Federated Learning Research: Trends and Bibliometric Analysis

Ali Farooq, Ali Feizollah, and Muhammad Habib ur Rehman

Abstract Federated learning (FL) allows machine learning algorithms to gain insights into a broad range of datasets located at different locations, enabling a privacy-preserving model development. Since its announcement in 2016, FL has gained interest from a variety of entities – both, in academia and industry. To understand what are the research trends in this area, a bibliometric analysis is conducted to objectively describe the research profile of the FL area. In this regard, 476 documents written in English were collected through a thorough systematic search in the Scopus database and examined from several perspectives (e.g., growth trends, top-cited papers, subject area), productivity measures of authors, institutions, and countries. Further, a co-word analysis through VOSviewer was carried out to identify the evolving research themes in FL. There has seen exponential growth in FL literature since 2018. There are five research themes, namely internet of things, wireless communication, privacy and security, data analytics, and learning and optimization, which were surfaced in the analysis. We also found that most of the documents related to FL were published in computer science, followed by engineering disciplines. It was also observed that China is at the forefront in terms of the frequency of documents in this area followed by the United States of America and Australia.

Ali Farooq
Department of Computing, University of Turku, 20500 Turku, Finland e-mail: alifar@utu.fi

Ali Feizollah
University of Malaya Halal Research Centre (UMHRC), University of Malaya, 50603 Kuala Lumpur, Malaysia, e-mail: ali.feizollah@um.edu.my

Muhammad Habib ur Rehman
Center for Cyber-Physical Systems, Khalifa University of Science and Technology, 127788 Abu Dhabi, UAE, e-mail: mhrehman@ieee.org

1 Introduction

Machine learning has become very popular recently due to an increase in computing power and abundance of data [1]. Data is considered the fuel for machine learning algorithms and it has a great impact on the model's performance. Although many datasets are publicly available, the hunger for more data persists. However, users' data is normally inaccessible due to organizational policies, otherwise, direct sharing of personal data raises privacy concerns. Considering this, various governments are regulating the privacy laws such as GDPR (EU), HIPPA (USA), and PIPEDA (Canada), to name a few.

Normally, the data is collected from users and a machine learning model is trained on that data. But, this method makes personal data exposed to the people and devices who do not own the data. To effectively address this issue, Google announced Federated Learning (FL) in 2016 [3]. FL is the concept of training an algorithm by running it locally on a user's device. It brings the algorithm to the data, rather than data to the algorithm. This operation is performed under the supervision of a central server. Upon completion of the training, only the updated parameters of the model are sent to the central server. The server aggregates all the received parameters from various devices to produce a global model [4, 21].

The FL has many applications in sales, finance, and other industries where data can not be directly collected because of intellectual property rights, privacy requirements, and data security policies. For example, in the retail industry, creating a personalized experience for users is the ultimate goal. To achieve this goal, the organization needs to have an access to users' data and shopping history to understand their behavior and shopping patterns. However, due to data privacy reasons, conventional methods are considered not to be secured. FL solves this issue by bringing the algorithm to users' data, by performing training on it, and by collecting the resultant model updates. This way, the data stays on users' devices but the retailers get the required model updates.

Although it has only been four years since the introduction of the FL, it has attracted the attention of many researchers. Therefore, it is important to examine the research progress and trends to identify the research gaps and suggest future directions. To this end, a bibliometric analysis can provide a macroscopic view of the entire field in the global context of related and neighboring fields. The understanding of the bigger picture will allow an individual to rationally choose a specific starting point for more detailed investigations in their areas of interests [5]. This analysis technique allows examining the evolution of the research domain, both topic-wise and authorship-wise [6]. Bibliometrics is the use of scholarly data to analyze publication patterns according to the author(s), topics including keywords, subject index or classification codes, affiliations of the author(s), the location where the research was conducted, sources of publications such as the journal or conference, the location where the research was published, date of publication, etc. [7]. In the recent years, this technique has been used in various fields (for example, knowledge management [8], medicine [9], business [6]) and domains (for example, strategic management [10],

information security [17], corporate social responsibility [12], application of AI in marketing [13], and COVID-19 [14]) for examining the research evolution.

Therefore we present a thorough bibliometric analysis of research on FL and perform domain profiling, to better understand this research area. The rest of this chapter is organized as follows: Section 2 discusses the materials and methods, section 3 presents the results and discussion, section 4 presents a few relevant studies, and the chapter is concluded in section 5.

2 Material and Method

In this section, data collection and analysis are discussed.

2.1 Data Collection

Since the purpose of the study was to examine the current research on FL, we first identified the keywords that could be used to search relevant literature. For this purpose, we examined the current literature and expert opinion was sought from two subject-matter experts. We used the following keywords: "federated learning", "federated machine learning", "federated ML", federated artificial intelligence", "federated AI", "federated intelligence", and "federated training". A Boolean operator OR was used in between the keywords to form a search string. These broad keywords allowed us to cover a broad range of studies with a focus on FL.

The selection of data sources is the next important step in data collection. Researchers used a variety of sources in bibliometric studies such as ours. However, most often, Web of Science, Scopus, and Google scholar are mostly selected as data sources due to their coverage, citations, accuracy, and consistency. For example, Web of Science has better coverage of journals [15], however, Scopus has a wider coverage of publications as compared to Web of science [16]. Google Scholar has better citation coverage as compared to the other two, however, there has been a criticism for inconsistency [16]. All in all, Scopus has been regarded as a better source for data due to citation counts, coverage of disciplines and consistency [17]. Further, Scopus has been found stronger than Web of Science in the area of science and technology [18], which is particularly relevant considering our study. In line with this, we selected Scopus as a data source. To have a broader set of publications we search in titles, abstracts, and keywords of publications such as journal articles, conference proceedings, book chapters, reviews, short papers, and editorials. We refer to these publications as documents in the rest of the paper. The search was conducted on October 4, 2020, without any lower timelimit. We included documents that were published in English. In this way, we identified 476 documents from the database.

2.2 Data Analysis

Bibliometric data of 476 documents were downloaded from the Scopus in a comma-separated values (CSV) file that can be used for further analysis in a variety of software tools for scientific mapping and profile analysis [19]. The data file was checked for missing data values such as authors, title, publication year, the title of the source, authors' affiliations, keywords, and citation information. No significant deficiency was found, however, we examined the keywords and change the abbreviations to full titles (for example, from IoT or iot to internet of things), used standard terminology (for example, internet of thing (iot) were changed to internet of things), and removed duplicate of terms in unique records.

Since the purpose of the study was to examine the profile and conceptual evolution of the research domain, we analyzed two phases. In Phase I, Microsoft Excel was used for performance analysis where essential characteristics of publications, such as year of publication, type of document, most prolific document sources, authors, and countries, were analyzed. VOSviewer [20], a visualization tool, was used for identifying conceptual themes based on co-occurrence of keywords, also known as co-word analysis. Co-word analysis identifies the topic themes based on the majority of documents in a dataset and considered better in comparison to co-citation analysis [17]. During the co-word analysis, the VOSviewer identifies the noun phrases (terms) from documents' titles, abstracts, and keywords and create a similarity matrix. This matrix shows the frequency with which two terms have appeared together in the dataset. Thereafter, the terms are clustered and mapped based on "visualization of similarities" (VOS) mapping techniques[20]. Here it is pertinent to mention that we removed the keywords (that is, "federated learning", "federated machine learning", "federated ML", federated artificial intelligence", "federated AI", "federated intelligence", and "federated training") used in data search as it was likely that when visualizing the data the selected keywords would overshadow other terms. The parameters used in Table 1 are used for preparing the visualization shown in the result section.

Table 1: Parameters Used for Visualization in VOSviewer.

Method	Value
Counting Method	Binary
Threshold (Minimum number of occurrences of a term	5
Number of terms (most relevant)	183 (100%)
Normalization Method	Factorization

3 Results and Discussion

This section presents the results of the bibliometric analysis on the FL. This research domain is still new and we found 476 documents in the Scopus database. These findings are important since they map the current research and unravel potential gaps in the research domain.

3.1 Growth Pattern Over the Years

Figure 1 shows the growth in the domain of FL.

Fig. 1: Growth Pattern Over the Years

As it can be seen, after the introduction of the concept in 2016, there is a massive growth in this topic. Specifically, after 2018, there is a considerable jump in published articles in this domain. The main idea of FL proposed by Google was to train machine learning models using distributed datasets scattered across multiple devices. This idea has been improved as the popularity of FL has been rising. Recently, the research works have been focusing on personalizing the FL [22, 23], the security aspect of the domain[24, 25], and improving statistical challenges [23]. Due to the challenging nature of FL, which includes users' devices' reliability and unbalanced data distribution, the research works have been moving towards on-device FL [27].

Most of the publications within FL were in the area of computer science and engineering. The topic generated interest in other subjects as well. For example, as shown in Table 2, a significant number of papers were published in the decision sciences which is close to business and economics, physics and astronomy, material sciences, as well as medicine. The list shown in Table 2 is directly taken from Scopus. The sum of the number of papers may exceed the total number of documents used for analysis in this chapter as one paper can belong to more than subject areas.

Table 2: Subject Area with Highest Number of Papers

Subject Area	Number of Papers
Computer Science	422
Engineering	204
Mathematics	97
Decision Sciences	77
Physics and Astronomy	23
Materials Science	22
Medicine	22
Social Sciences	12
Energy	8
Biochemistry, Genetics and Molecular Biology	6
Health Professions	5
Business, Management and Accounting	4
Chemistry	3
Agricultural and Biological Sciences	2
Chemical Engineering	2
Multidisciplinary	2
Dentistry	1
Economics, Econometrics and Finance	1
Neuroscience	1
Pharmacology, Toxicology and Pharmaceutics	1

3.2 Top Cited Papers

As we can see in Figure 1, the publications in the area of FL saw steady growth from 2017. We identified the top 10 most cited papers in the area and are shown in Table 3. The citations were taken from the Scopus database and may differ from the Web of Science or Google Scholar. We did not normalize the citations based on years. The paper titled, "Communication-efficient learning of deep networks from decentralized data" [26], is the most cited paper in the research area, followed by paper titled, "Practical secure aggregation for privacy-preserving machine learning" [28], and "Federated machine learning: Concept and applications" [29].

Table 3: Ten Top Cited Papers

Number	Title	Year	Source	Citations
1	Communication-efficient learning of deep networks from decentralized data	2017	Proceedings of the 20th International Conference on Artificial Intelligence and Statistics	194
2	Practical secure aggregation for privacy-preserving machine learning	2017	Proceedings of the ACM Conference on Computer and Communications Security	147
3	Federated machine learning: Concept and applications	2019	ACM Transactions on Intelligent Systems and Technology	102
4	Federated multi-task learning	2017	Advances in Neural Information Processing Systems	84
5	Adaptive Federated Learning in Resource Constrained Edge Computing Systems	2019	IEEE Journal on Selected Areas in Communications	53
6	In-edge AI: Intelligentizing mobile edge computing, caching and communication by federated learning	2019	IEEE Network	42
7	Federated Learning over Wireless Networks: Optimization Model Design and Analysis	2019	Proceedings - IEEE INFOCOM	33
8	Client Selection for Federated Learning with Heterogeneous Resources in Mobile Edge	2019	IEEE International Conference on Communications	29
8	Federated learning of predictive models from federated Electronic Health Records	2018	International Journal of Medical Informatics	29
8	Personalizing access to learning networks	2018	ACM Transactions on Internet Technology	29
9	Federated Learning for Ultra-Reliable Low-Latency V2V Communications	2018	2018 IEEE Global Communications Conference, GLOBECOM 2018 - Proceedings	26
10	VerifyNet: Secure and Verifiable Federated Learning	2020	IEEE Transactions on Information Forensics and Security	22

3.3 Productivity Measures

In this sub-section, we describe the domain profile in terms of productivity measures such as prolific authors, institutions with the highest research production, countries with most papers produced, and type of sources.

3.3.1 Authors

We show the top 10 authors with the highest number of published papers in Table 4. As it can be seen, Niyato, D. published 12 papers that is the highest number among other authors [30][31][32][33][34][35][36][37][38][39][40][41]. His papers are mostly about FL in mobile networks as well as mobile edge network. He also published a paper related to 5G and FL. Three authors published 9 papers, namely Bennis, M., Saad, W., and Yu, S. Similarly, four authors published 8 papers, namely Chen, M., Yang, Q., Yu, H., and Zhang, J. The last four authors published 7 papers whose names are Hong, C.S., Li, H. Tran, N.H.n, and Xu, G.

Table 4: Top 10 Authors with Highest Number of Papers

Position	Authors	Number of Papers
1	Niyato, D.	12
2	Bennis, M.	9
3	Saad, W.	9
4	Yu, S.	9
5	Chen, M.	8
6	Yang, Q.	8
7	Yu, H.	8
8	Zhang, J.	8
9	Hong, C.S.	7
10[1]	Li, H.	7
10[1]	Tran, N.H.	7
10[1]	Xu, G.	7

[1]Published equal number of articles

3.3.2 Institutes

This section provides a list of the top 10 institutions that published their work on FL. Table 5 provides a list of institutions in descending order.

Table 5: Top 10 Institutions with Highest Number of Published Papers

Institution	Number of Published Papers
University of Electronic Science and Technology of China	20
Beijing University of Posts and Telecommunications	20
Nanyang Technological University	19
Princeton University	17
University of Technology Sydney	16
Hong Kong University of Science and Technology	15
Tsinghua University	15
School of Computer Science and Engineering	14
IBM Thomas J. Watson Research Center	13
Kyung Hee University	11
Imperial College London	11
Chinese Academy of Sciences	11

Note: last three institutions have equal number of publications.

As it can be seen, the University of Electronic Science and Technology of China and Beijing University of Posts and Telecommunications have published 20 papers, which is the highest number of papers among institutions. The third-place goes to the Nanyang Technological University with 19 published papers. The Princeton University and the University of Technology Sydney are ranked fourth and fifth by publishing 17 and 16 published papers, respectively. The next two institutions published the same number of papers, namely Hong Kong University of Science and Technology, and Tsinghua University by 15 papers. Fourteen papers were published

by the School of Computer Science and Engineering, and thirteen papers were published by the IBM Thomas J. Watson Research Center. The last 3 institutions published 11 papers. They are Kyung Hee University, Imperial College London, and the Chinese Academy of Sciences.

3.3.3 Countries

In the previous section, we saw the top 10 institutions producing research in the area of FL. Table 6 shows a list of the top 10 countries that have produced research in FL. Most documents were produced from China, followed by the United States. Australia remained at the 3rd position. In the top 10 list, four countries are from Asia, three from Europe, and two from North America.

Table 6: Top 10 Countries in terms of Publications

Country	Number of documents
China	149
United States	128
Australia	36
United Kingdom	35
Singapore	24
South Korea	23
Hong Kong	16
Canada	14
Finland	11
France	10

3.3.4 Sources

In terms of publication types, out of 476 documents, 300 were conference proceedings, followed by 138 documents from journal publications. Table 7 lists different types of documents and the corresponding number of documents. In terms of sources, most (36) documents were published in Springer's lecture notes series, followed by the IEEE International Conference on Communication. The top 10 sources contain three journals: IEEE Access, IEEE Intelligent Systems, and IEEE Internet of Things Journal. Table 8 lists top sources and the corresponding number of documents published in these sources.

Table 7: Number of Different Document Types

Document Type	Number of Documents
Conference Paper	300
Article	138
Conference Review	25
Review	7
Book Chapter	3
Editorial	1
Letter	1
Short Survey	1
Total	**476**

Table 8: Top 10 Document Sources in Descending Order

Source Name	Number of documents
Lecture Notes In Computer Science Including Subseries Lecture Notes In Artificial Intelligence And Lecture Notes In Bioinformatics	36
IEEE International Conference On Communications	20
ICASSP IEEE International Conference On Acoustics Speech And Signal Processing Proceedings	13
ACM International Conference Proceeding Series	12
Ceur Workshop Proceedings	11
IEEE Access	11
IEEE Intelligent Systems	11
IEEE Internet of Things Journal	10
Communications In Computer And Information Science	7
Proceedings IEEE INFOCOM	7

3.4 Domain Profile

Based on the documents we found in our research, we clustered them according to the keywords related to each paper. Figure 2 shows the retrieved document based on the clusters. We identified 5 clusters in the figure. By clustering the documents, we get a clearer picture of the ongoing research in the FL domain. As we can see, the clusters in red and blue are larger. Therefore, they include more research papers than other clusters. Table 9 shows the keywords in each cluster and our suggested name for each cluster.

In each of the following sub-sections, we describe the cluster domain and mention some papers published in the cluster.

3.4.1 Cluster 1: Internet of Things

The keywords in the first cluster suggest that this cluster deals with the IoT and FL. A collection of sensors constitutes the IoT, which is a suitable application for machine learning to improve the efficiency of the network. However, due to the distributed

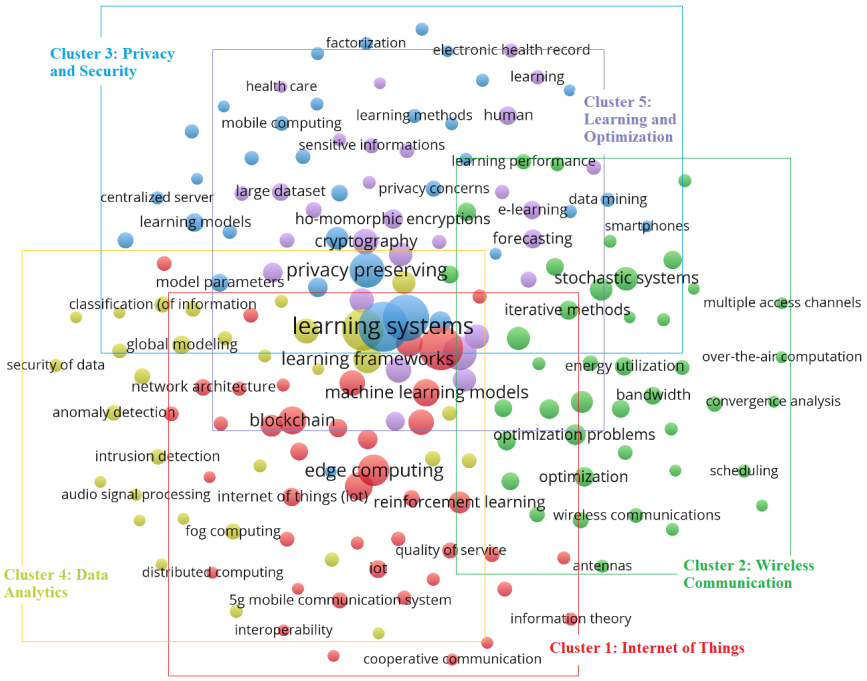

Fig. 2: Research Themes in FL

nature of IoT as well as privacy issues, this domain is a great candidate for FL application.

Among many kinds of research conducted in IoT and FL, Duan et. al [42] focused on internet video service providers (IVSPs) that are used to recommend videos to users in different geo-distributed manner. Due to its distributed nature, the authors applied FL to their network. They found that because the agents in the distributed network need to receive federated training instructions and to send the weights and trained model, the network traffic is high. Therefore, they combined low-rank matrix factorization and the 8-bit quantization method to reduce the used network bandwidth.

Feraudo et. al [43] considered edge computing as a candidate for FL. Edge computing is designed to speed up the IoT networks. The edge devices are an important part of an IoT network since they transfer data from IoT devices to the cloud. Therefore, applications like intrusion detection are crucial in this type of network. FL is a great option to train a model without compromising the security of the edge devices. The authors proposed a system based on the open-source manufacturer usage description (MUD) and the PySyft framework.

Lu et. al [44] examined the use of FL in blockchain architecture and the industrial internet of things (IIoT). They first designed a blockchain system for secure data

Table 9: Different Themes in FL Research

Cluster Color	Top 20 Most Occurring Terms	Name of Theme
Red (1)	machine learning, edge computing, blockchain, internet of things, machine learning models, artificial intelligence, data sharing, distributed machine learning, reinforcement learning, network security, distributed computer systems, IoT, transfer learning, IoT, security and privacy, network architecture, 5G mobile communication systems, decision making, quality of service, computation theory.	IoT
Green (2)	stochastic systems, communication overheads, gradient methods, wireless networks, optimization problems, economic and social effects, optimization, iterative methods, stochastic gradient descent, communication efficiency, energy utilization, incentive mechanism, bandwidth, communication rounds, efficiency, mobile telecommunication systems, resource allocation, signal processing, numerical results, energy efficiency.	Wireless Communication
Blue (3)	learning systems, data privacy, privacy preserving, privacy, privacy preservation, neural networks, learning models, model parameters, state of the art, adversarial networks, privacy concerns, mobile computing, privacy-preserving, data mining, learning methods, privacy leakages, sensitive data, training process, benchmark datasets, centralized server.	Privacy and Security
Yellow (4)	deep learning, learning frameworks, deep neural networks, communication cost, global modeling, classification (of information), fog computing, poisoning attacks, anomaly detection, central servers, intrusion detection, real-world datasets, benchmarking, data distribution, training data, computer aided instruction, data handling, automation, collaborative training, computation costs.	Data Analytics
Purple (5)	learning algorithms, cryptography, differential privacy, big data, distributed learning, privacy protection, digital storage, collaborative learning, homomorphic encryptions, forecasting, human, e-learning, large dataset, cloud computing, clustering algorithms, sensitive information, diagnosis, distributed data, medical imaging, secure multi-party computation.	Learning and Optimization

sharing for distributed users. They then introduced FL to securely share the data model in the network. We just mentioned a couple of research works related to IoT and FL research. There is an opportunity in this domain for research since the FL is a great solution to the distributed agents in IoT that involving sensitive and personal data.

3.4.2 Cluster 2: Wireless Communication

The second cluster is found to be related to wireless communication. Specifically, focusing on communication costs and data transfer over the slow and expensive network. Wu et. al [45] researched a common application, which is a mobile keyboard suggestion. It aims at predicting the next word or phrase to ease the user interaction with the devices. The problem occurs when users' inputs are needed to train a

model. Since these input data are personal and sensitive, the authors proposed an FL approach to preserve the privacy of users and to reduce communication costs. They achieved a solution by using adaptive aggregation of weights in model training, mediation incentive scheme, and Top-K strategy. They conducted experiments on three datasets such as Penn Treebank, WikiText-2, and Yelp. The authors reported robust performance with results outperforming baseline approaches.

Luping et. al [46] specifically focused on communication overhead in FL. The problem occurs with mobile devices where the data plan is limited and the network connection to the central server is slow. The authors mentioned that the current works focus on compressing the data to reduce the bit transferred. Therefore, they proposed an orthogonal approach that identifies irrelevant updates from the client and excludes them from uploading to reduce the network traffic. The authors experimented with their proposed method by comparing it with the traditional FL. They reported that their method improves communication efficiency by 5.7 times with 4% higher prediction accuracy.

3.4.3 Cluster 3: Privacy and Security

This cluster includes research works related to privacy and security. This topic is the main characteristic of FL. Chandiramani et. al [47] compared basic machine learning, distributed machine learning, and FL to investigate how they perform in terms of data privacy and security. Their results showed that the FL method maintained privacy, and performed well with the fast deployment of models in mobile and low-compute devices.

3.4.4 Cluster 4: Data Analytics

This cluster is focused on data analytics, which is a necessary part of data science. The role of FL in this cluster is significant. The data is stored on users' devices and using FL, the analysis is performed with preserving the privacy of the users. Zhao et. al [50] researched intrusion detection. With the fact that data analytics need to take data from many users into the account, privacy issues rises. It is not possible to train a model on a single user, and it violates the privacy of users to access their data and collect them for model training. The authors proposed an intrusion detection model based on FL and long short-term memory (LSTM) algorithm. First, the model from the central server is sent to all users. Then, each user trains the algorithm using their data, and they upload the model's parameters to the central server. Finally, the central server aggregates the model parameters and updates the global model. The authors experimented with their proposed method and achieved higher accuracy and better consistency than conventional models.

Schneible and Lu [51] considered anomaly detection in edge computing where there are many small devices and sensors with intermittent connectivity. The authors proposed an autoencoder, specialized deep learning neural networks, model to detect

anomalies in edge devices. The model is deployed to devices to perform data analytics and detection of anomalies in a distributed fashion. Once the devices have access to network connectivity, they upload their observations to the central server. The central server then aggregates all the results from the devices and updates the global model. Then, the updated global model is sent back to the devices. This ensures that the devices have updated models. This method reduces bandwidth and connectivity requirements for the edge devices.

3.4.5 Cluster 5: Learning and Optimization

This final cluster is about learning and optimization. It includes learning algorithms and optimizing them for some applications. Balachandar et. al [48] researched medical imaging. Patients' data is very sensitive and sharing this data to train the algorithms is risky. Therefore, the authors tried to use FL to share patients' data. They mentioned that "optimizations included proportional local training iterations, cyclical learning rate, locally weighted minibatch sampling, and cyclically weighted loss. We evaluated our optimizations on simulated distributed diabetic retinopathy detection and chest radiograph classification." Their results showed 98.6% of accuracy.

In another application, Doku et. al [49] researched big data and distributed data sources to disrupt data centralization. They mentioned that majority of data is in the possession of several big companies. They proposed an approach to combine blockchain and FL to store data in a decentralized manner.

4 Related Work

To establish the position of this chapter among similar works, and to further clarify the contributions of this chapter, we examine similar research works in this section. The available literature suggests that the current review papers fall into two categories. In the first category, papers focus on the general overview of the FL domain. As an example, Yang et. al [52] published a research work that discussed the concept and applications of FL. The authors proposed a secure FL framework that supplements the original proposal by Google in 2016. They introduced horizontal FL, vertical FL, and federated transfer learning. They then surveyed existing works that fall into the proposed framework.

Li et. al [53] discussed the general characteristics of FL and its challenges. They also provided a general overview of the current research directions. The authors mentioned four challenges of FL. They are expensive communication, systems heterogeneity, statistical heterogeneity, and privacy concerns. Furthermore, the current research works were classified based on their objectives to solve one of the introduced challenges. Finally, the authors mentioned future directions in the FL research domain. Another work that focuses on reviewing the general aspect of the FL was

published by Lo et. al [54]. In their paper, the author systematically reviewed FL, and from the software engineering perspective. The authors examined 231 primary studies. The results suggest that most of the motivation of the published works target the FL challenges, including data privacy, communication efficiency, and statistical heterogeneity.

The second category is the papers that focused on a specific aspect of the FL. For instance, Lim et. al [55] focused on the applications of the FL in mobile edge networks. They first discussed the general concept of the FL and then narrowed the discussion down to the mobile edge networks. Security and privacy is another niche area that Mothukuri et. al [56] discussed. This paper aimed at providing a comprehensive study on the security and privacy aspects of FL. The author reviewed various styles of implementation of the security and privacy frameworks in FL. They concluded that there are fewer privacy-specific threats associated with FL compared to security threats. The most specific security threats currently are communication bottlenecks, poisoning, and backdoor attacks while inference-based attacks are the most critical to the privacy of FL.

This research work falls into the first category, which is a general overview of the current research works. However, we focused on the bibliometric study of the FL, which bridges the gap in the journey of the FL research domain. It is important to be aware of the prominent authors and influential institutes in this field. Additionally, we grouped all the published works into five categories and discussed each group. This helps us understand which group needs more attention from the research community.

5 Conclusion and Future Research Directions

The purpose of this study was to examine and describe the intellectual profile of research focused on federated learning (FL) over the past 25 years. In this regard, through a systematic search, 476 relevant documents published in the English language were extracted from the Scopus and examined. FL as a research area witnessed an exponential growth since 2018. Most of FL research is published in form of conference proceedings in the subject area of computer science and engineering. China has been at the forefront of publishing the most number of documents. In terms of universities, the University of Electronic Science and Technology, Beijing University of Posts and Telecommunications from China, and Nanyang Technological University from Singapore were the top three producers of research in FL. The current research in FL can be divided into five thematic areas: Internet of things, wireless communication, privacy and security, data analytic, and learning and optimization. While this study provides state-of-the-art in the research area, it is also affected by some limitations. For example, only one source has been used to identify related documents. We expect some publications that are not indexed in the Scopus database will be left out. Secondly, we took a quantitative approach to examine the research and bibliometric data was used. We did not run a content analysis of the publication.

References

1. Alpaydin, E. (2020). Introduction to machine learning. MIT press.
2. 2019. General Data Protection Regulation GDPR. https://gdpr-info.eu/
3. H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise AgÃijera y Arcas. 2016. Communication-Efficient Learning of Deep Networks from Decentralized Data. arXiv:1602.05629
4. Canh Dinh, Nguyen H Tran, Minh NH Nguyen, Choong Seon Hong, Wei Bao, Albert Zomaya, and Vincent Gramoli. 2019. Federated Learning over Wireless Networks: Convergence Analysis and Resource Allocation. arXiv preprint arXiv:1910.13067 (2019).
5. Kostoff, R. N., Toothman, D. R., Eberhart, H. J., & Humenik, J. A. (2001). Text mining using database tomography and bibliometrics: A review. Technological Forecasting and Social Change, 68(3), 223-253.
6. Donthu, N., Kumar, S., & Pattnaik, D. (2020). Forty-five years of journal of business research: a bibliometric analysis. Journal of Business Research, 109, 1-14.
7. McBurney, M. K., & Novak, P. L. (2002, September). What is bibliometrics and why should you care?. In Proceedings. IEEE international professional communication conference (pp. 108-114). IEEE.
8. Gaviria-Marin, M., Merigó, J. M., & Baier-Fuentes, H. (2019). Knowledge management: A global examination based on bibliometric analysis. Technological Forecasting and Social Change, 140, 194-220.
9. Liao, H., Tang, M., Luo, L., Li, C., Chiclana, F., & Zeng, X. J. (2018). A bibliometric analysis and visualization of medical big data research. Sustainability, 10(1), 166.
10. Ferreira, J. J. M., Fernandes, C. I., & Ratten, V. (2016). A co-citation bibliometric analysis of strategic management research. Scientometrics, 109(1), 1-32
11. Olijnyk, N. V. (2015). A quantitative examination of the intellectual profile and evolution of information security from 1965 to 2015. Scientometrics, 105(2), 883-904
12. Bhattacharyya, S. S.,& Verma, S. (2020). The intellectual contours of corporate social responsibility literature. International Journal of Sociology and Social Policy.
13. Mekhail, M., Salminen, J., Ple, L., & Wirtz, J. (2021). Artificial Intelligence in Marketing; Bibliometric Analysis, Topci Modeling and Research Agenda.Journal of Business Research. forthcoming
14. Verma, S., & Gustafsson, A. (2020). Investigating the emerging COVID-19 research trends in the field of business and management: A bibliometric analysis approach. Journal of Business Research, 118, 253-261.
15. Adriaanse, L. S., & Rensleigh, C. (2013). Web of science, scopus and Google Scholar. The Electronic Library
16. Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, web of science, and Google scholar: strengths and weaknesses. The FASEB journal, 22(2), 338-342.
17. Olijnyk, N. V. (2015). A quantitative examination of the intellectual profile and evolution of information security from 1965 to 2015. Scientometrics, 105(2), 883-904.
18. de Moya-Anegón, F., Chinchilla-Rodríguez, Z., Vargas-Quesada, B., Corera-Álvarez, E., Muñoz-Fernández, F., González-Molina, A., & Herrero-Solana, V. (2007). Coverage analysis of Scopus: A journal metric approach. Scientometrics, 73(1), 53-78.
19. Cobo, M.J.; López-Herrera, A.G.; Herrera-Viedma, E.; Herrera, F. Science mapping software tools: Review, analysis, and cooperative study among tools. J. Am. Soc. Inf. Sci. Technol. 2011, 62, 1382–1402.
20. Van Eck, N.J.; Waltman, L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 2010, 84, 523–538.
21. Peter Kairouz, H. Brendan McMahan, Brendan Avent, AurÃl'lien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, AdriÃă GascÃşn, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang

He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub KoneÄŊnÃ¡, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, TancrÃĺde Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer ÃŰzgÃijr, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian TramÃĺr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. 2019. Advances and Open Problems in Federated Learning. arXiv:1912.04977

22. title=Federated Meta-Learning with Fast Convergence and Efficient Communication, author=Fei Chen and Mi Luo and Zhenhua Dong and Zhenguo Li and Xiuqiang He, year=2019, eprint=1802.07876, archivePrefix=arXiv, primaryClass=cs.LG
23. Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S. Talwalkar. 2017. Federated multi-task learning. In Advances in Neural Information Processing Systems 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4424–4434. http://papers.nips.cc/paper/ 7029-federated-multi-task-learning.pdf.
24. Robin C. Geyer, Tassilo Klein, and Moin Nabi. 2017. Differentially private federated learning: A client level perspective. CoRR abs/1712.07557 (2017). arxiv:1712.07557 http://arxiv.org/abs/1712.07557.
25. Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2017. Practical secure aggregation for privacy-preserving machine learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS'17). ACM, New York, NY, 1175–1191. DOI:https://doi.org/10.1145/3133956.3133982
26. McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In Artificial Intelligence and Statistics (pp. 1273-1282). PMLR.
27. Rehman,M. H., Salah, K., Damiani, E., and Svetinovic, D., "Towards Blockchain-Based Reputation-Aware Federated Learning," IEEE INFOCOM 2020 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Toronto, ON, Canada, 2020, pp. 183-188, doi: 10.1109/INFOCOMWKSHPS50562.2020.9163027.
28. Bonawitz, Keith, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. "Practical secure aggregation for privacy-preserving machine learning." In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 1175-1191. 2017.
29. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST), 10(2), 1-19.
30. Zou, Y., Feng, S., Niyato, D., Jiao, Y., Gong, S., & Cheng, W. (2019, July). Mobile device training strategies in federated learning: An evolutionary game approach. In 2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) (pp. 874-879). IEEE.
31. Feng, S., Niyato, D., Wang, P., Kim, D. I., & Liang, Y. C. (2019, July). Joint service pricing and cooperative relay communication for federated learning. In 2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) (pp. 815-820). IEEE.
32. Kang, J., Xiong, Z., Niyato, D., Yu, H., Liang, Y. C., & Kim, D. I. (2019, August). Incentive design for efficient federated learning in mobile networks: A contract theory approach. In 2019 IEEE VTS Asia Pacific Wireless Communications Symposium (APWCS) (pp. 1-5). IEEE.
33. Zou, Y., Feng, S., Xu, J., Gong, S., Niyato, D., & Cheng, W. (2019, August). Dynamic Games in Federated Learning Training Service Market. In 2019 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM) (pp. 1-6). IEEE.

34. Anh, T. T., Luong, N. C., Niyato, D., Kim, D. I., & Wang, L. C. (2019). Efficient training management for mobile crowd-machine learning: A deep reinforcement learning approach. IEEE Wireless Communications Letters, 8(5), 1345-1348.
35. Kang, J., Xiong, Z., Niyato, D., Xie, S., & Zhang, J. (2019). Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory. IEEE Internet of Things Journal, 6(6), 10700-10714.
36. Yu, Han, Zelei Liu, Yang Liu, Tianjian Chen, Mingshu Cong, Xi Weng, Dusit Niyato, and Qiang Yang. "A fairness-aware incentive scheme for federated learning." In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 393-399. 2020.
37. Kang, J., Xiong, Z., Niyato, D., Zou, Y., Zhang, Y., & Guizani, M. (2020). Reliable federated learning for mobile networks. IEEE Wireless Communications, 27(2), 72-80.
38. Yu, Han, Zelei Liu, Yang Liu, Tianjian Chen, Mingshu Cong, Xi Weng, Dusit Niyato, and Qiang Yang. "A Sustainable Incentive Scheme for Federated Learning." IEEE Intelligent Systems (2020).
39. Lim, Wei Yang Bryan, Nguyen Cong Luong, Dinh Thai Hoang, Yutao Jiao, Ying-Chang Liang, Qiang Yang, Dusit Niyato, and Chunyan Miao. "Federated learning in mobile edge networks: A comprehensive survey." IEEE Communications Surveys & Tutorials (2020).
40. Liu, Y., James, J. Q., Kang, J., Niyato, D., & Zhang, S. (2020). Privacy-preserving Traffic Flow Prediction: A Federated Learning Approach. IEEE Internet of Things Journal.
41. Liu, Y., Peng, J., Kang, J., Iliyasu, A. M., Niyato, D., & El-Latif, A. A. A. (2020). A Secure Federated Learning Framework for 5G Networks. arXiv preprint arXiv:2005.05752.
42. Duan, S., Zhang, D., Wang, Y., Li, L., & Zhang, Y. (2019). JointRec: A Deep Learning-based Joint Cloud Video Recommendation Framework for Mobile IoTs. IEEE Internet of Things Journal, 1–1. doi:10.1109/jiot.2019.2944889.
43. Feraudo, Angelo, Poonam Yadav, Vadim Safronov, Diana Andreea Popescu, Richard Mortier, Shiqiang Wang, Paolo Bellavista, and Jon Crowcroft. "CoLearn: enabling federated learning in MUD-compliant IoT edge networks." In Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking, pp. 25-30. 2020.
44. Lu, Yunlong, Xiaohong Huang, Yueyue Dai, Sabita Maharjan, and Yan Zhang. "Blockchain and federated learning for privacy-preserved data sharing in industrial IoT." IEEE Transactions on Industrial Informatics 16, no. 6 (2019): 4177-4186.
45. Wu, Xing, Zhaowang Liang, and Jianjia Wang. "Fedmed: A federated learning framework for language modeling." Sensors 20, no. 14 (2020): 4048.
46. Luping, W. A. N. G., Wei, W. A. N. G., & Bo, L. I. (2019, July). Cmfl: Mitigating communication overhead for federated learning. In 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS) (pp. 954-964). IEEE.
47. Chandiramani, K., Garg, D., & Maheswari, N. (2019). Performance Analysis of Distributed and Federated Learning Models on Private Data. Procedia Computer Science, 165, 349-355.
48. Balachandar, Niranjan, Ken Chang, Jayashree Kalpathy-Cramer, and Daniel L. Rubin. "Accounting for data variability in multi-institutional distributed deep learning for medical imaging." Journal of the American Medical Informatics Association 27, no. 5 (2020): 700-708.
49. Doku, Ronald, Danda B. Rawat, and Chunmei Liu. "Towards federated learning approach to determine data relevance in big data." In 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI), pp. 184-192. IEEE, 2019.
50. Zhao, Ruijie, Yue Yin, Yong Shi, and Zhi Xue. "Intelligent intrusion detection based on federated learning aided long short-term memory." Physical Communication 42 (2020): 101157.
51. Schneible, Joseph, and Alex Lu. "Anomaly detection on the edge." In MILCOM 2017-2017 IEEE Military Communications Conference (MILCOM), pp. 678-682. IEEE, 2017.
52. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST), 10(2), 1-19.
53. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50-60.
54. Lo, Sin Kit, Qinghua Lu, Chen Wang, Helen Paik, and Liming Zhu. "A systematic literature review on federated machine learning: From a software engineering perspective." arXiv preprint arXiv:2007.11354 (2020).

55. Lim, Wei Yang Bryan, Nguyen Cong Luong, Dinh Thai Hoang, Yutao Jiao, Ying-Chang Liang, Qiang Yang, Dusit Niyato, and Chunyan Miao. "Federated learning in mobile edge networks: A comprehensive survey." IEEE Communications Surveys & Tutorials (2020).
56. Mothukuri, Viraaji, Reza M. Parizi, Seyedamin Pouriyeh, Yan Huang, Ali Dehghantanha, and Gautam Srivastava. "A survey on security and privacy of federated learning." Future Generation Computer Systems 115 (2020): 619-640.