(23):11191. https://doi.org/10.3390/app112311191
3. Mammen, Priyanka Mary. 2021. Federated Learning: Opportunities and Challenges. In Proceedings of ACM Conference (Conference'17). ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn
4. Malik, H., Naeem, A., Naqvi, R. A., & Loh, W. K. (2023). DMFL_Net: A Federated Learning-Based Framework for the Classification of COVID-19 from Multiple Chest Diseases Using X-rays. Sensors (Basel, Switzerland), 23(2), 743. https://doi.org/10.3390/s23020743
5. Chowdhury, A., Kassem, H., Padoy, N., Umeton, R., Karargyris, A. (2022). A Review of Medical Federated Learning: Applications in Oncology and Cancer Research. In: Crimi, A., Bakas, S. (eds) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2021. Lecture Notes in Computer Science, vol 12962. Springer, Cham. https://doi.org/10.1007/978-3-031-08999-2_1
6. Zolotareva O, Nasirigerdeh R, Matschinske J, Torkzadehmahani R, Bakhtiari M, Frisch T, Späth J, Blumenthal DB, Abbasinejad A, Tieri P, Kaissis G, Rückert D, Wenke NK, List M, Baumbach J. 2021, Dec 14. Flimma: a federated and privacy-aware tool for differential gene expression analysis. Genome Biol. 22(1):338. doi: 10.1186/s13059-021-02553-2. PMID: 34906207; PMCID: PMC8670124.
7. Wang T, Du Y, Gong Y, Choo KR, Guo Y. 2023, May 1. Applications of Federated Learning in Mobile Health: Scoping Review. J Med Internet Res. 25:e43006. doi: 10.2196/43006. PMID: 37126398; PMCID: PMC10186185.
8. Fathima AS, Basha SM, Ahmed ST, Mathivanan SK, Rajendran S, Mallik S, et al. (2023) Federated learning based futuristic biomedical big-data analysis and

standardization. PLoS ONE 18(10): e0291631. https://doi.org/10.1371/journal.pone.0291631

9. Chalamala, S.R., Kummari, N.K., Singh, A.K. et al. 2022. Federated learning to comply with data protection regulations. CSIT 10, 47–60. https://doi.org/10.1007/s40012-022-00351-0