Financial Privacy Computing Applications with Distributed Machine Learning

Guanhao Feng

International School of Information Science and Engineering, Dalian University of Technology, Dalian, 116024, China

Corresponding author email: 2966382581@mail.dlut.edu.cn

ABSTRACT. With the further development of big data applications, data privacy and security have attracted great attention from all countries in the world. There are nearly 100 countries and regions in the world that have laws related to data security protection. Through legislation, data users have more control to control their personal data. At the same time, most of the industry data present the phenomenon of data island, how to carry out cross-organizational data cooperation under the premise of meeting user privacy protection, data security and government regulations is a major problem, and federated machine learning will become the key technology to solve this industry problem.

Keywords: Federated learning, private computing, distributed machine learning, big data

1 Introduction

With the wave of digitalization sweeping the world, a large amount of data has been generated, which contains huge commercial value behind these data. How to dig into the potential commercial value behind the data has become a business issue faced by various data organizations. However, with the increasing attention of various countries to data privacy and security constraints, the use of data is limited by laws and regulations, personal privacy and other data privacy security constraints, resulting in the data between various data organizations facing the problem of data island, the rational use of data difficulties, namely difficulty 1: data security sharing is difficult, enterprises and institutions are often difficult to establish a trust relationship, cannot achieve data security sharing; Difficulty 2: It is difficult to use data effectively. Data files between enterprises and institutions usually have different data formats, which leads to a time-consuming and laborious data fusion step. Difficulty 3: data transmission is difficult. Data files between enterprises and institutions are transmitted through secure channels, and high frequency transmission has high requirements for transmission efficiency and transmission cost.

In response to the dilemma of data island and data privacy, experts have proposed the concept of federated machine learning, and various data institutions have followed

C. Chen et al. (eds.), *Proceedings of the 3rd International Conference on Digital Economy and Computer Application (DECA 2023)*, Atlantis Highlights in Computer Sciences 17,
https://doi.org/10.2991/978-94-6463-304-7_58

up. Especially in the financial field, various data institutions have built a federated learning platform to build privacy computing service capabilities, realize the ability of data "available invisible, controllable and measurable", and empower the innovation and development of industry business.

2 Machine Learning Architecture

2.1 Centralized Machine Learning Architectures

In a centralized machine learning architecture, the data provider processes the local data, encrypts it, sends the processed data to a trusted execution environment for training, and then transmits the trained model back to the local machine. During model training, a key challenge is to minimize the lifetime of high-priority model training tasks. [1] Since the data of each party is encrypted, it cannot be seen by the platform or other participants. When multiple participants or others have authorized access to the shared model, they can use the corresponding programming interface (API) to input information and get the return value. Before providing data, participants (data providers) verify the correctness and validity of the software running in the trusted execution environment through remote attestation. After the data is provided, an encryption mechanism is negotiated between the participant and the trusted execution environment. The participant encrypts the data using the negotiation mechanism and sends the cipher text to the trusted execution environment. In order to ensure that the calculation model will not be leaked, the trained model is encrypted after the whole calculation is completed.

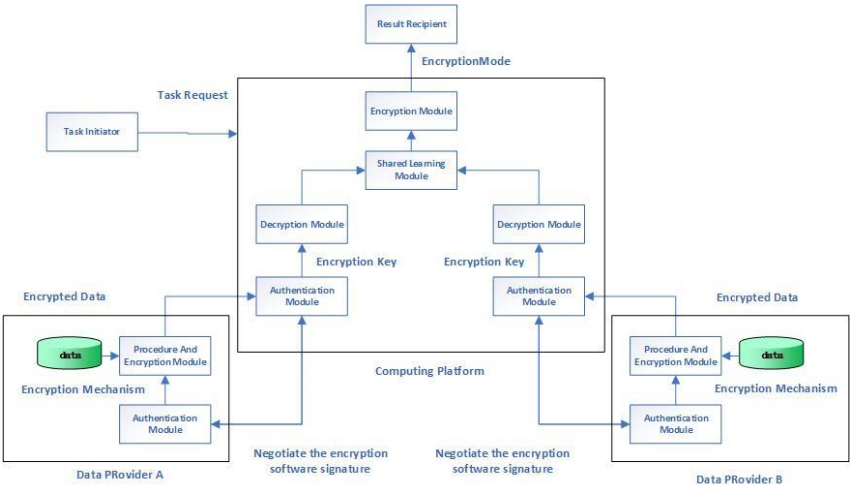

Fig. 1. Architecture of Centralized Learning System

As shown in Fig.1, in the centralized machine learning system architecture, each data provider consists of a process and encryption module, an authentication module and data. The data provider provides this data as input to the computing platform or to other

data providers (which also have computing power). Subsequently, the encrypted secret key obtained from the authentication module processes and encrypts the data of each data provider through the authentication module, and then uploading the encrypted data to the trusted execution environment computing platform. The authentication module of the computing platform decrypts the encrypted data through the decryption module, and then sends the decrypted data to the shared learning module. The shared learning module performs shared machine learning operations on the decrypted data from multiple data providers and sends the encrypted data to the data receiver through the encryption module of the computing platform.

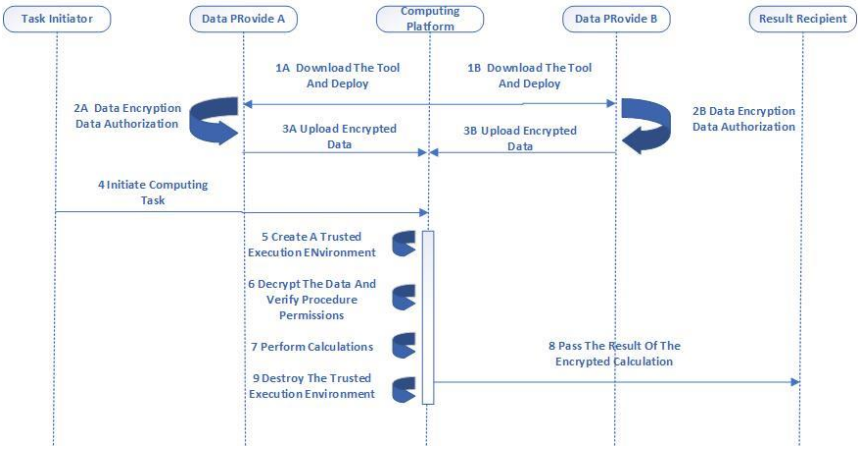

Fig. 2. The process of a centralized learning system

As shown in the Fig.2 above, in the process of a centralized learning system, any data provider or computing platform can initialize computing tasks, and then the computing platform will create a trusted execution environment. The local data provided by each data provider can be processed, encrypted, and uploaded to the computing platform. The platform decrypts the received encrypted data sent by each data provider in the trusted execution environment, and executes module training based on this decrypted data to obtain a shared model. The data processing, encryption, decryption, and training steps can be repeated multiple times. Finally, the trusted execution environment is destroyed to ensure the security and privacy of the data.

2.2 Distributed machine learning architectures

Distributed machine learning technology is to deploy huge data and computing resources to multiple machines to improve the scalability and computational efficiency of the system. It mainly refers to splitting the data set and then sending each data block to different devices for training. Each device only needs to transmit part of the data, so distributed learning can significantly reduce the communication overhead. Distributing machine learning (ML) tasks also seems to be the only way to cope with growing datasets. In distributed machine learning, multiple workers cooperate and communicate

with each other to train a model. [2] A common way to distribute learning tasks is through the now classic parameter server architecture. [3] In addition, distributed learning can also better protect data privacy because each device only needs to access part of the data instead of the whole data set.

Under a distributed machine learning system, each party needs to deploy a learning module locally and transfer the data to the local learning module. Learning modules between different data providers exchange parameters by using different encryption methods to achieve data sharing without original data sharing to protect data privacy. The computing platform helps trigger the learning module update for each participant and coordinates the relationship between all parties.

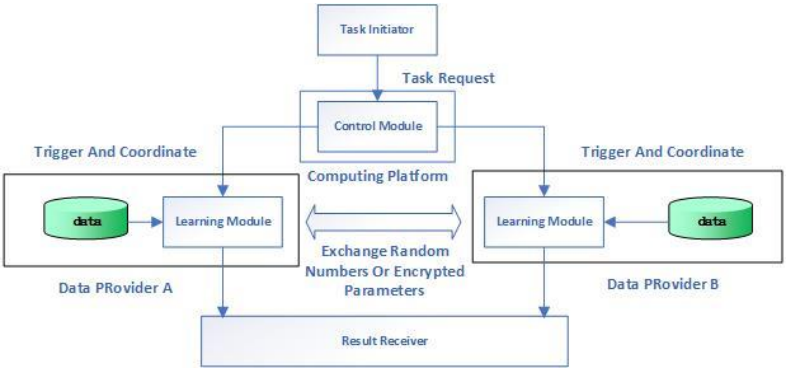

Fig. 3. Architecture of Distributed Learning System

The technical architecture of a distributed machine learning system mainly consists of a computing platform, a result receiver, a task initiator, and multiple data providers, as shown in fig.3. The computing platform mainly includes a control module, which distributes computing tasks to different data providers, and coordinates the learning modules in the data providers. Then, the learning modules in each data provider exchange random numbers or encrypted parameters to perform shared machine learning operations, while the data provider is mainly composed of data and learning modules. In this process, the multi-terminal data is placed in the virtual security domain-trusted execution environment for the calculation of the learning module, and the secret text is sent to the control module through the security protocol. At the same time, the control module of the computing platform is also in the trusted execution environment to ensure the security and reliability of the data calculation process of all parties in the end side and the cloud. Finally, the result receiver received the encrypted result from the data provider and got the final result.

Federated learning is a distributed machine learning approach in which multiple users collaboratively train a model while keeping the raw data dispersed without moving to a single server or data centre. In federated learning, raw data or data generated by secure processing based on raw data are used as training data. Federated learning only allows the transfer of intermediate data between distributed computing resources, while

avoiding the transfer of training data. Distributed computing resources refer to the mobile devices of the end users or the servers of multiple organizations.

As one of the most commonly used technical means of private computing, it technically solves the basic problems of privacy, ownership and data location. Federated learning is a distributed algorithm framework that shares intermediate statistical results without revealing the original data during the calculation process, which realizes the secure sharing of data. The privacy protection of data in multi-centre collaborative computing, so that federated learning can use multiple users to train a model while meeting the legal data restrictions, while protecting the privacy of the original data, and ensuring the accuracy and precision of the calculation results, while achieving data sharing and privacy protection.

2.3 Distributed machine Learning architectures vs. centralized machine learning architectures

Centralized learning and distributed learning are two very important learning strategies, and they are often used to solve different problems. For small data sets, centralized learning is usually preferable because it can train the model quickly and keep the model consistent, in addition, Model transmission in this architecture is smooth and elegant, and the monocentric node has a dedicated system that can be modified to suit customization requirements. [4] For scenarios with a large amount of data, distributed learning is usually a better choice because it can reduce communication overhead and protect data privacy and in some scenarios, it may be necessary to dynamically select the learning strategy to continuously optimize the performance of the model.

In the future, with the continuous development of learning technology, these two learning strategies may be further extended to better meet the requirements of data privacy and learning performance.

3 Federated Learning Models

The federated learning model is mainly used for win-win cooperation between enterprises. It has the potential to use client computing and storage resources, the enhancement of privacy at the same time, extend to large, distributed data set. [5] It requires that on the basis of protecting the data security of each enterprise, the resources of multiple parties are coordinated, so that each participant in the alliance can achieve better learning results than independent modelling. The design goal of Federated Learning is to carry out efficient machine learning among multiple participants or computing nodes on the premise of ensuring the information security during massive data exchange, protecting the privacy of terminal data. Many machine learning tasks rely on centralized learning (CL), which requires the transfer of local datasets from the client to the parameter server (PS), which incurs significant communication overhead. To overcome this, joint learning (FL) has been recognized as a promising tool [6]. According to the relationship between the data islands in the federated learning model, there are three types

of federated learning: horizontal federated learning, vertical federated learning and transfer federated learning.

3.1 Horizontal Federated Learning

Horizontal Federated Learning, also known as Sample-Partitioned Federated Learning or Example-Partitioned federated learning. In various federal study method, level of federal study (HFL) is the most studied class of isomorphism feature space. [7] It can be applied to federated learning where the data sets of each party have the same feature space and different sample Spaces.

In horizontal federated learning, user features overlap more, but users intersect less. The training data of each participant is horizontally divided, and multiple rows of samples from multiple participants with the same characteristics are combined for federated learning. The total amount of training samples is increased by horizontal federation.

3.2 Vertical Federated Learning

Also known as feature-based federated learning, it can be applied to scenarios where the datasets of each party in the federated learning have different feature Spaces and the same sample space. Vertical joint learning (VFL) describes the situation of ml private data model is based on the different separation characteristics of participation in the same case, suitable for many real-world collaboration tasks. [8] Vertical means that the data is split vertically (by columns), and the ID space of the samples is unchanged

In vertical federated learning, sample alignment is needed first, that is, to find the common samples owned by participants. Also known as entity resolution (a.k.a. entity alignment), vertical federation increases the feature latitude of the training samples.

3.3 Federated Transfer Learning

Federated transfer learning refers to the learning process of applying a model learned in the source domain to the target domain by exploiting the similarity between data, tasks, or models, is a technique used in machine learning, by moving from a related information to improve the learners in the field of a model. [9] The essence of federated transfer learning is to discover the invariance (or similarity) between the resource-rich source domain and the resource-scarce target domain, and use this invariance to transfer knowledge between the two domains.

4 Privacy Computing

Privacy Computing is a computing paradigm to protect personal privacy. It is a computing theory and method for the whole life cycle protection of private information, which uses a series of encryption technologies, security protocols and algorithms to protect data privacy. It is a series of information technologies that analyse and calculate the data under the premise that the original data is not leaked by the data provider.

In privacy computing, the process of data encryption and decryption is carried out locally, and no plaintext data is transmitted over the network, so data privacy can be effectively protected. Existing privacy protection schemes mostly focus on the relatively isolated application scenarios and technical points, for a given application scenario of problem solution is put forward. [10] Private computing can also realize data sharing, that is, data can be used jointly by multiple institutions without revealing the data source. In financial services, private computing can protect users' transaction information and assets.

5 Financial Privacy Computing Applications

5.1 Case of joint anti-telecom fraud by banks based on privacy computing

5.1.1 Case background

The provinces located in China's southwest border region and adjacent to the northern region of Myanmar with high risk of telecom fraud have become increasingly serious in recent years, and the incidence of telecom fraud has continued to be high. According to the requirements of national and local supervision, in order to effectively combat, prevent and curb new types of illegal crimes in telecom networks, build a strong defence line against telecom fraud in the field of payment and settlement, and protect the property safety and legitimate rights and interests of the people, To strengthen the anti-telecommunication fraud work of banks and improve the efficiency of account supervision, banks and China Mobile cooperated to launch a joint anti-telecommunication fraud project. The project is based on the privacy computing product of the bank. Under the condition of ensuring data security and meeting regulatory requirements, the valuable feature data of the mobile side is introduced according to the business requirements of the bank, and a big data anti-fraud modelling platform is constructed. By breaking through the industry data barriers, the closed-loop management process of joint telecom fraud risk identification was formed, and collaborative anti-fraud was carried out, which effectively improved the accuracy and timeliness of the identification and control of illegal users involved in fraud in banks.

5.1.2 Case ideas

The project conducts feature screening according to the needs of banking business departments, and then introduces external multi-channel risk feature data and labels based on privacy computing products to improve the input volume and output accuracy of the existing model. The model outputs a list of prediction results, and the business department carries out risk control tasks and post-evaluation feedback after excluding zombie users from the list.

5.1.3 Case objectives

The project is based on the privacy computing capabilities of privacy computing products, chase the data barriers between the financial and communication industries, legally and compliantly carry out cross-institutional data collaboration against fraud,

further improve the bank's ability to accurately identify the crimes of telecom fraud cases, strengthen the efficiency of the control of suspicious accounts, and effectively solve the pain points and difficulties of the bank's current anti-telecom fraud work. By jointly building a big data anti-telecom fraud modelling platform, on the basis of "data not out of the domain", data cooperative calculation and data asset capability integration are carried out safely and compliance, and federated learning joint modelling technology is applied to continuously strengthen the comprehensive identification ability of telecom and network fraud suspicious accounts and personnel. The joint modelling exploration of big data for the full traceability and tracking of the fraud process was carried out to assist the two sides to carry out a more accurate attack on telecom and network fraud.

5.1.4 Case plan

5.1.4.1 Secure intersection of privacy data, anti-fraud model output model for verification, improve the accuracy of its model output

Fig.4 illustrates the plan. The prediction results (ID numbers) of the fraud identification model and the fraud identification model are encrypted with irreversible algorithms, and the privacy computing products at the bank end are used to perform the set privacy intersection operation between the two sides. The obtained intersection customer galaxies are used to improve the recall rate and precision rate of the fraud identification model and the fraud identification model respectively.

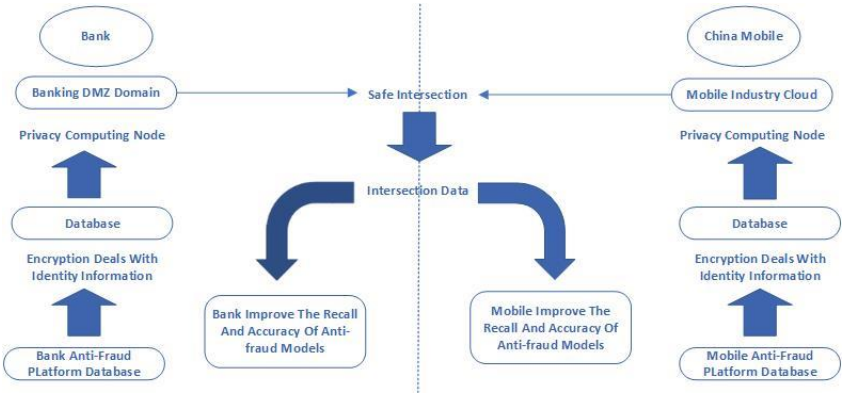

Fig. 4. Structure of the Case of joint anti-telecom fraud by banks based on privacy computing

After securely intersecting the blacklist ID (encrypted) and the full ID (encrypted) of the other party, the bank and China Mobile used the intersecting data to expand the training sample set of the model and the analysis and expansion of the blacklist user characteristic variables to retrain and optimize their respective models. The privacy of the whole process is protected, and the pre-risk identification ability is improved.

5.1.4.2 Obtain the risk list through the model, and push the list to the business department for abnormal risk investigation and related control and processing

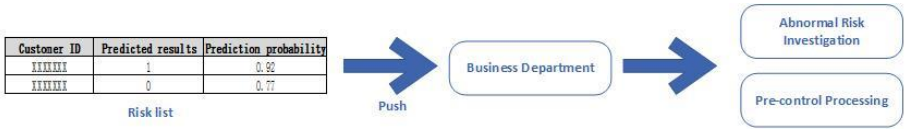

Fig. 5. Architecture of the process

The whole process is illustrated in the Fig.5 clearly.

5.1.5 Case summary

The introduction of external data based on privacy computing products solves the pain point of the single data source of the bank's current anti-fraud model. By jointly building a big data anti-fraud modelling platform with China Mobile, data cooperation and data asset capability integration can fully release the data dividend, accelerate the development and utilization of data resources, form a new paradigm of cross-industry and cross-field ecological data security cooperation, and realize ecological data secret value co-construction.

5.2 The mining project of high-value potential customers based on privacy computing in the bank

5.2.1 Case background

With the development of economy and the increase of information circulation, the demand for data interaction between various institutions is becoming more and more intense. At the same time, the social attention to personal information protection has also risen to an unprecedented height, and the regulatory requirements for personal data management have become stricter, which also requires enterprises to adopt a more secure and effective way for data fusion. At the same time, affected by the early epidemic, the customer flow of bank branches is small, the acquisition of new customers is difficult, and the loss of existing retail customers is serious. How to effectively identify high-value potential customers among existing customers has become the pain point and difficulty of banking business. In this context, the bank introduces third-party tags to realize joint analysis, joint modelling and joint prediction through the bank and China Unicom company based on privacy computing. The comprehensive evaluation index formed by Unicom based on multi-dimensional analysis is obtained to help business personnel accurately identify high-value customer groups in existing users, realize precise marketing of high-end products, improve operational efficiency, and revitalize and deeply cultivate existing customers.

5.2.2 Case thinking

By using the privacy computing product at the bank end, the bank and China Unicom company conduct joint modelling based on the valuable labels stored in their respective

environments, as shown in Fig.6. No plaintext data transmission occurs during the modelling process, only the encrypted gradient value of the features calculated during the modelling process is transmitted, and this process cannot be decrypted. Neither party can parse the encrypted information into data with actual meaning, which can fully protect the security of user information and user tags. After the iterative optimization of joint modelling is completed, the bank uses the output results to mine the existing customers, and screens out the high-end customers and potential basic customers.

5.2.3 Case objectives

The affluent label is a comprehensive affluent degree index formed by Unicom based on the analysis of its users 'residential value, asset situation and online behaviour. It can help banking institutions accurately identify high-end people in the existing users, and achieve precise marketing of high-end products for banks, which greatly improves operational efficiency and reduces marketing costs.

In the process of insight into the bank's stock users, in order to better protect user characteristics and historical business labels and other information, and avoid customer privacy data leakage, under the premise of ensuring "data available but not visible" through federated learning technology, privacy computing technology is used to cooperate with China Unicom to realize the deep and efficient completion of the user portrait of the bank's target customer group.

5.2.4 Case scheme

After the deployment of privacy computing products at the bank end, China Unicom will provide marketing user tags for joint modelling, obtain the intersection customer group through privacy intersection with the encrypted mobile phone numbers of the bank's existing customers, and realize joint analysis, joint modelling and joint prediction based on this intersection customer group, generate target customer prediction results, and complete the secondary screening according to the business demand. The high-value potential customers among them are marketed.

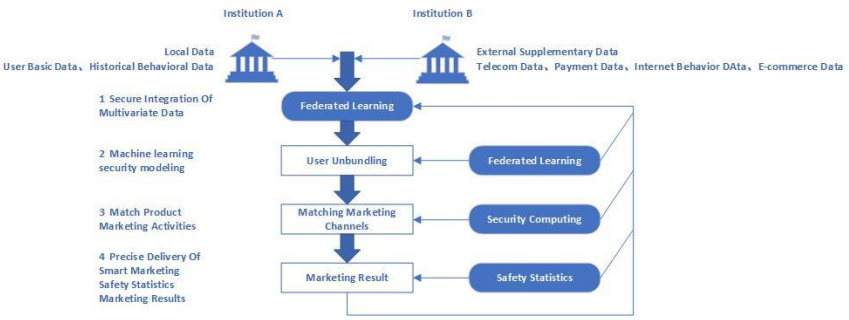

Fig. 6. Basic thinking of the second case

5.2.5 Case summary

Through privacy calculation, the bank uses the customer labels provided by Unicom to evaluate the retail customers of the bank. By deeply excavating the existing customers of the bank, the potential customers are selected from them, and the high-value customers among them are given special marketing to help marketing improve customer level management. Truly realize digital drive, promote efficiency improvement by innovation, and drive business development by innovation.

6 Conclusion

Federated Machine learning is a machine learning framework that can effectively assist multiple agencies in data usage and machine learning modelling which also promises to overcome the input privacy challenge, data security, and government regulations in machine learning. [11] The privacy computing technology based on federal machine learning expands the coverage of data services, provides large-scale secure and scalable available data for various industries and fields, makes up for the missing fragments of customer profiles that cannot be filled by relying on their own data in the past, and increases the richness of business scenarios that can be completed by internal and external cooperation. It assists various industries and fields to achieve data enrichment, scene diversification, revenue maximization, and service ecology. The application of privacy computing in the financial field solves the cost waste and security problems caused by data handling in the past, and expands the scope of high-value data, so that data partners can complete safe and compliant data cooperation and data asset capability integration, fully release the data dividend, accelerate the development and utilization of data resources, and provide safe and compliant data value sharing channels. It accelerates the development and utilization of data resources, promotes the circulation of high-value data elements, and realizes the effective management of global data assets.

References

1. Gao, C (Gao, Ce) ; Ren, R (Ren, Rui) ; Cai, HM (Cai, Hongming);C. Baier, J-P. Katoen, Principles Of Model Checking, MIT Press, 2008.(2018). GAI: A Centralized Tree-Based Scheduler For Machine Learning Workload In Large Shared Clusters. ALGORITHMS AND ARCHITECTURES FOR PARALLEL PROCESSING, ICA3PP 2018, PT Iivolume 11335 Page 611-629
2. Marozzo, F (Marozzo, Fabrizio) ; Orsino, A (Orsino, Alessio) ; Talia, D (Talia, Domenico); Trunfio, P (Trunfio, Paolo).(2022). Edge Computing Solutions For Distributed Machine Learning. 2022 IEEE Intl Conf On Dependable, Autonomic And Secure Computing, Intl Conf On Pervasive Intelligence And Computing, Intl Conf On Cloud And Big Data Computing, Intl Conf On Cyber Science And Technology Congress (Dasc/Picom/Cbdcom/Cyberscitech). Page (1148-1155)

3. El-Mhamdi, EM (El-Mhamdi, El-Mahdi); Guerraoui, R (Guerraoui, Rachid) ; Guirguis, A (Guirguis, Arsany); Hoang, LN (Hoang, Le-Nguyen); Rouault, S (Rouault, Sebastien).(2022). Genuinely Distributed Byzantine Machine Learning. DISTRIBUTED COMPUTING.Volume 35. Issue 4. Page 305-331
4. Zhang, HY (Zhang, Hongyi); Bosch, J (Bosch, Jan); Olsson, HH (Olsson, Helena Holmstrom.(2020). Federated Learning Systems: Architecture Alternatives. 2020 27TH ASIA-PACIFIC SOFTWARE ENGINEERING CONFERENCE (APSEC 2020). Page 385-394
5. Isaksson, M (Isaksson, Martin) [1] , [3] ; Zec, EL (Zec, Edvin Listo) [3] , [5] ; Coster, R (Coster, Rickard) [2] ; Gillblad, D (Gillblad, Daniel) [4] , [7] ; Girdzijauskas, S (Girdzijauskas, Sarunas.(2023). Adaptive Expert Models For Federated Learning. TRUSTWORTHY FEDERATED LEARNING, FL 2022.Volume 13448 Page 1-16
6. Elbir, AM (Elbir, Ahmet M.) ; Coleri, S (Coleri, Sinem) ; Papazafeiropoulos, AK (Papazafeiropoulos, Anastasios K.) ; Kourtessis, P (Kourtessis, Pandelis); Chatzinotas, S (Chatzinotas, Symeon).(2022). A Hybrid Architecture For Federated And Centralized Learning. IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING.Volume 8 Issue 3 Page 1529-1542
7. Mori, J (Mori, Junki) ; Teranishi, I (Teranishi, Isamu) ; Furukawa, R (Furukawa, Ryo) .(2022). Continual Horizontal Federated Learning For Heterogeneous Data. 2022 INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS (IJCNN)
8. Fu, FC (Fu, Fangcheng) ; Xue, HR (Xue, Huanran) ; Cheng, Y (Cheng, Yong) ; Tao, YY (Tao, Yangyu) ; Cui, B (Cui, Bin).(2022). BLINDFL: Vertical Federated Machine Learning Without Peeking Into Your Data. PROCEEDINGS OF THE 2022 INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA (SIGMOD '22). Page 1316-1330
9. Suzuki, J (Suzuki, Jordan) ; Lameh, SF (Lameh, Saba F.) ; Amannejad, Y (Amannejad, Yasaman).(2021). Using Transfer Learning In Building Federated Learning Models On Edge Devices. 2021 SECOND INTERNATIONAL CONFERENCE ON INTELLIGENT DATA SCIENCE TECHNOLOGIES AND APPLICATIONS (IDSTA) Page 105-113
10. Li, FH (Li, Fenghua) ; Li, H (Li, Hui) ; Niu, B (Niu, Ben) ; Chen, JJ (Chen, Jinjun).(2019). Privacy Computing: Concept, Computing Framework, And Future Development Trends. Volume 5 Issue 6 Page 1179-1192
11. Ekmefjord, M (Ekmefjord, Morgan) ; Ait-Mlouk, A (Ait-Mlouk, Addi) ; Alawadi, S (Alawadi, Sadi) ; Akesson, M (Akesson, Mattias) ; Singh, P (Singh, Prashant) ; Spjuth, O (Spjuth, Ola) ; Toor, S (Toor, Salman) ; Hellander, A (Hellander, Andreas).(2022). Scalable Federated Machine Learning With Fedn. 2022 22ND IEEE/ACM INTERNATIONAL SYMPOSIUM ON CLUSTER, CLOUD AND INTERNET COMPUTING (CCGRID 2022). Page 555-564