Vertical Federated Learning

Vertical Federated Learning: A Structured Literature Review

Afsana Khan*, Marijn ten Thij and Anna Wilbik

Abstract—Federated Learning (FL) has emerged as a promising distributed learning paradigm with an added advantage of data privacy. With the growing interest in having collaboration among data owners, FL has gained significant attention of organizations. The idea of FL is to enable collaborating participants train machine learning (ML) models on decentralized data without breaching privacy. In simpler words, federated learning is the approach of "bringing the model to the data, instead of bringing the data to the model". Federated learning, when applied to data which is partitioned vertically across participants, is able to build a complete ML model by combining local models trained only using the data with distinct features at the local sites. This architecture of FL is referred to as vertical federated learning (VFL), which differs from the conventional FL on horizontally partitioned data. As VFL is different from conventional FL, it comes with its own issues and challenges. In this paper, we present a structured literature review discussing the state-of-the-art approaches in VFL. Additionally, the literature review highlights the existing solutions to challenges in VFL and provides potential research directions in this domain.

Index Terms—Federated Learning, Vertically Partitioned Data, Privacy-Preserving Machine Learning.

1 Introduction

The expansion of machine learning (ML) has been largely driven by several factors, including the nearly infinite quantity of data that is currently accessible, the availability of reasonably priced data storage, and the development of processing that is both less expensive and more powerful. Currently, a wide variety of industries are working to develop more robust models that are able to analyze data that is larger and more complex than ever before, all while delivering results that are faster and more accurate on an unprecedented scale [1]. The use of ML has enabled organizations to more quickly identify potentially profitable opportunities as well as risks that may be involved. As more and more data becomes accessible over time, there has been a corresponding rise in interest in the application of machine learning across a variety of fields. One of the fields where machine learning has had far-reaching social effects is healthcare. The use of machine learning to analyze the health data generated by the growing number of wearable devices like smart watches and fit bits is gaining momentum [2]. Moreover, the growing use of ML in financial systems has transformed industries and societies. From traditional hedge fund management firms to FinTech service providers, many financial firms are investing in data science and ML expertise [3]. ML has also made a significant contribution to the agriculture sector by creating new opportunities to unravel, quantify, and understand data intensive processes in agricultural operational environments [4].

While organizations can benefit from applying machine learning techniques to their own data, doing the same with data from other comparable organizations could result in significant improvements to the existing organizational processes. In order to build sophisticated machine learning models for improving consumer service and acquisition, substantial emphasis has been placed on integrating data from various organizations, indicating the importance of collaboration. However, the traditional approach of bringing data located at different sites into a central server for training machine learning models is not always feasible as it raises numerous concerns. At present, sharing data among organizations has become critical due to concerns about privacy, maintaining competitive advantages, and/or other constraints. Data security and privacy are issues that are being prioritized not just by individuals or organizations but also by the larger society. The General Data Protection Regulations (GDPR), which the European Union put into place on May 25, 2018 [5] aims to protect users' personal privacy and data security. To address this issue, federated learning (FL), a new distributed learning paradigm, has recently received a lot of attention. FL allows collaboration among organizations to train machine learning models while ensuring that private data of these organizations are not disclosed [6]. Kairouz et al. in [7] formally defined federated learning as

• *The Authors are with the Department of Advanced Computing Sciences, Maastricht University, Maastricht 6229 EN, The Netherlands. E-mail: a.khan@maastrichtuniversity.nl.*

"A machine learning setting where multiple entities (clients) collaborate in solving a machine learning problem, under the coordination of a central server or service provider. Each client's raw data is stored locally and not exchanged or transferred; instead focused updates intended for immediate aggregation are used to achieve the learning objective"

Depending on how the data is partitioned or distributed among organizations, FL can be classified into three scenarios; Horizontal, Vertical and Hybrid. Horizontal federated learning (HFL) is suitable in scenarios when organizations share the same attribute space but differ in samples (Figure 1a). An example of HFL is a group of hospitals collaborating to build a ML model used to predict health risks for their patients, based on agreed-upon data. However, HFL sometimes has limited applications in practical scenarios, for example, in fostering collaboration among organizations with competing interests. Due to business reasons, it is more likely that organizations will not be willing to collaborate with their competitors [8]. On the other hand, vertical federated learning (VFL) is suitable for scenarios where organizations have the same set of samples as their data but differ in feature space (Figure 1b). VFL promotes collaboration among non-competing organizations with vertically partitioned data. In such cases, typically one organization has the ground truth, or labels, with some of the features of a number of samples. The rest of the participants take part in the federation by providing additional feature information of the same sample space but at the same time ensuring that their data is not disclosed directly to other participants. In return these participants are compensated with monetary and/or reputational rewards. Examples where VFL is applicable could be a telecom company collaborating with a home entertainment company (cable TV provider) or an airline collaborating with a car rental agency. Hybrid FL refers to the hybrid situation of horizontally and vertically partitioned data (Figure 1c). In this scenario, the data owners hold different attributes for different data instances. However, hybrid FL is not yet explored significantly in the literature.

Although VFL is a promising paradigm for privacy-preserving learning, limited research exists which explored the core challenges and methodologies of VFL due to the fact that, FL itself is a comparatively new concept. In this article, we present a structured literature review (SLR) designed to dive deeper into the critical aspects and existing methodologies of VFL as well as to pinpoint potential future directions in order to address the challenges.

2 METHODOLOGY

The goal of the following structured literature review is to investigate major challenges and existing solutions for vertical federated learning. The study not only provides an overview of the major publications in VFL, but also to identify potential gaps and opportunities for further research. The review was planned, conducted, and reported in accordance with the SLR process proposed by Armitage et al. [9]. The SLR consists of five main steps including defining research questions, designing search strategy, selecting studies, extracting data and finally synthesis of data. Below we explain these steps in more detail.

2.1 Research Questions

Taking into consideration the objective of the review, the following set of research questions were formulated as the initial stage of this SLR.

- **RQ1:** What are the existing methods in VFL and what problems do they tackle?
- **RQ2:** What are the existing applications of VFL?
- **RQ3:** What are the potential future directions for research in VFL?

2.2 Search Strategy

After the formulation of the research questions, a plan was made to design the search strategy for the SLR. The search strategy includes the initial task of selecting literature databases. Kitchenham et al. listed several high-quality databases for searching research resources [10]. Since VFL is a trending research topic, there are many articles which are pre-prints. We chose to include both published (conference and journal papers) and pre-prints as a part of the SLR. The selected databases were Google Scholar, Web of Science (WoS), IEEE Xplore and arXiv. We experimented with different search terms on the chosen databases. Finally, it was decided that the following search term would yield the most relevant results:

("Vertical federated learning") OR ("Vertical" AND ("Federated Learning" OR "privacy-preserving federated learning" OR "Heterogeneous federated learning"))

2.3 Study Selection

The results obtained using the defined search strategy were filtered based on a set of criteria. Only the articles which met the following criteria were considered further for first round screening:

- Published after 2015 since the term *"Federated Learning"* was first coined in 2016
- Written in English language
- Availability of full text
- Title and abstract specifically mention the focus on vertical federated learning

After the initial screening, the chosen articles were checked for redundancy. The unique articles were then investigated by going through the full text. Additionally, we used "snowballing" [11] technique to identify relevant articles. This was done by identifying relevant articles from the reference section of the previously selected articles. The final articles which are included in this SLR were selected based on the fact that they answered the following questions

- Did the article provide an answer to any of the research questions?
- Was the article focused on VFL and not general FL?
- Was the method proposed in the article evaluated?
- Were the experiments and results properly documented?

In addition, survey papers were also considered where vertical federated learning had been addressed.

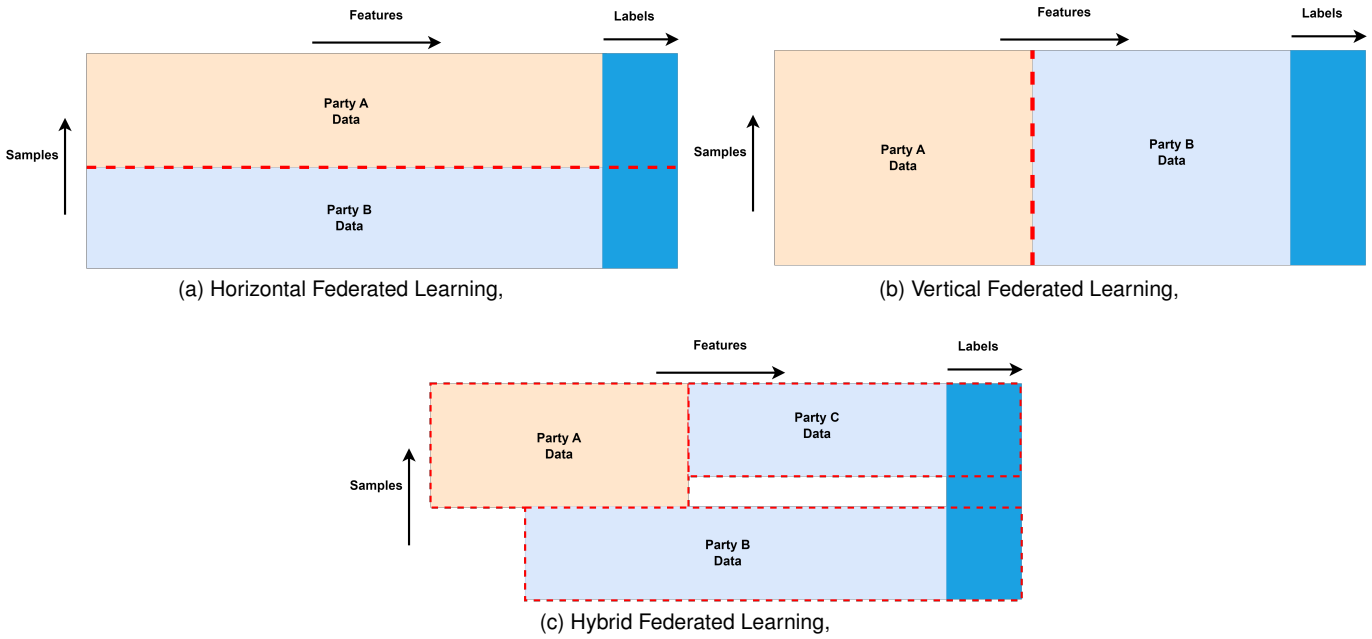
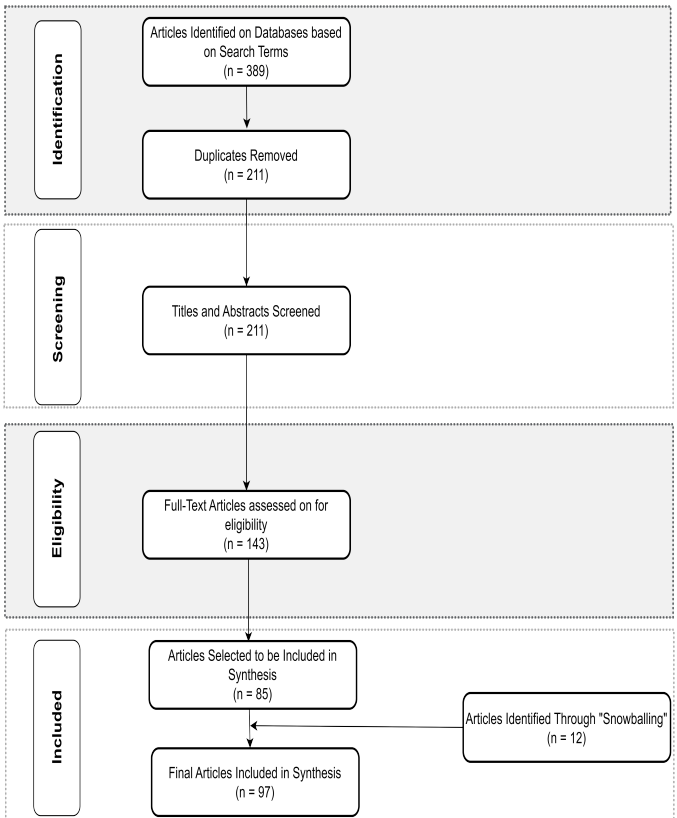


Fig. 1: Federated Learning on Partitioned Data

Search Terms	Google Scholar	WoS	IEEE Xplore	arXiv	Articles per Search Term	Unique Articles per Search Term	Total Articles	Total Unique Articles
"Vertical federated learning"	113	30	59	48	250	171		
"Vertical" AND "Federated Learning"	93	33	83	94	303	197	389	211
"Vertical" AND "privacy-preserving federated learning"	15	2	4	2	23	16		
"Vertical" AND "Heterogeneous federated learning"	5	0	0	0	5	5		

TABLE 1: Search Results



2.4 Data Extraction

We extracted data from each of the included papers and organized it in a manner such that we could provide an analysis of the reviewed literature and use them for synthesis in the next section. The data which were extracted from the articles were title & year of publication, source of publication, research question/problem solved, proposed method, availability of theoretical analysis, dataset evaluated on and model evaluated with.

2.5 Data Synthesis

As a last step of the literature review, we performed an analysis of 97 articles relevant to VFL and clustered them based on the problem those have addressed and solved. A detailed review of these articles have been provided in the further section.

3 Research Results

We conduct statistical analysis and present the results of a structured literature review by reading and analyzing articles related to FL that are found in the four major

databases mentioned earlier. These results of the study provide answers to the research questions that were presented in Section 2.1. To understand the research trend of VFL, we conducted statistics for the publication year of literature as shown in Figure 3. Although the concept of federated learning was first introduced in 2016, research on FL until 2018 primarily concentrated on horizontal federated learning. The publication of articles with a VFL focus started growing from 2019. In 2019, there were just 8 articles. But by 2021, 40 articles had been published, and 36 articles had been published in the first half of 2022. Therefore, it can be concluded that VFL is still in its early stages of development. Observing from the sources of publications, we found that

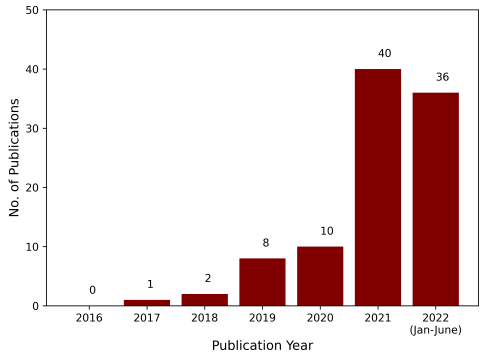


Fig. 3: Publication Year of Articles

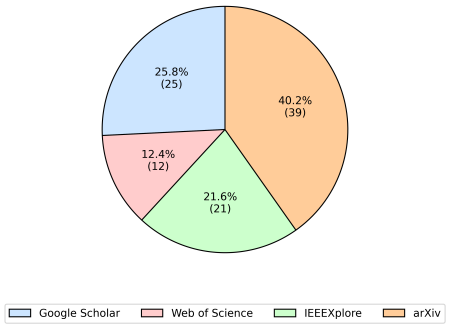


Fig. 4: Publications in Databases

about 60% of the included papers in this literature review are published work in journals and conferences which implies that VFL has produced mature results.Emerging topics frequently have a substantial number of pre-prints available in databases like arXiV. A total of 97 articles has been included in this review, of which around 40 percent are pre-prints and available on arXiv. This indicates the rapid development in VFL.

3.1 Vertical Federated Learning

In a vertical federated learning setting, the features of data points are distributed among different partitions. There are two basic architectures for VFL; with co-ordinator [12] and without co-ordinator [13]. Hardy et al. in [12] proposed a framework comprised of one trusted coordinator and two parties, each of which represents a single client. The task of the coordinator was to compute training loss as well as generate homomorphic encryption key pairs for privacy. Later on, some research works proposed [13], [14], [15] proposed a two-party architecture that eliminated the need for a trusted coordinator, thus reducing the complexity of the system. This architecture was further extended by implementing it in case of multiple collaborating parties/clients [16], [17]. Figure 5 illustrates the basic protocol for VFL with multiple parties where the party holding the labels is the active party/guest party and rest others are passive parties/host parties. The active party is responsible for computing training loss and generating key pairs to preserve privacy.

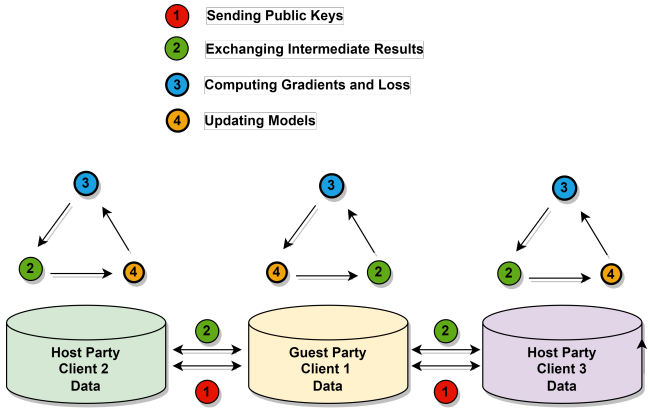


Fig. 5: Vertical Federated Learning Architecture (Adapted from [18])

The vertical federated learning problem can be defined in a more formal manner. Let $\{(x_i, y_i), i = 1, 2, .., n\}$ be a dataset where $x_i \ \epsilon \ \mathbb{R}^d$ and y_i denotes the feature vector and output labels respectively. The feature dimension is represented by d. In case of a VFL setting, the dataset is partitioned vertically across M parties/clients where each of the M parties possesses a disjoint subset of features vector $x_{[i,m]}(x_{[i,m]} \ \epsilon \ \mathbb{R}^{d_m}(i = 1, 2, .., M))$, where d_m is the feature dimension of the mth party and $\sum_{m=1}^{M} d_m = d$. Similarly, $\Theta = (\theta_1, \theta_2, .., \theta_M)$ can be defined where, $\theta_m \epsilon \mathbb{R}^{d_m}$ denotes the model parameter of the m-th party. Ideally, in VFL one of the collaborating parties is assumed to have the data labels. The party possessing label information is referred to as active party and the ones without label information as passive party. Considering the M-th party as the active party that holds the label information $y_{[i,M]}(y_{[i,M]} \ \epsilon \ \mathbb{R})$, the following function is minimized while training models in a VFL setting.

$$\mathcal{L}(\Theta) = \frac{1}{n} \sum_{i=1}^{n} l(\sum_{m=1}^{M} x_{[i,m]} \theta_i, y_{[i,M]}) + \lambda R(\Theta) \qquad (1)$$

Here $l(.)$ and $R(.)$ denote loss function and regularizer respectively while λ is the tuning parameter. VFL has been known to be used in solving regression [14] and classification [19] problems. The following algorithm [20] shows the process of solving a logistic regression problem in a VFL setting.

Algorithm 1 Vertical Federated Logistic Regression

$M \rightarrow$ Number of Clients
$T \rightarrow$ Number of Communication Rounds
$B \rightarrow$ Number of Batches
Training Data $X = \{X^1, X^2, .., X^M\}$
Client $1 \rightarrow$ Active Party/Guest Client
Client $(2..M) \rightarrow$ Passive Party/Host Client
Initialize local model ${\theta_0}^m, m \; \epsilon \; (1..M)$
for *each communication round* $t = 1, 2, \ldots, T$ **do**
  for *each batch data* ${X_b}^m \; \epsilon \; ({X_1}^m, {X_2}^m, .., {X_B}^m)$ **do**
    for *each Client* $m = 1, 2, ..M$ **do**
      Compute ${z_b}^m = X_b^m {\theta_t}^m$
      if $m \neq 1$ **then**
        send ${z_b}^m$ to Client 1
      end if
    end for
    Compute $\hat{y}_b = \sum_{m=1}^{M} {z_b}^m$ and $L(y_b, \hat{y}_b)$ on Client 1
    Compute $\frac{\partial L}{\partial z_b}$ on Client 1
    Send $\frac{\partial L}{\partial z_b^m}$ to Client $m, m\epsilon(2, k)$
    for *each Client* $m = 1, 2, ..M$ **do**
      ${\theta_t}^m \leftarrow {\theta_t}^m - \eta \frac{\partial L}{\partial z_b^m} \frac{\partial z_b^m}{\partial \theta_t^m}$
    end for
  end for
end for

VFL has also been proposed for other machine learning algorithms such as linear regression [21], [22], decision trees [23], random forests [24] and neural networks [25] etc.

Extensions for Vertical Federated Learning: There has been some existing research conducted focusing on implementing the idea of VFL in extended scenarios. For instance, current VFL systems are developed on the assumption that the labels are possessed by only one client i.e. the active party. However, there might be practical cases where multiple collaborating clients possess labels which arises the need to apply VFL in a modified manner. The Multi-VFL proposed in [26] makes use of split learning in a scenario where there are multiple data and label owners. Here, forward propagation is performed by the data owners on their corresponding partial models until the cut layer and then, their activations are sent to the label owners. These activations are concatenated by the label owners in order to complete their forward propagation. Subsequently, the losses are computed and back propagation is performed to compute the gradients. The gradients are then send back to the data owners who are supposed to use them for completing their back propagation. Moreover, Zhu et al [27] proposed a secure vertical FL framework PIVODL, to train gradient boosting decision trees (GBDTs) with data labels distributed on multiple devices. PIVODL presents a more realistic setting of training XGBoost decision tree models in VFL in which each participating client holds parts of data labels that cannot be shared and exchanged with others during the training process. A similar approach is also observed in [28] in order to deal with VFL when labels are distributed among multiple parties.

3.2 Improvements to Vertical Federated Learning

The existing literature on vertical federated learning can be categorized into four groups; communication, learning, privacy & security and business value. A brief overview of recent research on VFL in these fields have been presented in this review.

3.2.1 Communication

- **Communication Efficiency:** When following the conventional VFL approach, each of the passive/host clients share their updated gradients or intermediate results with the active/guest client during every training iteration. The total communication for each client can significantly increase over the course of hundreds or thousands of training iterations for very large data sets. As a result, the learning process might become inefficient due to communication cost and bandwidth constraints. Some existing research [29], [30], [31], [32], [33] deals with the communication overhead problem in VFL by reducing the number of local model updates during training. A Federated Stochastic Block Coordinate Descent (FedBCD) algorithm was proposed in [29] for vertically partitioned data,wherein each party performs multiple local updates before each communication in order to reduce the number of client communication rounds significantly. Furthermore, Quasi-Newton method based vertical federated learning systems [30], [34] proposed where descent steps scaled by approximate Hessian information are performed leading to faster convergence than Stochastic Gradient Descent (SGD)-based methods This allowed significant reduction in the number of communication rounds. Zhang et al. [35] proposed an algorithm to minimize the training time of VFL by adaptive selection of the number of local training updates for each party.

  Another widely used strategy to achieve communication efficiency in VFL is the application of compression schemes to the data that is being shared among the clients. [36] demonstrates efficiency of VFL improves when the data to be transmitted such as gradients of clients are compressed (using quantization or sparsification) before sharing. Based on this idea, Yang et al. [37] proposed a method of gradient sharing and compression in VFL where only the gradients greater than a certain threshold are selected and then compressed by each of the clients before sharing in order to reduce communication bandwidth. Similarly, [38] proposes an efficient vertical federated learning framework with gradient prediction and double-end sparse compression, where the compression occurs at the local models to reduce training time as well as transmission cost. Compression can also be directly applied on the local data of the host clients in a way that the compressed data contains relevant information of the local data. Later on these compressed local data are send to and aggregated by the guest client to train the model. Satisfactory outcomes to this approach has been observed in some of the research works where the compression of local data by extracting relevant information is done

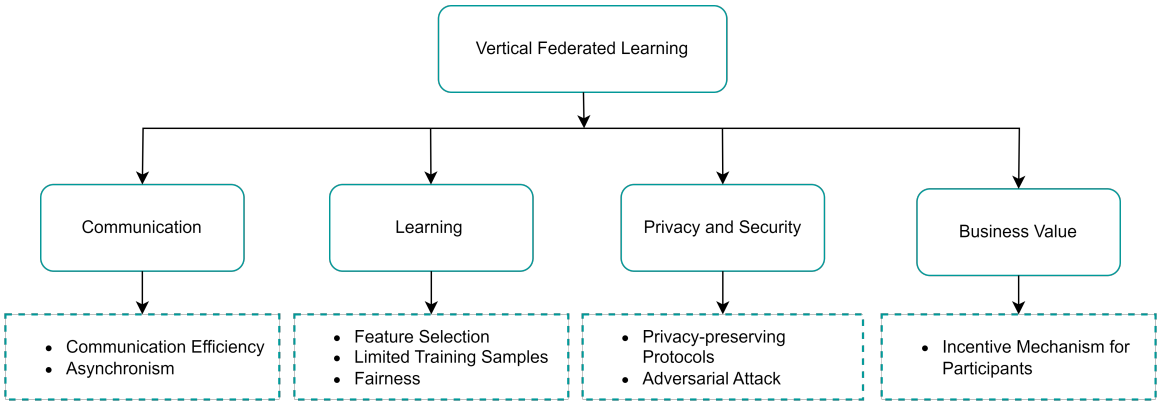


Fig. 6: Overview of Federated Learning Literature

	Article	Method	Model	Dataset
Modification in Local Updates	[29]	Stochastic Block Coordinate Descent with multiple update of local models	Logistic Regression, Neural Network	MIMIC-III, NUS-WIDE, MNIST, Default-Credit
	[30], [34]	Quasi-Newton Method	Logistic Regression	Default-Credit
	[31]	Eliminates need for peer to peer communication among clients by using functional encryption schemes	Linear regression, Logistic regression, linear SVM	Website phishing, Ionosphere, Landsat satellite, Optical recognition of handwritten digits, . MNIST
	[32]	Allowed multiple local updates in each round by using alternating direction of multipliers	Convolutional Neural Network	MNIST, CIFAR-10, NUS-WIDE, ModelNet40
	[33]	Cache enabled local updates at each client	Neural Network	$Criteo^5$, $Avazu^6$
	[35]	Adaptive selection of local updates	Logistic Regression, Neural Network	a9a, MNIST, Citeseer
Compression	[36]	Arbitrary compression scheme on gradients of local models	Neural Network	MIMIC-III, CIFAR-10, ModelNet40
	[37]	Transmission of selective gradients after compression	Logistic Regression	Default Credit
	[38]	Double-end sparse compression on local models	Logistic Regression, Neural Network	Default Credit, Insurance claim dataset
	[39]	Compression on local data using Autoencoders	Logistic Regression, SVM	Adult income, Wine-quality, Breast cancer, Rice MSC
	[40]	Compression on local data using Autoencoders	Logistic Regression	Bank loan dataset
	[41]	Compression on local data using Autoencoders	Neural Network	Adult income, Vestibular Schwannoma Dataset, The eICU Collaborative Research Database
	[42]	Compression on local data containing images using feature maps	Neural Network	CIFAR-10, CIFAR-100, CINIC-10
	[43]	Compression on local data using unsupervised representation learning	Neural Network	NUS-WIDE, MNIST

TABLE 2: Overview of Existing Methods to Reduce Communication Overhead in VFL

by using unsupervised techniques like Autoencoders [39], [40], [41], Feature Maps [42] and Representation Learning [43].

- **Asynchronism:** Vertical federated learning setting involves collaboration of multiple clients having different features of a single data instance to train a machine learning model. But it is not always practical to assume that all the clients would be identical in terms of in storage, hardware, network connectivity, etc. Due to such variability among the clients there may be cases when one or more clients aren't participating in model updates at the same time typically resulting in asynchronous updates. Thus asynchronism can pose challenges in the proper functioning of VFL which is addresses by some of the research works in [44], [45], [46]. A vertical asynchronous FL scheme was proposed [47] incorporating a backward updating mechanism and a bi-level asynchronous parallel architecture. The

two level parallel architecture: the inner level between active (available to share gradients) clients and the intra level within each client. The updates at both the levels are performed asynchronously which improves the efficiency and scalability. Moreover, [48] solved vertical FL in an asynchronous manner by allowing each client to run stochastic gradient algorithms without coordination of other clients. Thus, temporary inactivity of any client does not pose any problem in the overall training.

3.2.2 Learning

Most of the research on VFL is conducted assuming that all the participating clients possess exactly same number of samples but different features. However, in real world scenario, this assumption may not be suitable as clients may not have identical records of data. The necessity to determine the common set of samples among the clients arises but also the aspect of privacy has to be considered such that no raw data is revealed. A common technique to solve this problem is Private Set Intersection (PSI) [49], [50], [51] which is a multi-party computation cryptographic technique that allows parties, where each hold a set of elements, to compute the intersection of these elements, without revealing anything to the other party except for the elements in the intersection. Over the past few years, different PSI protocols have been proposed in [52], [53], [54], [55].

- **Limited Samples:** It is more practical to apply any of the PSI protocols before the implementation of VFL in a real world application in order to determine the common set of data records. But a crucial fact which is to be considered is that, there might not always be enough common or overlapping samples of data available among the clients. As VFL might not produce satisfactory results due to lack of sufficient data. To address this problem a solution could be to expand the training data which has to be done also in a privacy preserving manner. A data augmentation method, FedDA proposed in [56] uses generative adversarial network (GAN) to generate more overlap data by learning the features of finite overlap data and many locally existing non-overlap data among the clients. Similarly, the semi-supervised learning approach FedCVT in [57] improves the performance of the VFL model with limited aligned samples by expanding the training data through estimation of representations for missing features and predicting pseudo-labels. Some other approaches to tackle the issue with limited samples include determining inferences from non-overalpping data by using Federated Transfer Learning [58] and Oblivious Transfer [59].
- **Feature Selection:** To improve accuracy and training time of machine learning models, feature selection is a widely used strategy. It refers to the is the practice of choosing a subset of relevant features (predictors and variables) for use in a model construction. Conventional feature selection methods like Principal Component Analysis are simpler to apply in HFL setting compared to VFL. Since features are distributed across multiple clients, it becomes challenging to use the typical feature selection methods. Federated PCA approaches have been proposed [60], [61] where feature selection is achieved in a VFL setting at each client end by sharing of eigen vectors and eigen values of the host clients with the guest client. Furthermore, [62] VFL-based feature selection method that leverages deep learning models as well as complementary information from features in the same samples at multiple parties without data disclosure. Several other feature selection approaches [19], [63] have also been investigated in VFL settings which resulted in better performance when compared to traditional VFL.
- **Fairness:** Machine learning models in some cases may manifest unexpected and erratic behaviors. These behaviors when have undesirable effects on users, the model can be deemed as "unfair" based on some criteria. The existing bias in the training data is one of the key causes of a model becoming unfair. As real-world data encodes bias on sensitive features such as age, gender, and so on, VFL models may adopt bias from data and become unfair to particular user groups [64]. Again due to features being decentralized across different parties, applying existing fair ML methods to VFL models becomes challenging. Qi et al. [65] have addressed this issue by proposing a FairVFL framework where unified and fair representations of samples are learned based on the decentralized features in a privacy-preserving way. In order to obtain fair representations adversarial learning has been used to eliminate bias from the data. A superior performance in training fair VFL model was achieved [66] in which the fair learning task was modeled as a non-convex constrained optimization problem. The equivalent dual form of the optimization problem was considered and subsequently, an asynchronous gradient coordinate descent ascent algorithm was proposed to solve the dual problem.

3.2.3 Privacy and Security

- **Privacy-preserving Protocols:** Federated learning ensures privacy of data while federation among clients since training of models occurs locally and data never leaves their local sites. In vertical federated learning process, the federated model is trained by sharing of gradients or intermediate results among clients. However, some studies conclude that, there are still possibilities of sensitive private data being leaked through the local gradients in [67], and participants' data can be inferred through a generative adversarial network during the prediction stage in VFL [68]. According to recent studies on VFL, the widely used privacy-preserving protocols include homomorphic encryption (HE) is and differential privacy (DP).

  Homomorphic encryption (HE) [69] is a cryptography technique which allows specific types of computations to be carried out on ciphertexts and generates an encrypted result which, when decrypted, matches the result of operations performed on the plaintexts. For the purpose of encrupting intermediate results (e.g. gradients), VFL typically utilizes additively homomorphic encryption like Paillier [70]. Additively homomorphic encryption allows participants to

encrypt their data with a known public key and perform computation with the encrypted data by other participants with the same public key. The encrypted data needs to be sent to the private key holder so that it can be decrypted. A secure cooperative learning framework was proposed [12] for vertically partitioned data using additional HE. The framework was evaluated to be precise as the non-private solution of centralized data. Moreover, it scaled to problems with large number of samples and features. Similarly, [71] proposed a privacy-preserving DNN model training scheme based on homomorphic encryption is for vertically segmented datasets. Moreover, Several other studies [13], [30], [72] also have used HE as a privacy preserving protocol while proposing vertical federated learning approaches.

Differential privacy(DP) is a privacy-preserving protocol for bounding and quantifying the privacy leakage of sensitive data when performing learning tasks. It relies on adding noise to original data or training results to protect privacy. Too much noise can degrade the model's performance, while too less data can breach privacy. Hence, a balance between performance and privacy has to be achieved here. Wang et al. [73] designed a DP-based privacy-preserving algorithm to ensure the data confidentiality of VFL participants. The algorithm, when implemented, was quantitatively and qualitatively similar to generalized linear models, learned in an idealized non-private VFL setting. A multiparty learning framework for vertically partitioned datasets proposed in [74] achieves differential privacy of the released model by incorporating noise to the objective function. In this case, the framework requires only a single round of noise addition and secure aggregation. In addition to using DP during model training, it can also be used during the model evaluation phase in VFL since there is also possibility of leaking private label information. Sun et al. proposed two evaluation algorithms in [75] that accurately computes the widely used AUC (area under curve) metric when using label DP [76] in VFL.

- **Adversarial Attacks:** For the sake of computing efficiency, most VFL protocols choose to overlook data security and user privacy. This tends to make joint model training more vulnerable to the adversary, resulting in the leakage of private data. Because no raw data is shared between the two parties, VFL initially appears to be private. At the end of a passive party, a considerable amount of information still exists in the cut layer embedding that can be used by the active party to leak raw data. In addition, the gradient update mechanism of VFL can also be exploited by a malicious participant to gain the power to infer the privately owned labels [77]. Sun et al. designed an adversarial training based framework for VFL which simulates the game between an attacker (i.e. the active party) who actively reconstructs raw input from the cut layer embedding and a defender (i.e. the passive party) who aims to prevent the input leakage. A similar approach is observed in [68] which deals with malicious attack by active party during model evaluation stage of VFL.

3.2.4 Business Value

The concept of fairness in FL also includes collaboration fairness which implies the incentive mechanism in a FL setting. Since FL is based on collaboration, rewarding mechanism is crucial for the allocating rewards to current and potential participants of FL. To achieve that, FL needs a fair evaluation mechanism to give agents reasonable rewards. Moreover, to analyse business value of VFL in real world application determing a proper incentive mechanism for the participating clients is crucial since not doing do may result in lack of motivation from the clients to collaborate. The Shapley value (SV) [78] is a provably fair contribution valuation metric originated from cooperative game theory. The contribution of participants can be measured by Shapley Values to calculate grouped feature importance [79]. This idea can also be extended to both synchronous and asynchronous vertical federated algorithms [80].

3.3 VFL Applications

Google initiated project in 2016 in order to establish federated learning among Android mobile users [81]. The goal was to improve the keyboard input prediction quality, while at the same time ensure the security and privacy of users. In the later stage, many use cases were addressed in several surveys and studies where FL could be implemented. There has been significant efforts in designing federated learning frameworks for industrial usage but most of them are still in their development stage. Among the existing FL frameworks, only few of them basically support VFL either completely or partially. Table 3 provides an overview of the existing VFL frameworks. It can be observed from the overview that, most listed frameworks (FedML [82], FedLearner, FederatedScope) do not support a variety of ML models for implementation. Besides, few of the frameworks don't have complete privacy features incorporated and don't possess complete documentations. Thus, it can be concluded that VFL frameworks are still in its developing stage which indicates a lot of scope to work on.

Federated learning is a cutting-edge modeling mechanism which is capable of training a unified machine learning model using decentralized data while maintaining privacy and security. Thus, it has a promising application in financial, healthcare and many other industries where direct aggregation of data for training models is not feasible due to concerns like data security, privacy protection and intellectual property. Most applications are focused on horizontal federated learning and it is quite difficult to observe VFL being used in real-world applications yet. However, in our literature review, we found a limited number of studies proposing and as well as implementing VFL in applications. [84] considered a scenario involving two hospitals; Inner and Outer hospital possessing records of daily performance and clinical test of patients. Due to patients privacy regulations, it was not allowed to share raw data among the hospitals even though they had records of the same patients. Hence, the Inner- and Outer-hospital information had been

Framework	FATE	FedML	PaddleFL	FedLearner	FederatedScope	Crypten	FedTree
Model Support							
Regression	✓	✓	✓	-	✓	✓	-
Neural Network	✓	-	✓	✓	-	✓	-
Tree-Based Model	✓	-	-	✓	-	-	✓
Third Party Collaborator Requirement	✓	-	✓	✓	-	✓	✓
Privacy							
Complete Privacy of Model Parameters	✓	-	✓	-	-	✓	✓
Complete Privacy of Model Gradients	✓	✓	✓	-	✓	✓	✓
Availability of Complete Documentation	✓	-	-	-	✓	✓	✓

TABLE 3: Overview of Federated Learning Frameworks Supporting Vertical Partitioning (Adjusted from [83])

bridged via vertical Federated Learning for perioperative complications prognostic prediction. In [85] a VFL scheme is developed for the purpose of human activity recognition (HAR) across a variety of different devices from multiple individual users by integrating shareable features from heterogeneous data across different devices into a full feature space. VFL has also been useful in the field of e-commerce as observed in [86] where a method based on clustering and latent factor model under the vertical federated recommendation system was implemented. Taking into account the diversity of a large number of different users in each participant and the complexity of the matrix factorization of the user-item matrix, the users were clustered to reduce the dimension of the matrix and improve the accuracy of user recommendations. Similarly, efficient online advertising was achieved through the application of VFL [87]. Vertical federated learning when applied to financial institutions boosted their profits by collaboration of data among them maintaining privacy [88], [89]. A special use case of VFL was observed in the aviation domain [90] where a flight delay prediction model based on federated learning was designed by integrating horizontal and vertical federated frameworks.

4 Open Challenges and Future Directions

The structured literature review has provided insights to existing approaches adapted to improve different aspects of VFL while also pointing out potential research directions. The research directions concluded from the review are discussed as follows:

- **Communication Efficiency:** The total communication and computation cost in VFL is proportional to the size of the training data. With the growing amount of data accross multiple platforms, VFL becomes more challenging due to significant increase in local model computations, updates and as well as communication cost. To tackle this problem several sufficient studies discussed earlier have been computation and communication efficient methods that reduce the communication and computation complexity. But then again data in massively increasing day by day, on the other hand computing resources of participants are not increasing at the same rate. Hence, making VFL more efficient by ensuring low communication rounds and computation cost will still remain a challenge.

- **Asynchronism:** The participating clients in VFL may vary in their resources like storage, network or power which can often result in transmission delay of gradients. In HFL scenarios it does not cause much problem but in case of VFL where no client has a complete features set of a data instance, significant performance degradation can be observed. Under asynchronous settings, conventional VFL algorithms cannot always ensure convergence because only active parties are able to update the gradient of the loss function, whilst passive parties cannot due to absence of labels, resulting in partial model parameters that are not optimized during the training process. Although there have been several works studying asynchronous VFL algorithms, it is still an open problem to design asynchronous algorithms for solving real-world VFL tasks.

- **Privacy and Security:** Privacy and Security has always been a concern in FL since the assurance of the no exposure of raw data is a major motivation for data owners to participate in the federation. Researchers in their studies so far have rigorously tried to identify potential privacy leakage and malicious attacks in FL setup. However, the aspect of security in VFL is not enough explored as most privacy preserving protocols like homomorphic encryption, secret sharing and differential privacy are used in HFL scenarios. There is still scope for improvements in dealing with privacy leakage and backdoor attacks in VFL.

- **Limited Training Data:** As a preprocessing step for VFL, the overlapping samples among the clients are determined. Often the overlapping samples are insufficient for achieving a good performance on VFL models. Expanding the training data is a solution but then again that also comes with privacy constraints since local raw data cannot be disclosed. Again not utilizing the non-overlapping samples among the participants would mean making the data useless.

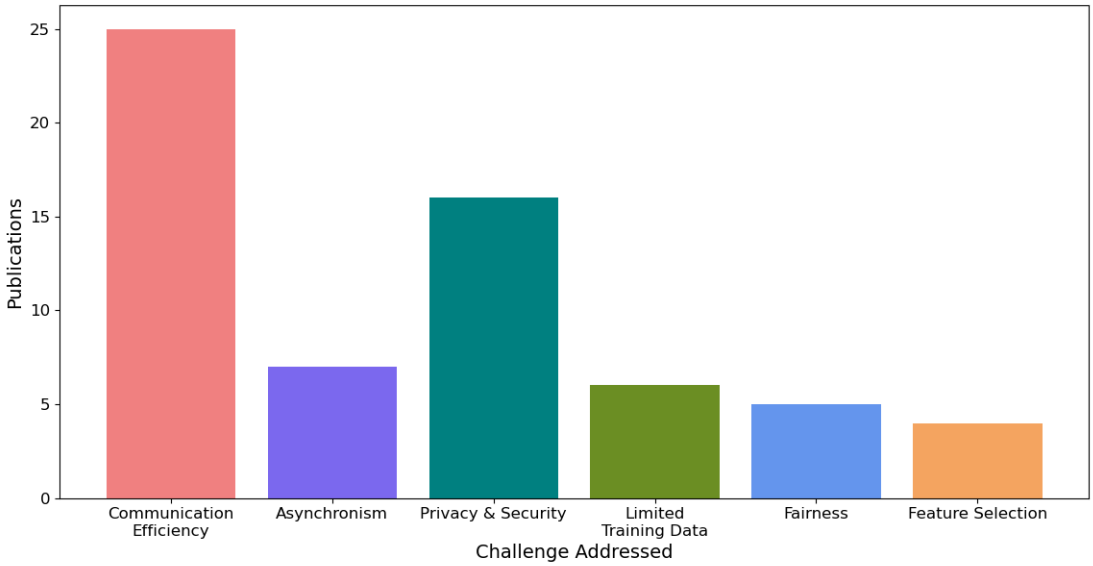


Fig. 7: VFL Challenges Addressed in the Selected Studies

Since data is expensive and difficult to obtain, new approaches could be designed such that relevant information is inferred from the non-overlapping data as well.

- **Feature Selection:** Feature selection, which has been a topic of research in both methodology and practice for decades, is used in a variety of different fields like text mining, image recognition, fault diagnosis, intrusion detection, and so on. VFL has promising potential in many feature selection applications with privacy preservation. Feature selection combined with VFL would allow business from different organizations to collaboratively perform feature selection without exposing private data. Privacy-preserving feature selection in VFL has not been fully explored yet, although only a few solutions were presented (Section 3.2.2). Hence, designing efficient and effective privacy-preserving feature selection protocols for VFL could be an interesting direction for future research.

- **Model Fairness:** There has been rising interest in developing fair methods for machine learning [91]. In practical VFL setups, the aspects like data being distributed across multiple platforms and asynchronous parallelized updates makes model fairness even more challenging to enhance. However, such concerns have been less addressed in vertical federated learning settings. The application of fair VFL methods for ensuring bias free model training is an open opportunity for future vertical federated learning research. It is particularly important as VFL has potential to be implemented in applications involving real populations of users without knowledge of their sensitive identities.

- **Incentive Mechanism:** Motivating data owners to participate in a data federation is a significant challenge in federated learning. Even in vertically federated setups, it is essential to encourage more qualified clients to participate. In order to achieve this goal, incentive schemes are to be designed which are able to fairly share the profits generated due to the collaboration among participants. This cannot be done without first establishing a mechanism for assessing each data owner's contribution to the federated model. Despite the fact that several works have focused on the design of incentive mechanisms for vertical federated learning, one crucial aspect, i.e. security, has been overlooked. Even while evaluating contributions of participants, it is important that privacy is preserved from all ends by ensuring no exposure of data. Improvised incentive mechanism could be designed which not only deals with distribution of profit but also penalizes the participants in case erroneous data is provided by them misleading the federated learning process.

- **Explainability:** An area of study that has gained a great deal of attention recently is explainable artificial intelligence (XAI). Models must be explainable in high stake applications like healthcare, when there is a strong need to justify decisions made. The same applies in case of VFL as well. VFL models must be explainable, especially when dealing with sensitive data. In VFL models, each party's data is kept private and is not accessible to third parties for analysis. For the deployment of VFL, it is crucial to interpret models while making sure the data is stored locally only. Explainability of VFL models may also be attributed to potential privacy violations from unintended data leaks. Thus, there is a potential scope for future research for explainable AI in context of VFL where a trade-off between explainability and

privacy is achieved.

5 Conclusion

With the advancement of big data and artificial intelligence, the expectations of privacy are becoming increasingly stringent. As a result, federated learning, a novel solution to cross-platform privacy protection, was developed. Apart from privacy, FL now needs to deal with a number of other challenges when applied to data partitioned in various formats. While most studies focus on FL on horizontally partitioned data, this review aims to provide researchers in FL domain with the current state-of-the-art in vertical federated learning. We not only mentioned the challenges observed in VFL in our study, but also clustered them into four categories; communication, learning, privacy & security and business value. The study also discussed the approaches adopted by earlier studies to overcome these challenges and looked into the applicability of VFL in real-world scenarios. Finally, a set of eight prospective future directions for research in this domain have been identified.

References

[1] Iqbal H Sarker. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3):1–21, 2021.

[2] Rohan Bhardwaj, Ankita R Nambiar, and Debojyoti Dutta. A study of machine learning in healthcare. In *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, volume 2, pages 236–241. IEEE, 2017.

[3] John W Goodell, Satish Kumar, Weng Marc Lim, and Debidutta Pattnaik. Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *Journal of Behavioral and Experimental Finance*, 32:100577, 2021.

[4] Konstantinos G Liakos, Patrizia Busato, Dimitrios Moshou, Simon Pearson, and Dionysis Bochtis. Machine learning in agriculture: A review. *Sensors*, 18(8):2674, 2018.

[5] Jan Philipp Albrecht. How the gdpr will change the world. *Eur. Data Prot. L. Rev.*, 2:287, 2016.

[6] Chuan Ma, Jun Li, Ming Ding, Howard H Yang, Feng Shu, Tony QS Quek, and H Vincent Poor. On safeguarding privacy and security in the framework of federated learning. *IEEE network*, 34(4):242–248, 2020.

[7] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[8] Yong Cheng, Yang Liu, Tianjian Chen, and Qiang Yang. Federated learning for privacy-preserving ai. *Communications of the ACM*, 63(12):33–36, 2020.

[9] Andrew Armitage and Diane Keeble-Allen. Undertaking a structured literature review or structuring a literature review: Tales from the field. In *Proceedings of the 7th European Conference on Research Methodology for Business and Management Studies: ECRM2008, Regent's College, London*, page 35, 2008.

[10] Barbara A Kitchenham. Systematic review in software engineering: where we are and where we should be going. In *Proceedings of the 2nd international workshop on Evidential assessment of software technologies*, pages 1–2, 2012.

[11] Rosemarie Streeton, Mary Cooke, and Jackie Campbell. Researching the researchers: using a snowballing technique. *Nurse researcher*, 12(1):35–47, 2004.

[12] Stephen Hardy, Wilko Henecka, Hamish Ivey-Law, Richard Nock, Giorgio Patrini, Guillaume Smith, and Brian Thorne. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. *arXiv preprint arXiv:1711.10677*, 2017.

[13] Shengwen Yang, Bing Ren, Xuhui Zhou, and Liping Liu. Parallel distributed logistic regression for vertical federated learning without third-party coordinator. *arXiv preprint arXiv:1911.09824*, 2019.

[14] Daojing He, Runmeng Du, Shanshan Zhu, Min Zhang, Kaitai Liang, and Sammy Chan. Secure logistic regression for vertical federated learning. *IEEE Internet Computing*, 26(2):61–68, 2021.

[15] Huizhong Sun, Zhenya Wang, Yuejia Huang, and Junda Ye. Privacy-preserving vertical federated logistic regression without trusted third-party coordinator. In *2022 The 6th International Conference on Machine Learning and Soft Computing*, pages 132–138, 2022.

[16] Kewei Cheng, Tao Fan, Yilun Jin, Yang Liu, Tianjian Chen, Dimitrios Papadopoulos, and Qiang Yang. Secureboost: A lossless federated learning framework. *IEEE Intelligent Systems*, 36(6):87–98, 2021.

[17] Di Zhao, Ming Yao, Wanwan Wang, Hao He, and Xin Jin. Ntp-vfl-a new scheme for non-3rd party vertical federated learning. In *2022 14th International Conference on Machine Learning and Computing (ICMLC)*, pages 134–139, 2022.

[18] Qiang Yang, Yang Liu, Yong Cheng, Yan Kang, Tianjian Chen, and Han Yu. Federated learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 13(3):1–207, 2019.

[19] Siwei Feng and Han Yu. Multi-participant multi-class vertical federated learning. *arXiv preprint arXiv:2001.11154*, 2020.

[20] Hangyu Zhu, Haoyu Zhang, and Yaochu Jin. From federated learning to federated neural architecture search: a survey. *Complex & Intelligent Systems*, 7(2):639–657, 2021.

[21] Adrià Gascón, Phillipp Schoppmann, Borja Balle, Mariana Raykova, Jack Doerner, Samee Zahur, and David Evans. Privacy-preserving distributed linear regression on high-dimensional data. *Proc. Priv. Enhancing Technol.*, 2017(4):345–364, 2017.

[22] Hiroaki Kikuchi, Chika Hamanaga, Hideo Yasunaga, Hiroki Matsui, Hideki Hashimoto, and Chun-I Fan. Privacy-preserving multiple linear regression of vertically partitioned real medical datasets. *Journal of Information Processing*, 26:638–647, 2018.

[23] Fatemeh Khodaparast, Mina Sheikhalishahi, Hassan Haghighi, and Fabio Martinelli. Privacy preserving random decision tree classification over horizontally and vertically partitioned data. In *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*, pages 600–607. IEEE, 2018.

[24] Yang Liu, Yingting Liu, Zhijie Liu, Yuxuan Liang, Chuishi Meng, Junbo Zhang, and Yu Zheng. Federated forest. *IEEE Transactions on Big Data*, 2020.

[25] Jun Zhou, Chaochao Chen, Longfei Zheng, Huiwen Wu, Jia Wu, Xiaolin Zheng, Bingzhe Wu, Ziqi Liu, and Li Wang. Vertically federated graph neural network for privacy-preserving node classification. *arXiv preprint arXiv:2005.11903*, 2020.

[26] Vaikkunth Mugunthan, Pawan Goyal, and Lalana Kagal. Multi-vfl: A vertical federated learning system for multiple data and label owners. *arXiv preprint arXiv:2106.05468*, 2021.

[27] Hangyu Zhu, Rui Wang, Yaochu Jin, and Kaitai Liang. Pivodl: Privacy-preserving vertical federated learning over distributed labels. *IEEE Transactions on Artificial Intelligence*, 2021.

[28] Rui Wang, Oğuzhan Ersoy, Hangyu Zhu, Yaochu Jin, and Kaitai Liang. Feverless: Fast and secure vertical federated learning based on xgboost for decentralized labels. 2021.

[29] Yang Liu, Xinwei Zhang, Yan Kang, Liping Li, Tianjian Chen, Mingyi Hong, and Qiang Yang. Fedbcd: A communication-efficient collaborative learning framework for distributed features. *IEEE Transactions on Signal Processing*, 2022.

[30] Kai Yang, Tao Fan, Tianjian Chen, Yuanming Shi, and Qiang Yang. A quasi-newton method based vertical federated learning framework for logistic regression. *arXiv preprint arXiv:1912.00513*, 2019.

[31] Runhua Xu, Nathalie Baracaldo, Yi Zhou, Ali Anwar, James Joshi, and Heiko Ludwig. Fedv: Privacy-preserving federated learning over vertically partitioned data. In *Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security*, pages 181–192, 2021.

[32] Chulin Xie, Pin-Yu Chen, Ce Zhang, and Bo Li. Improving privacy-preserving vertical federated learning by efficient communication with admm. *arXiv preprint arXiv:2207.10226*, 2022.

[33] Fangcheng Fu, Xupeng Miao, Jiawei Jiang, Huanran Xue, and Bin Cui. Towards communication-efficient vertical federated learning

training via cache-enabled local updates. *Proceedings of the VLDB Endowment*, 15(10):2111–2120, 2022.

[34] Song WenJie and Shen Xuan. Vertical federated learning based on dfp and bfgs. *arXiv preprint arXiv:2101.09428*, 2021.

[35] Jie Zhang, Song Guo, Zhihao Qu, Deze Zeng, Haozhao Wang, Qifeng Liu, and Albert Y Zomaya. Adaptive vertical federated learning on unbalanced features. *IEEE Transactions on Parallel and Distributed Systems*, 33(12):4006–4018, 2022.

[36] Timothy J Castiglia, Anirban Das, Shiqiang Wang, and Stacy Patterson. Compressed-vfl: Communication-efficient learning with vertically partitioned data. In *International Conference on Machine Learning*, pages 2738–2766. PMLR, 2022.

[37] Kuihe Yang, Ziying Song, Yingchao Zhang, Yufan Zhou, Xiaohan Sun, and Jianxuan Wang. Model optimization method based on vertical federated learning. In *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE, 2021.

[38] Ming Li, Yiwei Chen, Yiqin Wang, and Yu Pan. Efficient asynchronous vertical federated learning via gradient prediction and double-end sparse compression. In *2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pages 291–296. IEEE, 2020.

[39] Afsana Khan, Marijn ten Thij, and Anna Wilbik. Communication-efficient vertical federated learning. *Algorithms*, 15(8):273, 2022.

[40] Pratik Ratadiya, Khushi Asawa, and Omkar Nikhal. A decentralized aggregation mechanism for training deep learning models using smart contract system for bank loan prediction. *arXiv preprint arXiv:2011.10981*, 2020.

[41] Dongchul Cha, MinDong Sung, Yu-Rang Park, et al. Implementing vertical federated learning using autoencoders: Practical application, generalizability, and utility study. *JMIR medical informatics*, 9(6):e26598, 2021.

[42] Tianchi Sha, Xiao Yu, Zhiwei Shi, Yuan Xue, Shouxin Wang, and Sikang Hu. Feature map transfer: Vertical federated learning for cnn models. In *International Conference on Data Mining and Big Data*, pages 37–44. Springer, 2021.

[43] Zhaomin Wu, Qinbin Li, and Bingsheng He. Practical vertical federated learning with unsupervised representation learning. *IEEE Transactions on Big Data*, 2022.

[44] Qingsong Zhang, Bin Gu, Cheng Deng, Songxiang Gu, Liefeng Bo, Jian Pei, and Heng Huang. Asysqn: Faster vertical federated learning algorithms with better computation resource utilization. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3917–3927, 2021.

[45] Bin Gu, An Xu, Zhouyuan Huo, Cheng Deng, and Heng Huang. Privacy-preserving asynchronous vertical federated learning algorithms for multiparty collaborative learning. *IEEE transactions on neural networks and learning systems*, 2021.

[46] Qianjun Wei, Qiang Li, Zhipeng Zhou, ZhengQiang Ge, and Yonggang Zhang. Privacy-preserving two-parties logistic regression on vertically partitioned data using asynchronous gradient sharing. *Peer-to-Peer Networking and Applications*, 14(3):1379–1387, 2021.

[47] Qingsong Zhang, Bin Gu, Cheng Deng, and Heng Huang. Secure bilevel asynchronous vertical federated learning with backward updating. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10896–10904, 2021.

[48] Tianyi Chen, Xiao Jin, Yuejiao Sun, and Wotao Yin. Vafl: a method of vertical asynchronous federated learning. *arXiv preprint arXiv:2007.06081*, 2020.

[49] Michael J Freedman, Kobbi Nissim, and Benny Pinkas. Efficient private matching and set intersection. In *International conference on the theory and applications of cryptographic techniques*, pages 1–19. Springer, 2004.

[50] Yan Huang, David Evans, and Jonathan Katz. Private set intersection: Are garbled circuits better than custom protocols? In *NDSS*, 2012.

[51] Emiliano De Cristofaro and Gene Tsudik. Practical private set intersection protocols with linear complexity. In *International Conference on Financial Cryptography and Data Security*, pages 143–159. Springer, 2010.

[52] Prasad Buddhavarapu, Andrew Knox, Payman Mohassel, Shubho Sengupta, Erik Taubeneck, and Vlad Vlaskin. Private matching for compute. *Cryptology ePrint Archive*, 2020.

[53] Mihaela Ion, Ben Kreuter, Ahmet Erhan Nergiz, Sarvar Patel, Shobhit Saxena, Karn Seth, Mariana Raykova, David Shanahan, and Moti Yung. On deploying secure computing: Private intersection-sum-with-cardinality. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 370–389. IEEE, 2020.

[54] Melissa Chase and Peihan Miao. Private set intersection in the internet setting from lightweight oblivious prf. In *Annual International Cryptology Conference*, pages 34–63. Springer, 2020.

[55] Linpeng Lu and Ning Ding. Multi-party private set intersection in vertical federated learning. In *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 707–714. IEEE, 2020.

[56] JianFei Zhang and YuChen Jiang. A data augmentation method for vertical federated learning. *Wireless Communications and Mobile Computing*, 2022, 2022.

[57] Yan Kang, Yang Liu, and Xinle Liang. Fedcvt: Semi-supervised vertical federated learning with cross-view training. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(4):1–16, 2022.

[58] Dashan Gao, Yang Liu, Anbu Huang, Ce Ju, Han Yu, and Qiang Yang. Privacy-preserving heterogeneous federated transfer learning. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 2552–2559. IEEE, 2019.

[59] Zhenghang Ren, Liu Yang, and Kai Chen. Improving availability of vertical federated learning: Relaxing inference on non-overlapping data. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2022.

[60] Yiu-ming Cheung, Jian Lou, and Feng Yu. Vertical federated principal component analysis on feature-wise distributed data. In *International Conference on Web Information Systems Engineering*, pages 173–188. Springer, 2021.

[61] Yiu-ming Cheung, Juyong Jiang, Feng Yu, and Jian Lou. Vertical federated principal component analysis and its kernel extension on feature-wise distributed data. *arXiv preprint arXiv:2203.01752*, 2022.

[62] Siwei Feng. Vertical federated learning-based feature selection with non-overlapping sample utilization. *Expert Systems with Applications*, 208:118097, 2022.

[63] Rui Zhang, Hongwei Li, Meng Hao, Hanxiao Chen, and Yuan Zhang. Secure feature selection for vertical federated learning in ehealth systems. In *ICC 2022-IEEE International Conference on Communications*, pages 1257–1262. IEEE, 2022.

[64] Chuhan Wu, Fangzhao Wu, Xiting Wang, Yongfeng Huang, and Xing Xie. Fairness-aware news recommendation with decomposed adversarial learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4462–4469, 2021.

[65] Tao Qi, Fangzhao Wu, Chuhan Wu, Lingjuan Lyu, Tong Xu, Zhongliang Yang, Yongfeng Huang, and Xing Xie. Fairvfl: A fair vertical federated learning framework with contrastive adversarial learning. *arXiv preprint arXiv:2206.03200*, 2022.

[66] Changxin Liu, Zirui Zhou, Yang Shi, Jian Pei, Lingyang Chu, and Yong Zhang. Achieving model fairness in vertical federated learning. *arXiv preprint arXiv:2109.08344*, 2021.

[67] Yoshinori Aono, Takuya Hayashi, Lihua Wang, Shiho Moriai, et al. Privacy-preserving deep learning via additively homomorphic encryption. *IEEE Transactions on Information Forensics and Security*, 13(5):1333–1345, 2017.

[68] Xinjian Luo, Yuncheng Wu, Xiaokui Xiao, and Beng Chin Ooi. Feature inference attack on model predictions in vertical federated learning. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pages 181–192. IEEE, 2021.

[69] Xun Yi, Russell Paulet, and Elisa Bertino. Homomorphic encryption. In *Homomorphic encryption and applications*, pages 27–46. Springer, 2014.

[70] Pascal Paillier. Public-key cryptosystems based on composite degree residuosity classes. In *International conference on the theory and applications of cryptographic techniques*, pages 223–238. Springer, 1999.

[71] Mingjun Dai, Annan Xu, Qingwen Huang, Zhonghao Zhang, and Xiaohui Lin. Vertical federated dnn training. *Physical Communication*, 49:101465, 2021.

[72] Wei Ou, Jianhuan Zeng, Zijun Guo, Wanqin Yan, Dingwan Liu, and Stelios Fuentes. A homomorphic-encryption-based vertical federated learning scheme for rick management. *Computer Science and Information Systems*, 17(3):819–834, 2020.

[73] Chang Wang, Jian Liang, Mingkai Huang, Bing Bai, Kun Bai, and Hao Li. Hybrid differentially private federated learning on vertically partitioned data. *arXiv preprint arXiv:2009.02763*, 2020.

[74] Depeng Xu, Shuhan Yuan, and Xintao Wu. Achieving differential privacy in vertically partitioned multiparty learning. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 5474–5483. IEEE, 2021.

[75] Jiankai Sun, Xin Yang, Yuanshun Yao, Junyuan Xie, Di Wu, and Chong Wang. Differentially private auc computation in vertical federated learning. *arXiv preprint arXiv:2205.12412*, 2022.

[76] Badih Ghazi, Noah Golowich, Ravi Kumar, Pasin Manurangsi, and Chiyuan Zhang. Deep learning with label differential privacy. *Advances in Neural Information Processing Systems*, 34:27131–27145, 2021.

[77] Chong Fu, Xuhong Zhang, Shouling Ji, Jinyin Chen, Jingzheng Wu, Shanqing Guo, Jun Zhou, Alex X Liu, and Ting Wang. Label inference attacks against vertical federated learning.

[78] Alvin E Roth. *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press, 1988.

[79] Guan Wang, Charlie Xiaoqian Dang, and Ziye Zhou. Measure contribution of participants in federated learning. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 2597–2604. IEEE, 2019.

[80] Zhenan Fan, Huang Fang, Zirui Zhou, Jian Pei, Michael P Friedlander, and Yong Zhang. Fair and efficient contribution valuation for vertical federated learning. *arXiv preprint arXiv:2201.02658*, 2022.

[81] Keith Bonawitz, 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*, pages 1175–1191, 2017.

[82] Chaoyang He, Songze Li, Jinhyun So, Xiao Zeng, Mi Zhang, Hongyi Wang, Xiaoyang Wang, Praneeth Vepakomma, Abhishek Singh, Hang Qiu, et al. Fedml: A research library and benchmark for federated machine learning. *arXiv preprint arXiv:2007.13518*, 2020.

[83] Xiaoyuan Liu, Tianneng Shi, Chulin Xie, Qinbin Li, Kangping Hu, Haoyu Kim, Xiaojun Xu, Bo Li, and Dawn Song. Unifed: A benchmark for federated learning frameworks. *arXiv preprint arXiv:2207.10308*, 2022.

[84] Weihao Sun, Yiqiang Chen, Xiaodong Yang, Jiangbei Cao, and Yuxiang Song. Fedio: Bridge inner-and outer-hospital information for perioperative complications prognostic prediction via federated learning. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3215–3221. IEEE, 2021.

[85] Xiaokang Zhou, Wei Liang, Jianhua Ma, Zheng Yan, I Kevin, and Kai Wang. 2d federated learning for personalized human activity recognition in cyber-physical-social systems. *IEEE Transactions on Network Science and Engineering*, 2022.

[86] JianFei Zhang and YuChen Jiang. A vertical federation recommendation method based on clustering and latent factor model. In *2021 International Conference on Electronic Information Engineering and Computer Science (EIECS)*, pages 362–366. IEEE, 2021.

[87] Wenjie Li, Qiaolin Xia, Hao Cheng, Kouyin Xue, and Shu-Tao Xia. Vertical semi-federated learning for efficient online advertising. *arXiv preprint arXiv:2209.15635*, 2022.

[88] Yuxin Liang, Zhiyong Liu, Yong Song, Aidong Yang, Xiaozhou Ye, and Ye Ouyang. A methodology of trusted data sharing across telecom and finance sector under china's data security policy. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 5406–5412. IEEE, 2021.

[89] Yusuf Efe. A vertical federated learning method for multi-institutional credit scoring: Mics. *arXiv preprint arXiv:2111.09038*, 2021.

[90] Li Guo, Qin Wei, and He Chenyu. Research on flight delay prediction based on horizontal and vertical federated learning framework. In *2021 IEEE 3rd International Conference on Civil Aviation Safety and Information Technology (ICCASIT)*, pages 38–44. IEEE, 2021.

[91] Seyma Yucer, Samet Akçay, Noura Al-Moubayed, and Toby P Breckon. Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–19, 2020.

Afsana Khan is a PhD candidate at the Department of Advanced Computing Sciences at Maastricht University, The Netherlands. She received her Bachelor's (2018) and Master's degree (2021) in Computer Science from Military Institute of Science and Technology, Bangladesh and University of Tartu, Estonia respectively. Her research interests include artificial intelligence, federated learning and data fusion.

Marijn ten Thij Marijn ten Thij is an assistant professor at the Department of Applied Computing Sciences at Maastricht University. He obtained a MSc in Applied Mathematics at the University of Twente and a PhD in Mathematics (Business Analytics) at the Vrije Universiteit Amsterdam. His research interests involve the usage of mathematical modelling to study and mimic human behavior through data obtained from social media. His current work is at the intersection of the fields Complex Networks, Computational Social Science, and Data Science.

Anna Wilbik is currently Professor in Data Fusion and Intelligent Interaction in the Department of Advanced Computing Sciences of Maastricht University, in the Netherlands. . Currently she is also a chair of The Fuzzy Systems Technical Committee (FSTC) within IEEE CIS. She received her PhD (with honors) in Computer Science from the Systems Research Institute, Polish Academy of Science, Warsaw, Poland, in 2010. In 2011, she was a Post-doctoral Fellow with the Department of Electrical and Computer Engineering, University of Missouri, Columbia, USA. Anna is an alumnus of the Stanford University TOP500 Innovators: Science - Management - Commercialization Program. From 2013 till 2020 she was an Assistant Professor in the Information Systems Group of the Department of Industrial Engineering and Innovation Sciences at Eindhoven University of Technology (TU/e). Her research interests are in data fusion and business intelligence. With her research she tries to bridge the gap between the meaning of data and human understanding in complex application environments, where data can be of various natures. She makes this connection in research projects collaborating with industry both on the national and the European level. She has published over 100 papers in international journals and conferences.