IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

FEDRETAIL: A Framework for Distributed Retail Data Analysis and Learning Toward E-commerce 5.0

Sumit Chongder
Department of Computer Science and Engineering
Maharashtra Institute of Technology - Art, Design and Technology University, Pune, India

Abstract: *Retail data analysis has been identified as a crucial component in the pursuit of E-commerce 5.0. The recent rapid development of Information and Communication Technology (ICT) has revolutionized retail data analysis by providing advanced technologies such as big data analysis and machine learning. However, the privacy of customers has become a significant concern, making retailers hesitant to share their customer data. This reluctance forms isolated data islands, hindering the realization of comprehensive retail data analysis. I propose a federated learning-based retail data analysis framework, FEDRETAIL, to address this challenge. This framework allows retail data analysis federations to be formed by several retailers. None of these retailers need to exchange their customer data with each other directly, and they always keep the data in their place to ensure their customers' privacy. I apply the FEDRETAIL framework to analyze a retail dataset via different federated learning paradigms. The experimental results show that our framework not only guarantees the customers' privacy but also effectively breaks the borders of data islands by achieving higher analysis quality. FEDRETAIL framework closely approaches the performance of centralized analysis, which requires data collection in a commonplace, posing a risk of privacy exposure.*

Keywords: Retail Data Analytics; Federated learning; Machine Learning; E-commerce 5.0.

I. INTRODUCTION

E-Commerce 5.0 is the emerging paradigm of modern and future retail, where various technologies, especially IT technologies, integrate with each other to enhance the customer experience and business outcomes. With the widespread adoption of IT technologies in retail, customers can access products and services anytime and anywhere, and more customer data are available in digital form. This potentially provides the retailers and business managers the possibility of applying IT technologies to unlock the insights for more effective retail, e.g., the customer behavior analysis, product recommendation, pricing optimization, inventory management, etc. Big retail data analysis recently has attracted impressively hot concern together with the development of big data analysis related technologies, especially machine learning [1].

With the development of big retail data analysis, the trade-off between data analysis needs and customer privacy becomes a critically important issue. Protecting the customers' privacy is always of paramount importance. A notorious example is the leakage of Target's customer data in 2013, which exposed the personal information of millions of customers to hackers. Many countries have enacted laws and regulations to safeguard the privacy of customer data, such as the General Data Protection Regulation (GDPR) in the European Union, the California Consumer Sequestration Act (CCSA) in the United States, and the Personal Data Protection Act (PDPA) in Singapore. To comply with these laws and protect the customers' privacy, the retailers often hesitate to share their data, and only keep the customers' data locally, creating isolated data silos. Such hard "isolation" though indeed highly ensures privacy, but contradicts the big retail data analysis, which requires a large volume of data in high variety. To tackle this problem, pioneering researchers have already proposed various privacy-preserving technologies, such as differential privacy, homomorphic encryption, and secure multi-party computation[2]. However, they more or less have problems like limited application scope, low-performance efficiency, and low analysis accuracy. How to exploit the big retail data to well support the retail business with customer's privacy preservation thus becomes a vital challenge in E-commerce 5.0.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

To address such challenges, federated learning emerges as a promising enabling technology. Federated learning is a machine learning technology that enables several participatory parties to collaboratively learn a common model by holding data locally, without sharing data. Clearly, in contrast to traditional centralized machine learning, federated learning does not require the data owners to transfer their data to a centralized server. Such a feature naturally suits the privacy-preserving needs of retail data and inspires us to apply federated learning to retail data analysis and learning for E-commerce 5.0 in this paper. This paper presents several key contributions to the field of federated learning within the retail sector:

It introduces the FEDeratedREtaildaTAanalysIs and Learning (FEDRETAIL) framework, a novel approach to retail data analysis and learning that leverages advanced federated learning technologies.

The paper details a series of real-world, trace-driven experiments designed to assess the framework's viability and performance. These experiments reveal that FEDRETAIL is capable of conducting privacy-preserving analysis and learning, and in certain instances, it achieves greater accuracy than traditional centralized machine learning methods.

It explores and identifies various prospective challenges associated with the practical implementation and application of FEDRETAIL, considering the combined perspectives of retail and information technology.

The paper is structured as follows. Section 2 reviews some existing work on retail data analysis. Section 3 describes the design of the proposed framework for federated retail data analysis and learning. Section 4 presents a comprehensive review and analysis of experiments based on the framework. Section 5 summarizes this work and discusses some future directions.

II. RELATED WORK

Retail data analysis has been gaining popularity recently, especially with the advancement of big data and machine learning technologies. For example, Xiao et al. evaluate 13 machine learning algorithms on the Fashion MNIST dataset [3]in their retail data analysis work [4]. Their investigations show that ANN, RNN, CNN, LSTM, and MLP can achieve the best accuracy in retail data analysis. Ferreira et al. estimate customer's purchase behavior based on "Decision Support Tool", "LP Bound Algorithm" and "Regression"[5]. The outcome shows that prediction accuracy is higher with non-linear kernel methods and neural networks. The application of various machine learning technologies in retail data analysis has been widely discussed [6]. Wang et al. developed a cascade of ensemble predictors to discover search relevance[7]. Some works, in contrast to the earlier studies, focus on choosing features for retail data. Venkatesh et al. propose three methods for feature selection, namely Filter method, Wrapper method and Embeddedmethod, to find the most important and inherent features [8]. Lima et al. employ a feature selection algorithm that blends different methods with various machine learning classifiers[9]. Both the experiment results indicate the importance of pre-processing on feature selection.

Churn prediction also attracts the attention of researchers. Raizada et al. and Saini et al. use retail data mining and different supervised learning techniques to evaluate the accuracy of different prediction models to find the profile of customers who are at risk of churning in a Brazilian e-commerce setting [10]. Ullah et al. use the random forest model and the real data set of Customer Relationship Management(CRM) to identify customers who are likely to churn early and retain them effectively [11]. Lalwani et al. propose a machine learning framework based on SVM for the prediction of churn in e-commerce platforms using only click stream data [12] in their customer churn prediction work.

Through a literature survey, we observe the popularity and high potential of applying various machine learning technologies in retail data analysis. However, they all perform locally on their data, failing to exploit the benefit of big data technology in E-commerce from the perspective of volume, variety, velocity, veracity and value. It is undeniable that more customer data implies higher accuracy in customers' behavior prediction and retail support. Nevertheless, the privacy concern prevents the sharing of customers' data between different retail companies, hindering the development of big retail data analysis. To solve this problem, I am inspired to apply federated learning and propose a federated retail data analysis framework. In this article, I further elaborate on the FEDRETAIL framework and provide a more comprehensive review by exploring two different federated learning paradigms, namely, horizontal federated learning (HFL) and vertical federated learning (VFL), to demonstrate the advantages of our framework.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

III. FEDERATED RETAIL DATA ANALYSIS AND LEARNING FRAMEWORK

Federated learning is a paradigm that allows many servers to train a common model with their local training data under the coordination of an aggregation server. This way, there is no data transfer or sharing between these servers, which reduces the risk of privacy breaches. In this section, I leverage this benefit to construct a FEDeratedREtaildaTAanalysIs and Learning (FEDRETAIL) framework.

A. Architecture

Figure 1 illustrates the architecture of the FEDRETAIL framework. Integrating FEDRETAIL requires no alterations

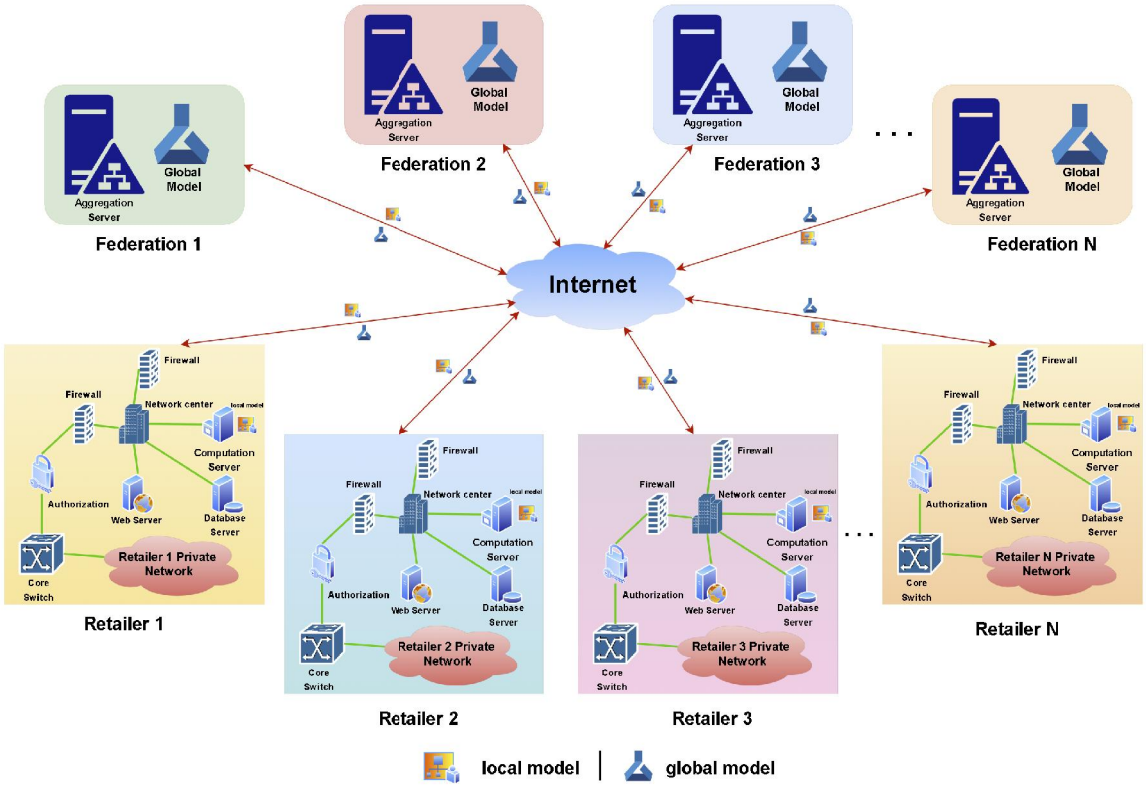


Fig. 1. Architecture of FEDRETAIL framework

to the pre-existing information systems utilized by various retail organizations. Each entity needs only to establish a local federated learning server and engage in communication with the central aggregation server. The remaining components, such as the local private network, web server, and database server, remain unchanged. Thus, FEDRETAIL can be seen as an evolutionary enhancement to the existing retail information management systems, offering a seamless and advantageous transition. Key features of FEDRETAIL's architecture include as follows:

Seamless Integration: No major alterations are required to existing information systems, ensuring a smooth transition to the federated learning framework.

Flexible Participation: Retail entities have the flexibility to participate in multiple federations simultaneously, allowing them to collaborate on different analytical projects as needed.

Data Privacy: FEDRETAIL ensures the privacy and security of sensitive retail data by keeping it within the boundaries of each organization's infrastructure and allowing controlled access during model training.

FEDRETAIL supports the coexistence of multiple retail federations, as depicted in Figure 1. A federation within FEDRETAIL represents a consortium of retail entities united by a shared objective. For instance, numerous E-commerce companies may wish to analyze the impact of diverse variables on consumer buying behavior. By forming a federation, these companies can pool their consumer data to collaboratively train a unified analytical model. In this

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

context, each retail entity acts as a participant within the federation. Notably, a participant has the flexibility to engage in multiple federations simultaneously.

A.1. Network Center

Network center is the main hub for coordinating federated learning activities for the retailer entities. It serves as the central point for communication and collaboration among the participating retail and E-commerce companies. At this center, updates to the model and training parameters from each retail entity are gathered and combined. This creates a comprehensive model that reflects insights from a wide range of data across the retail sector. This centralized method allows retailers to gain from shared intelligence while keeping their own data private and secure.

Additionally, the network center helps set up and manage groups of retailers, known as retail federations, who have common goals. It offers a space for these retailers to outline their joint efforts, agree on how to share data, and work together on training models. Through these federations, the network center enables retailers to tap into the shared knowledge and resources of the wider retail community. This collaboration fosters innovation and improvement in areas like analyzing customer behavior, predicting market demands, and creating targeted marketing campaigns.

A.2. Firewall

Firewall is a critical security component designed to protect the integrity of federated retail data analysis and learning processes. It acts as a barrier between each retailer's local network and the central aggregation server, ensuring that only authorized data and model updates are communicated. This safeguarding mechanism is essential for maintaining the confidentiality and integrity of sensitive retail data, which could include consumer behavior patterns, sales figures, and inventory levels.

The Firewall's configuration allows for the secure transmission of data while preventing unauthorized access, thus enabling retailers and E-commerce platforms to collaborate on data analysis and machine learning projects without compromising their proprietary information. In addition to security, the Firewall within FEDRETAIL Framework also plays a major role in managing network traffic to optimize the performance of data exchanges. It filters out irrelevant or malicious traffic, ensuring that the federated learning process is efficient and free from disruptions.

A.3. Computation Server

Computation Server is a pivotal element that facilitates the processing and analysis of retail data. It's the powerhouse where complex algorithms and machine learning models run to decipher trends and patterns from the data provided by participating retailers and e-commerce entities. This server is responsible for executing the computational tasks of federated learning, where it processes individual contributions from each retailer's dataset to enhance the collective retail model without compromising data privacy.The Computation Server also ensures that the insights gained are relevant and actionable. FATE is a federated learning framework at the industry level created by the AI team at Webank, which allows for AI collaboration among organizations while ensuring the protection of data security and privacy[13]. Google's TensorFlow Federated (TFF) is an open-source framework designed for machine learning and deep learning tasks[14]. NVIDIA has also introduced Clara Federated Learning, which is aimed at distributed and cooperative training within federated learning environments[15]. By providing a secure and robust platform for computation, the server supports the overarching goal of FEDRETAIL: to empower retailers with data-driven intelligence while safeguarding sensitive information within a collaborative, federated learning environment.

A.4. Web Server

The web server within the FEDRETAIL Framework plays a crucial role, acting as the intermediary that facilitates the exchange of information and learning models between the local and central servers. It ensures that the data remains within the private network of the retail organization, thus maintaining data privacy and security. The web server enables multiple retail entities to collaboratively train a single unified analytical model without sharing raw consumer data. This collaborative effort allows for a more comprehensive understanding of consumer buying behavior, as it draws on a diverse set of data points from various participants within the federation process.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

A.5. Database Server

The Database server plays an important role in the FEDRETAIL Framework for managing and storing vast amounts of retail data. It can employ robust database management systems like MySQL, Oracle, and PostgreSQL to ensure efficient handling of transactions and queries. This server is integral to the framework's ability to perform federated learning, as it can securely store the data while allowing for complex analytical operations. Retailers and e-commerce platforms rely on this server to store consumer interactions, inventory details, and sales records, which are then used to train machine learning models like logistics regression, decision tree, SVM, random forest, etc. without exposing raw data. It facilitates the framework's collaborative efforts by providing a unified view of processed data, ready for analysis. With advanced security measures in place, the server ensures that each participant's data remains isolated and protected, fostering a secure environment for collective data-driven insights and strategic decision-making in retail and e-commerce.

A.6. Authentication and Authorization

In the FEDRETAIL Framework, once data traverses from the network center towards a retailer's private network, it encounters a multi-layered security protocol. Initially, the Firewall authenticates the data, verifying its origin and ensuring it's from a trusted source within the federation. Post authentication, the data reaches the authorization phase where the system determines if the data has the necessary permissions to enter the retailer's private network. This process is crucial as it not only confirms the legitimacy of the data but also upholds the retailer's data governance policies, allowing only pertinent and authorized data through. This dual mechanism of authentication and authorization fortifies the framework's security, safeguarding sensitive retail data during federated learning collaborations.

A.7. Aggregation Server

The Aggregation Server within the FEDRETAIL Framework is responsible for collating individual model updates from each participant's local server, and integrating these updates to refine the global model. This global model is a comprehensive representation of shared insights, which benefits all participants by providing a more accurate analysis of retail data patterns without compromising individual data privacy. This global model is then distributed among all the retailer entities and thus resulting in continual learning.

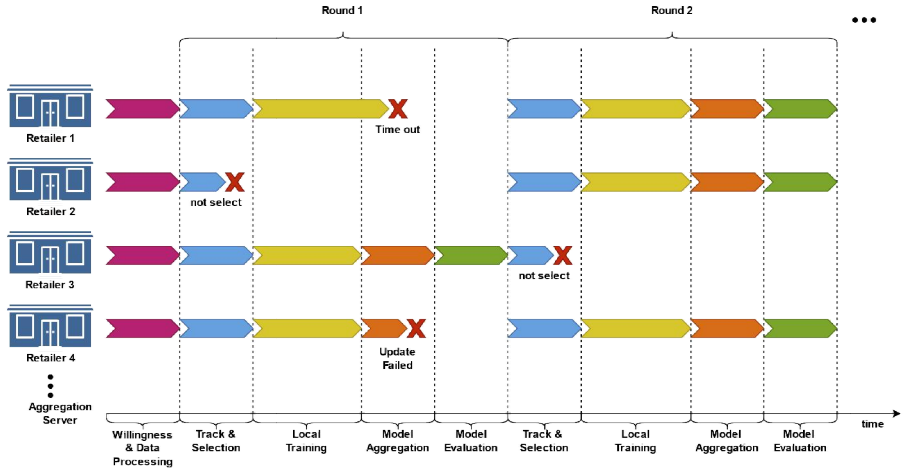


Fig. 2. The Federated Training Procedure in FEDRETAIL

The Aggregation Server ensures that the collective intelligence contributed by the federated network is effectively utilized to enhance the predictive capabilities of the global model, thus supporting data-driven decision-making in retail. It applies complex algorithms to merge local updates in a way that maximizes the value of the aggregated data while minimizing bias and variance. This process involves complex computations to ensure that the global model remains robust and representative of the diverse datasets. For Example, the Federated Averaging (FedAvg) algorithm,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

Federated Stochastic Variance Reduced Gradient (FSVRG) algorithm and Cooperative Machine Learning from Mobile Devices (CO-OP) algorithm for FEDRETAIL are described in Section 5.2.

B. Working Process

Following the architectural overview, let's focus on the working process of FEDRETAIL, which is grounded in federated learning. The key distinction between federated learning and conventional centralized machine learning lies in the training approach of the model. The application of the model for inference remains unchanged. Hence, my attention is directed towards the training methods within FEDRETAIL, as shown in Figure 2. Since federated learning's invention, numerous training frameworks have emerged, each varying in participant collaboration.

The FEDRETAIL Framework operates through a structured process to facilitate federated learning among retail entities. Initially, all participating retailers are gathered, and their willingness to contribute to the federated learning process is assessed. Once confirmed, the framework proceeds to data processing, where each retailer's data is prepared for analysis. Following this, a tracking and selection stage identifies the most relevant data for the current learningobjective. Subsequently, local training occurs at each retailer's server, utilizing their specific data to develop individual models. These models are then sent to the central server for aggregation, where they are combined to form a global model. This model encapsulates the collective knowledge of all participants. The final step in the process is the evaluation of the global model to ensure its accuracy and effectiveness.

In cases where a retailer's client times out during local training, the framework is designed to resume the process from

Dataset Characteristics	Details
Dataset Size	The dataset consists of 70.000 grayscale images
Image Dimensions	Each image in the dataset is a 28x28 pixel grayscale image
Categories	The dataset includes images of 10 types of fashion products such as t-shirt/top, trouser, pullover, dress, coat, sandal, shirt, sneaker, bag and ankle boot
Class Labels	Each image is associated with a label from 0 to 9, representing the 10 categories of clothing
Training and Testing Split	The dataset is divided into training set and a testing set. The training set consists of 60,000 images, and the test set consists of 10,000 images
Usage	The Fashion-MNIST dataset is commonly used for training and testing machine learning models, especially in the field of image classification

Table 1. Main feature information about Fashion MNIST

the tracking and selection stage in the next round. This ensures continuity and the inclusion of all participant's data in the learning process. Similarly, if a retailer fails to make a selection during the tracking and selection stage, or if there's a failure to update the aggregated model, the client will pick up from the respective stage in the subsequent round. This resilience in the workflow allows for robust learning and model development, despite any temporary setbacks.

IV. COMPREHENSIVE ANALYSIS AND REVIEW

In this section, I embark on an in-depth exploration of the FEDRETAIL Framework, a pioneering approach in thedomain of federated learning, specifically oriented for the retail industry. Utilizing the widely recognized Fashion MNIST dataset, a benchmark in machine learning for fashion article classification, my investigation delves into the nuances of decentralized data processing while upholding stringent privacy standards. This dataset, comprising a diverse array of clothing items, serves as the cornerstone for my analysis, enabling me to evaluate the efficacy and robustness of FEDRETAIL. Through meticulous examination, I aim to uncover insights that could revolutionize retail analytics and foster advancements in privacy-preserving federated learning paradigms like Horizontal Federated Learning (HFL) and Vertical Federated Learning (VFL).

A. Dataset Introduction

The FEDRETAIL framework leverages the Fashion MNIST dataset, a modern alternative to the traditional MNIST dataset, tailored for benchmarking machine learning algorithms in the context of fashion. The dataset comprises 70,000 (28x28) grayscale images of fashion articles, evenly distributed across 10 distinct classes: T-shirt/top, Trouser,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

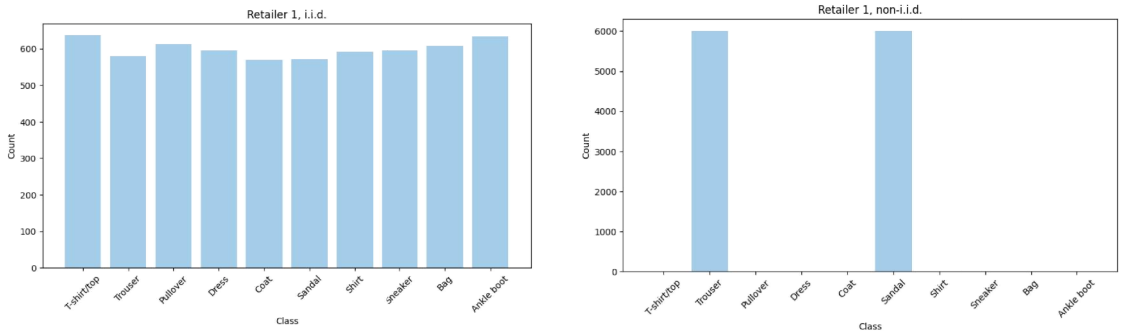


Fig. 3. Side-by-side comparison of an i.i.d. and non-i.i.d. distribution taken from one of the retailers.

Pullover, Dress, Coat, Sandal, Shirt, Sneaker, Bag, and Ankle boot. This diversity allows for a comprehensive analysis of classification algorithms within the retail sector.

For simulating a realistic retail environment, the dataset is partitioned into both independent and identically distributed (i.i.d.) [16]and non-independent and identically distributed (non-i.i.d.)[17] among the retailers, as demonstrated in Figure 3. This distribution ensures that each retailer's dataset is a representative subset of the whole, maintaining uniformity in class distribution and data volume. Such a setup is crucial for evaluating the performance of the FEDRETAIL framework in a decentralized and collaborative retail setting, allowing for comprehensive analysis across varying data distributions.The main features of the Fashion MNIST dataset are summarized in Table 1, which provides a detailed overview of the dataset's composition and structure.

The selection of the Fashion MNIST dataset for this study is strategic, as it serves as a benchmark for comparing various machine learning algorithms[3]. Its complexity and variety offer a challenging yet manageable benchmark for federated learning in E-commerce 5.0. Moreover, the dataset's relevance to the retail industry makes it an ideal candidate for demonstrating the potential of federated learning to revolutionize data analytics and machine learning in retail. By employing this dataset, the FEDRETAIL framework aims to showcase the efficacy of federated learning in handling diverse, decentralized data while preserving privacy and enabling collaborative insights.

B. Algorithms

In order to enhance the performance of the FEDRETAIL Framework, I embarked on various federated learning algorithms. My objective was to identify the algorithm that would yield the highest accuracy while minimizing loss, thereby optimizing the effectiveness of the framework. Experimentation was conducted on the Fashion MNIST dataset, a widely recognized benchmark in the machine learning community. I experimented on three distinct federated learning algorithms, each with its unique strengths and characteristics. The comparative analysis was based on two key metrics: accuracy and loss. The following sections will delve into the specifics of each algorithm, detailing their implementation, performance metrics, and the rationale behind the selection of the optimal algorithm for the FEDRETAIL Framework.

B.1. Federated Stochastic Variance Reduced Gradient (FSVRG)

The Federated Stochastic Variance-Reduced Gradient (FSVRG) algorithm is designed to address the challenges oftraining machine learning models in federated learning settings where data is distributed across multiple clients.Algorithm1 gives a complete description of FSVRG, where it commencesas follows: First, initializing the global model parametersθ_g, which represent the model to be optimized across all retailers. Additionally, the algorithm initializes local model θ_c on each retailer c, where c ranges from 1 to N, denoting the total number of retailers participating in the federated learning process.

Algorithm 1: Federated SVRG

1 initialize global model parameters: $\boldsymbol{\theta_g}$
2 initialize local models on each retailer: $\boldsymbol{\theta_c}$
3 initialize global gradient accumulator: $\boldsymbol{g}$

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

initialize global variance accumulator: $\boldsymbol{\vartheta}$
for each $\boldsymbol{epoch\ t = 1\ to\ T}$ do
  for each $\boldsymbol{retailer\ c = 1\ to\ N}$ do
    Sample a mini-batch from Retailer data ($\boldsymbol{D_c^B}$)
    $$\boldsymbol{\nabla f_c(\theta_c) = \frac{1}{B}\sum_{(x,y)\in D_c^B} \nabla f(\theta_c, x, y)}$$
    $$\boldsymbol{g = g + \nabla f_c(\theta_c)}$$
    $$\boldsymbol{\vartheta_c = \frac{1}{B}\sum_{(x,y)\in D_c^B} |\nabla f(\theta_c, x, y) - \nabla f_c(\theta_c)|^2}$$
    $$\boldsymbol{\vartheta = \vartheta + \vartheta_c}$$
  end for
  $$\boldsymbol{\nabla F(\theta_g) = \frac{g}{N}}$$
  $$\boldsymbol{V(\theta_g) = \frac{\vartheta}{N}}$$
  $\theta_g = \theta_g - \eta\left(\frac{\nabla F(\theta_g)}{V(\theta_g)}\right)$
  // Reset accumulators for next epoch
  $\boldsymbol{g = 0}$
  $\boldsymbol{\vartheta = 0}$
end for

FSVRG algorithm requires several hyperparameters to be specified such as, Learning rate (α), Number of retailers (N), Number of epochs (T), etc. The FSVRG algorithm iterates through multiple epochs of training, with each epoch consisting Retailer-Side Computation, Aggregation, Global Model Update and Accumulator Reset. In the Client-Side Computation phase each retailer c samples a mini-batch D_c^B from its local dataset. Using the sampled mini-batch, the client computes its local gradient $\nabla f_c(\theta_c)$ and local variance ϑ_c. In the Aggregation phase, retailers communicate their local gradients and variances to the server. The server aggregates the received gradients to compute the global gradient estimate $\nabla F(\theta_g)$ and the global variances estimate $V(\theta_g)$.

Post Aggregation, the server updates the global model parameters θ_g using the aggregated gradient and variance estimates. The update is performed according to the following equation:

$$\theta_g = \theta_g - \eta\left(\frac{\nabla F(\theta_g)}{V(\theta_g)}\right)$$

After updating the global model parameters, the server resets the accumulators $\boldsymbol{g}$ and $\boldsymbol{\vartheta}$ to zero in preparation for the next epoch. Upon completion of all the epochs, the FSVRG algorithm outputs the final global model parameters $\boldsymbol{\theta_g}$, which represents the optimized model learned from the federated dataset.

B.2.Cooperative Machine Learning from Mobile Devices (CO-OP)

The Cooperative Machine Learning from Mobile Devices (CO-OP) algorithm is designed to enable collaborative model training while preserving data privacy on decentralized mobile devices. It operates over a series of communication rounds, each comprising local model updates and global parameter aggregation. The algorithm requires several inputs: a training dataset$D$, partitioned into N subsets$D_1, D_2, \ldots, D_N$; a machine learning model M; the number of communication rounds T; a communication interval K; the number of local update iterations C; and a learning rate(α).

Algorithm2 gives a complete description of CO-OP, which commences with the initialization of model parameters theta, typically initialized with random values. Subsequently, it iterates through each communication round$r = 1\ to\ T$. Within each round, each retailer c selects its respective data subset D_c and initializes its local model parameters θ_c with the current global parameter θ_g. Local model updates are then performed on each retailer for a predefined number of iterations, denoted by C. This involves gradient descent optimization, where the local loss function L_c with

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

Algorithm 2: CO-OP

initialize local model parameters:θ_c
Initialize global model parameters: θ_g
for $t = 1$ *to* T **do**
  for each *retailer* $c = 1$ *to* N **do**
    Select subset D_c
  end for
  for each *retailer* $c = 1$ *to* N **do**
    for $j = 1$ *to* C **do**
      $\theta_c = \theta_c - \alpha \cdot \nabla L_c(\theta_c; D_c)$
    end for
  end for
  if $t \,\%\, K == 0$ **then**
    $\theta_g = \frac{1}{N}\sum_{c=1}^{N} \theta_c$
  end if
end for

Table 3. CO-OP algorithm for FEDRETAIL framework

respect to θ_c, computed on the data subset D_c. The update equation is given by:

$$\theta_c = \theta_c - \alpha \cdot \nabla L_c(\theta_c; D_c)$$

Following local updates, each retailer transmits its updated model parameters θ_c to a central server. Model parameter aggregation occurs at regular intervals specified by K. When the current round number t is divisible by K, the central server aggregates the received model parameters from all retailers, computing the average to obtain the updated global model parameter θ_g:

$$\theta_g = \frac{1}{N}\sum_{c=1}^{N} \theta_c$$

This process repeats for a total of T communication rounds. Finally, the algorithm outputs the final global model parameterθ_g, which have been collaboratively trained across all devices.

B.3. Federated Averaging (FedAvg)

Federated Averaging (FedAvg) is a distributed optimization algorithm designed for federated learning settings, where data is distributed across multiple retailers. FedAvg aims to train a global model by aggregating local updates from individual retailers while preserving data privacy.FedAvg begins by initializing the global model parametersθ_g, which represents the model to be optimized across all the retailers. Additionally, each retailer c initializes its localmodel parameters θ_c, where c ranges from 1 to N, denoting the total number of retailers participating in the federated learning process.

Algorithm3 gives a complete description of FedAvg algorithm, which iterates multiple communication rounds. At the beginning of each communication round, a subset of retailers S_r is selected, where $|S_r| = C \cdot K \geq 1$.In the next stage, the current global model parameters θ_g are distributed to all retailers in the selected subset S_r by the central server. Each retailer c updates its local model parameters θ_c to match the shared global model θ_g. Retailers then perform local training using their own data for E epochs with a mini-batch size of B.

After local training, retailers upload their trained local model parameters θ_c to the server. The server aggregates the received local models to compute the new global model parameters θ_{g+1} using federated averaging. The aggregation process is defined as follows:

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

$$\theta_{g+1} = \sum_{c=1}^{K} \frac{n_c}{N} \theta_c$$

The convergence condition is checked to determine whether the training process should continue or terminate. Once the change in the global model parameters falls below a predefined threshold (ε), the algorithm gets terminated as the convergence condition is satisfied. This convergence criterion ensures that the optimization process stops when the global model stabilizes and further iterations do not significantly improve the model performance.The process of checking whether the convergence condition is defined as follows:

$$\|\theta_{g+1} - \theta_g\| \leq \varepsilon$$

Algorithm 3: Federated Averaging (FedAvg)

initialize global model parameters: θ_g
initialize local model parameters: θ_c
for each $round\ r$ = 0, 1, 2, 3, … **do**
  // Subset of retailers S_r
  $|S_r|$ = C · K ≥ 1
  for each $retailer\ c\ in\ S_r$ **do**
    // Global model from server
    $\theta_c \leftarrow \theta_g$
    θ_c = RetailerUpdate()
  end for
  $\theta_{g+1} = \sum_{c=1}^{K} \frac{n_c}{N} \theta_c$
end for

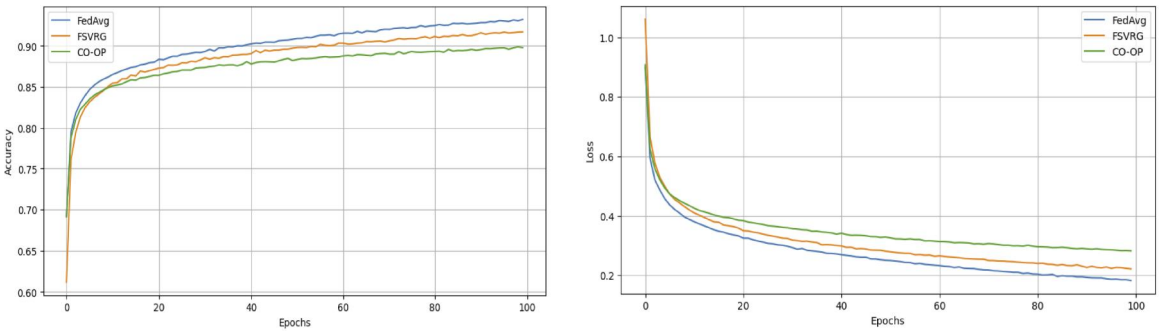
Upon satisfying the convergence condition, FedAvg outputs the final global model parameters θ_g, which represents the optimized model learned from the federated dataset.

C. Comparative Analysis of FedAvg, FSVRG, and CO-OP Algorithms

The implementation of the FedAvg, FSVRG, and CO-OP algorithms within the FEDRETAIL Framework has yielded insightful data, encapsulated in Figure 4. Figure 4 (a) illustrates the accuracy comparison across 100 epochs, showcasing the precision and reliability of each algorithm in predicting retail trends. Conversely, Figure 4 (b) delineates the loss comparison, providing a clear depiction of the efficiency and performance drawbacks inherent to each algorithm.

Based on the insights obtained from the comparative analysis depicted in Figure 4, the FedAvg algorithm has been integrated at the federation block of the FEDRETAIL Framework, as illustrated in Figure 1, due to its performance in benchmark tests with the Fashion MNIST dataset. It outperformed other algorithms regarding accuracy and loss minimization, which are critical for the framework's efficacy. Furthermore, FedAvg's architecture is particularly well-adapted for the federated learning environment of FEDRETAIL, which operates across both horizontal and vertical retail domains. Its straightforward implementation and decentralized optimization strategy enable efficient model training and updates across diverse retail clients, ensuring a scalable and privacy-preserving solution for collaborative learning within the retail sector.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024



(a) Accuracy comparison (b) Loss comparison

Fig. 4. Federated Learning algorithm performance comparison

D. Horizontal Retail Federated Learning (HRFL) Experiment

Horizontal Retail Federated Learning (HRFL) is introduced in scenarios where data is distributed across multipleentities that share the same feature space but differ in samples, as depicted in Figure 5. The core idea behind HRFL is to collaboratively train a model without exchanging raw data, thus preserving privacy and security. In the FEDRETAIL Framework, HRFL was implemented using a dataset evenly distributed among 10 retailers, with an i.i.d.distribution of the Fashion MNIST dataset's 60,000 training samples, as depicted in Figure 3. Each retailer received 6,000 samples covering all 10 classes. The data was then trained using a two-layer neural network and softmax regression (multinomial logistic regression) to evaluate the framework's performance.

In horizontal FEDRETAIL, the FedAvg algorithm is implemented at the federation block, as discussed in Section 4.3. The retailer participation probability was set at 10%, meaning that on average, 10 retailers participated in each training round. The learning rate, local epoch, and batch size were configured to 0.001, 2, and 64, respectively. To demonstrate horizontal FEDRETAIL's efficiency, its performance was benchmarked against two alternatives: one where model training was conducted locally with each retailer's own data, and another where all data was aggregated centrally on the server. This comparison highlighted the advantages of horizontal FEDRETAIL's federated approach in terms of efficiency and data privacy.

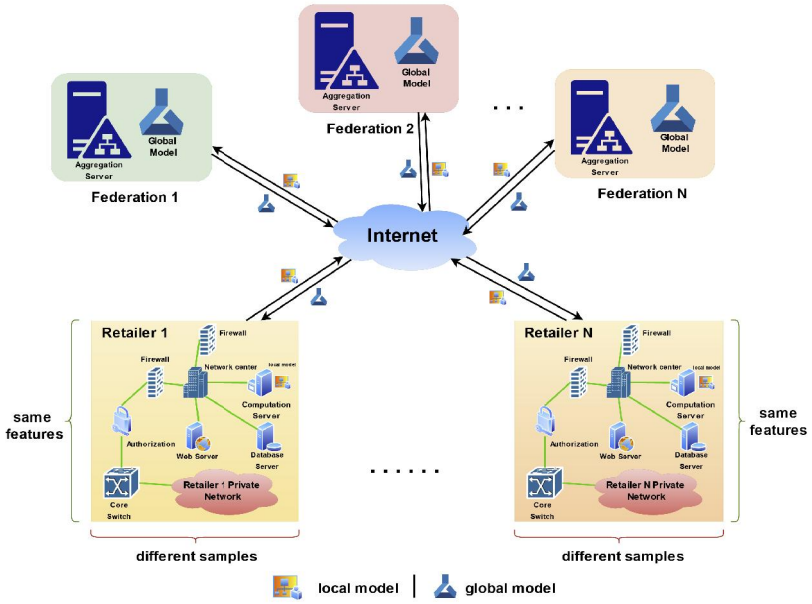


Fig. 5. Architecture overview of horizontal retail federated learning

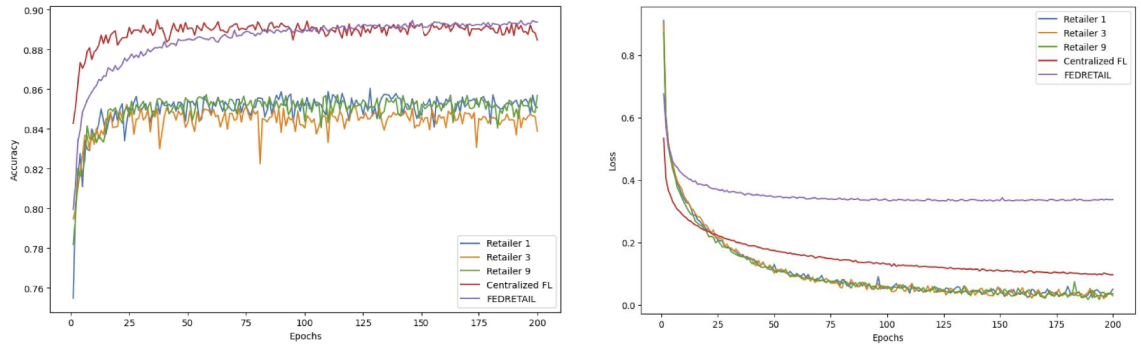
In the horizontal FEDRETAIL Framework, all algorithms were subjected to a rigorous 200-epoch evaluation to assess their accuracy and loss metrics, with the outcomes detailed in Figure 6 (a) and 6 (b) for accuracy and Figure 7 (a) and 7

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

(b) for loss. The framework's design allowed for the selection of a subset of retailers to provide a snapshot of local training results. The data presented in these figures illustrates a consistent increase in accuracy over successive epochs across all training methodologies, underscoring FEDRETAIL's capability to deliver precise classification of clothing items within the Fashion MNIST dataset. Notably, centralized training, which consolidates all retailer data onto a single server, demonstrated the quickest convergence in accuracy.

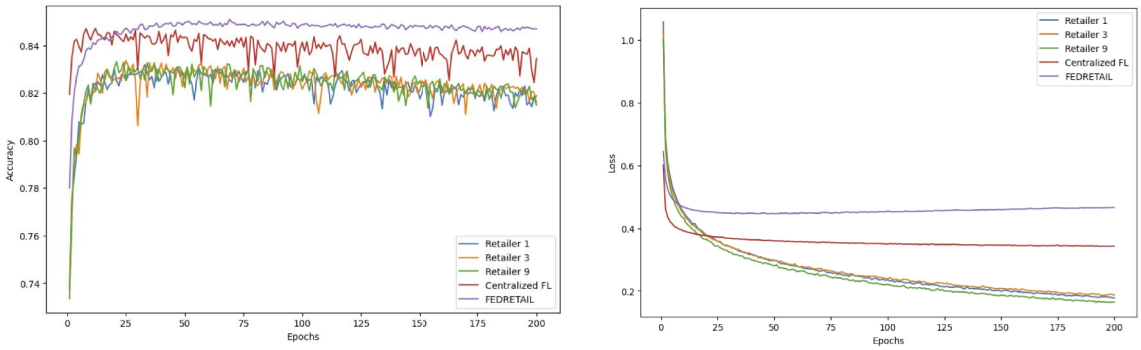
However, horizontal FEDRETAIL's performance was notably superior to scenarios where only local data was utilized for training. Over time, as horizontal FEDRETAIL's training progressed, it began to mirror the performance of the centralized approach, benefiting from the most favorable outcomes were recorded at the 0.7 probability level, which is reported in Figure 8 (a) and 8 (b) using two-layer neural network and Figure 9 (a) and 9 (b) using softmax regression respectively.

This empirical evidence supports the notion that a more extensive network of participants within a federation significantly enhances the training of the global model, thereby providing collective benefits to all members within the retail federation.



(a) Accuracy comparison (b) Loss comparison

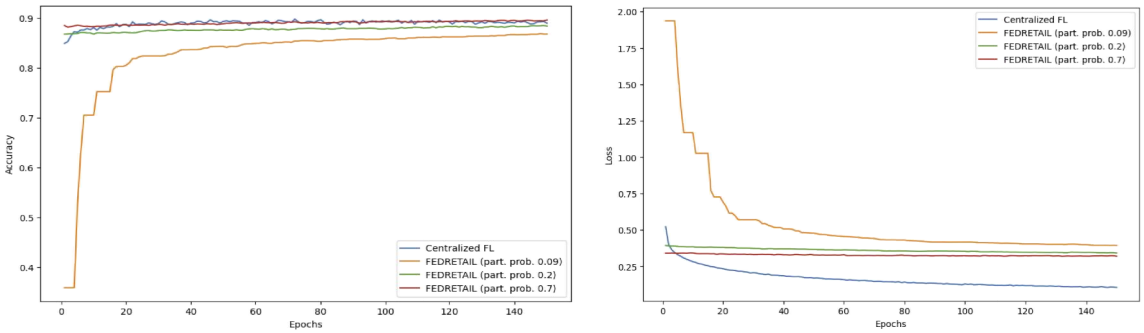
Fig. 6. Result comparison during HRFL training using two-layer neural network



(a) Accuracy comparison (b) Loss comparison

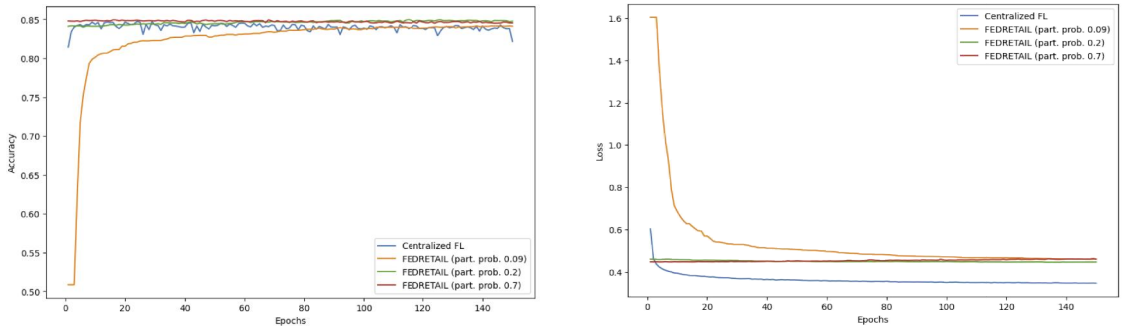
Fig. 7. Result comparison during HRFL training using softmax regression

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024



(a) Accuracy comparison (b) Loss comparison

Fig. 8. Impact of varied participation probabilities using two-layer neural network



(a) Accuracy comparison (b) Loss comparison

Fig. 9. Impact of varied participation probabilities using softmax regression

E. Vertical Retail Federated Learning (VRFL) Experiment

Vertical Retail Federated Learning (VRFL) is introduced as a collaborative approach where multiple retailers, eachwith unique features but sharing the same sample space, work together to improve machine learning models without sharing sensitive data, as depicted in Figure 10. The concept is particularly relevant in scenarios like the Fashion MNIST dataset, wheredifferent retailers may have distinct types of information for the same set of clothing items. VRFL allows these retailers to contribute to a collective learning process, enhancing the model's predictive accuracy while maintaining dataprivacy. Within the vertical FEDRETAIL Framework, the implementation was tested with the dataset distributed among 10 retailers in a non-independent and identically distributed (non-i.i.d.) manner across all 10 classes, as depicted in figure 3. This setup mimics real-world scenarios where data distribution is often skewed. The processed data was then trained using logistic regression to evaluate the framework's efficacy in handling such diverse data distributions.

Similarly, to demonstrate the efficiency of vertical FEDRETAIL, the performance efficiency of this framework is compared against two competitors. One approach involves aggregating all the data into a central server, while the other involves conducting model training locally with only the retailer's data. These algorithms were run for 200 epochs, and their loss values were evaluated. The training results are reported in Figure 11. It can be observed that the loss value gradually decreases with the training epochs for any training method, except for local training at centralized federated learning. Interestingly, it can also be noted that the decreasing trend of vertical FEDRETAIL's training loss mirrors that of the retailer's training method. This validates the feasibility of applying vertical retail federated learning,as it can indeed handle the case where the same customer's records are involved in different retailer's datasets, without

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

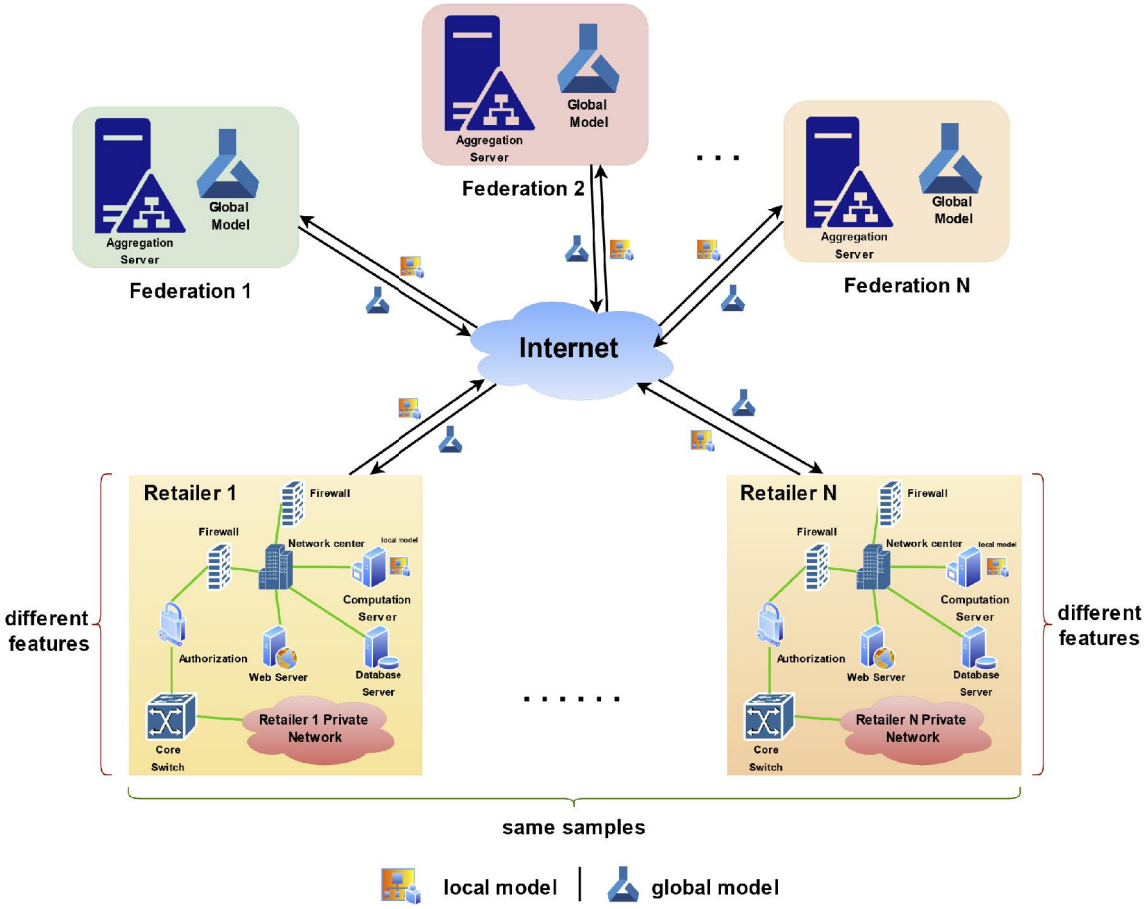


Fig. 10. Architecture overview of vertical retail federated learning

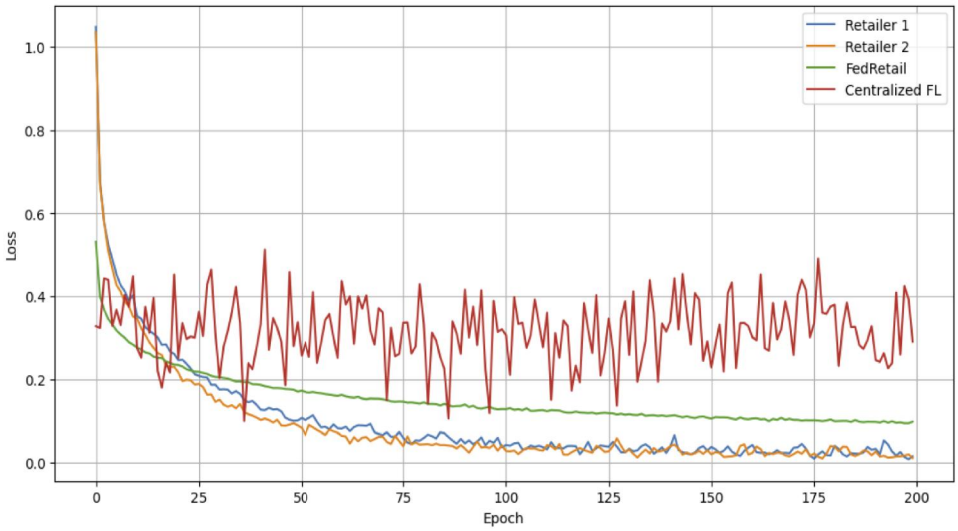


Fig. 11. Loss comparison during VRFL training

compromising customer privacy. Moreover, it can be observed that vertical FEDRETAIL exhibits superior performance compared to any other case with local data-only training. As the number of training epochs increases, vertical

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

FEDRETAIL gradually approaches the performance of the retailers' training method. This underscores the potential of FEDRETAIL in the retail industry.

V. CONCLUSION

In this research, the challenge of isolated retail data islands is addressed by introducing a federated retail data analysis framework, FEDRETAIL, based on federated learning principles. A case study has been conducted to evaluate retailer performance using a public retail dataset. The results demonstrate the feasibility of applying FEDRETAIL for collaborative analysis of retailers' data while ensuring data privacy. Furthermore, the benefits of eliminating the barriers between data islands are highlighted because the framework achieves superior prediction accuracy compared to individual training on local data alone. It is believed that FEDRETAIL offers a promising potential solution for retail data analysis, paving the way for forming retail federations in the era of E-commerce 5.0. This could be beneficial in enhancing retail practices and improving retailer performance. As the sole contributor to this research, the transformative impact of FEDRETAIL in the retail industry is eagerly anticipated.

REFERENCES

[1] S. Akter and S. F. Wamba, 'Big data analytics in E-commerce: a systematic review and agenda for future research', *Electronic Markets*, vol. 26, no. 2, pp. 173–194, May 2016, doi: 10.1007/s12525-016-0219-0.
[2] J. Park and H. Lim, 'Privacy-Preserving Federated Learning Using Homomorphic Encryption', *Applied Sciences (Switzerland)*, vol. 12, no. 2, Jan. 2022, doi: 10.3390/app12020734.
[3] H. Xiao, K. Rasul, and R. Vollgraf, 'Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms', Aug. 2017, [Online]. Available: http://arxiv.org/abs/1708.07747
[4] L. Eglite and I. Birzniece, 'Retail Sales Forecasting Using Deep Learning: Systematic Literature Review', *Complex Systems Informatics and Modeling Quarterly*, vol. 2022, no. 30, pp. 53–62, 2022, doi: 10.7250/csimq.2022-30.03.
[5] K. J. Ferreira, H. A. Lee, and D. Simchi-Levi, 'Analytics for an Online Retailer: Demand Forecasting and Price Optimization'. [Online]. Available: www.ruelala.com
[6] G. Chaubey, P. R. Gavhane, D. Bisen, and S. K. Arjaria, 'Customer purchasing behavior prediction using machine learning classification techniques', *J Ambient IntellHumanizComput*, vol. 14, no. 12, pp. 16133–16157, Dec. 2023, doi: 10.1007/s12652-022-03837-6.
[7] Q. Wang, 'E-commerce Sites Search Results Relevance Prediction Based on Ensemble Approach', 2017.
[8] B. Venkatesh and J. Anuradha, 'A review of Feature Selection and its methods', *Cybernetics and Information Technologies*, vol. 19, no. 1, pp. 3–26, 2019, doi: 10.2478/CAIT-2019-0001.
[9] H. C. S. C. Lima, F. E. B. Otero, L. H. C. Merschmann, and M. J. F. Souza, 'A Novel Hybrid Feature Selection Algorithm for Hierarchical Classification', *IEEE Access*, vol. 9, pp. 127278–127292, 2021, doi: 10.1109/ACCESS.2021.3112396.
[10] S. Raizada and J. R. Saini, 'Comparative Analysis of Supervised Machine Learning Techniques for Sales Forecasting', *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 11, pp. 102–110, 2021, doi: 10.14569/IJACSA.2021.0121112.
[11] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, 'A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector', *IEEE Access*, vol. 7, pp. 60134–60149, 2019, doi: 10.1109/ACCESS.2019.2914999.
[12] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, 'Customer churn prediction system: a machine learning approach', *Computing*, vol. 104, no. 2, pp. 271–294, Feb. 2022, doi: 10.1007/s00607-021-00908-y.
[13] Y. Liu, T. Fan, T. Chen, Q. Xu, and Q. Yang, 'FATE: An Industrial Grade Platform for Collaborative Learning With Data Protection', 2021. [Online]. Available: https://www.fedai.org.
[14] I. Kholod *et al.*, 'Open-source federated learning frameworks for IoT: A comparative review and analysis', *Sensors (Switzerland)*, vol. 21, no. 1, pp. 1–22, Jan. 2021, doi: 10.3390/s21010167.
[15] H. R. Roth *et al.*, 'Empowering Federated Learning for Massive Models with NVIDIA FLARE', Feb. 2024, [Online]. Available: http://arxiv.org/abs/2402.07792

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

[16] M. Arafeh, A. Hammoud, H. Otrok, A. Mourad, C. Talhi, and Z. Dziong, 'Independent and Identically Distributed (IID) Data Assessment in Federated Learning', in *Proceedings - IEEE Global Communications Conference, GLOBECOM*, 2022, pp. 293–298. doi: 10.1109/GLOBECOM48099.2022.10001718.
[17] H. Zhu, J. Xu, S. Liu, and Y. Jin, 'Federated learning on non-IID data: A survey', *Neurocomputing*, vol. 465, pp. 371–390, Nov. 2021, doi: 10.1016/j.neucom.2021.07.098.