ASCE

# Data Collaboration Analysis Framework Using Centralization of Individual Intermediate Representations for Distributed Data Sets

Akira Imakura[1] and Tetsuya Sakurai[2]



**Abstract:** This paper proposes a data collaboration analysis framework for distributed data sets. The proposed framework involves centralized machine learning while the original data sets and models remain distributed over a number of institutions. Recently, data has become larger and more distributed with decreasing costs of data collection. Centralizing distributed data sets and analyzing them as one data set can allow for novel insights and attainment of higher prediction performance than that of analyzing distributed data sets individually. However, it is generally difficult to centralize the original data sets because of a large data size or privacy concerns. This paper proposes a data collaboration analysis framework that does not involve sharing the original data sets to circumvent these difficulties. The proposed framework only centralizes intermediate representations constructed individually rather than the original data set. The proposed framework does not use privacy-preserving computations or model centralization. In addition, this paper proposes a practical algorithm within the framework. Numerical experiments reveal that the proposed method achieves higher recognition performance for artificial and real-world problems than individual analysis. **DOI: 10.1061/AJRUA6.0001058.** 

**Author keywords:** Data collaboration analysis; Distributed data sets; Intermediate representation.

## Introduction

Dimensionality reduction methods that project high-dimensional data to a low-dimensional space are successfully applied in several application areas to improve the prediction performance and accelerate machine learning algorithms, including gene expression data analysis (Tarca et al. 2006), chemical sensor data analysis (Jurs et al. 2000), social network analysis (Tichy et al. 1979), infrastructure analysis (Lasisi and Attoh-Okine 2018, 2020) and so on. Recently, there has been a rise in large and distributed data, and the costs of data collection have decreased. Centralizing distributed data sets and analyzing as one data set, which we refer to as centralized analysis, can enable us to obtain novel insights and achieve higher prediction performance than that of individual analysis on an individual distributed data set. However, it is generally difficult to centralize the original data sets because of large data size or privacy concerns.

For example, in the case of medical data analysis, the data sets in each medical institution may not be sufficient for generating a high-quality prediction result because of insufficiency and imbalance of the data samples. However, it is difficult to centralize the original medical data samples with those from other institutions because of privacy concerns. If the original data is transformed to another (e.g., low-dimensional) space by an appropriate mapping; however, the mapped data, which is referred to as an intermediate representation, can be centralized fairly easily because each feature of the intermediate representation lacks any physical interpretation.

Examples of overcoming the difficulties of centralized analysis include the usage of privacy-preserving computation based on cryptography (Jha et al. 2005; Kerschbaum 2012; Cho et al. 2018; Gilad-Bachrach et al. 2016) and differential privacy (Abadi et al. 2016; Ji et al. 2014; Dwork 2006). Federated learning (Konečnỳ et al. 2016; McMahan et al. 2016), in which a model is centralized while the original data sets remain distributed, has also been studied in this context.

In contrast to these existing methods, this paper proposes a data collaboration analysis framework for distributed data sets that centralizes only individually constructed intermediate representations. The proposed framework assumes that each institution uses a different mapping function for constructing intermediate representations. The framework does not centralize the mapping functions to avoid a risk of approximating the original data samples from their intermediate representations by using the (approximate) inverse of the mapping functions. The proposed data collaboration analysis framework also does not use privacy-preserving computation. Instead, using sharable data such as public data and randomly constructed dummy data, the proposed framework achieves a data collaboration analysis by mapping individual intermediate representations to incorporable representations referred to as collaboration representations.

This paper additionally proposes a practical algorithm and a practical operation strategy regarding the problem of privacy preservation. Using numerical experiments on artificial and real-world data sets, the recognition performance of the proposed method is evaluated and compared with centralized and individual analyses.

The main contributions of this paper are as follows:

- We propose a data collaboration analysis framework using centralization of the individual intermediate representations

[1]Associate Professor, Dept. of Computer Science, Univ. of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan (corresponding author). ORCID: https://orcid.org/0000-0003-4994-2499. Email: imakura@cs.tsukuba.ac.jp

[2]Professor, Dept. of Computer Science, Univ. of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan. Email: sakurai@cs.tsukuba.ac.jp

Note. This manuscript was submitted on July 3, 2019; approved on November 20, 2019; published online on February 28, 2020. Discussion period open until July 28, 2020; separate discussions must be submitted for individual papers. This paper is part of the ***ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering***, © ASCE, ISSN 2376-7642.

that does not centralize the original data sets for distributed data sets.
- The proposed framework differs from existing approaches as it does not use privacy-preserving computations or model centralization.
- The proposed data collaboration analysis achieves higher recognition performance than that produced by individual analysis.

## Data Collaboration Analysis Framework

In this section, we discuss the case that there are multiple institutions and each institution has an individual data set. We propose a data collaboration analysis framework for distributed data sets that does not centralize the original data. The proposed method can be considered a dimensionality reduction method for distributed data sets. The distributed original data sets are transformed into the collaboration representations via the intermediate representations. Therefore, after constructing the collaboration representations, we can use any machine learning algorithms including unsupervised, supervised, and semi-supervised learning.

Let $d$ be the number of institutions. Let $m$, $n_i$ be the numbers of features and training data samples of the $i$th institution and $n$ be the total number of training data samples, $n = \sum_{i=1}^{d} n_i$. In addition, let $\mathbf{X}_i = [\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{in_i}] \in \mathbb{R}^{m \times n_i}$ be the training data set of the $i$th institution. For supervised learning, we additionally let $\mathbf{L}_i = [\mathbf{l}_{i1}, \mathbf{l}_{i2}, \ldots, \mathbf{l}_{in_i}] \in \mathbb{R}^{l \times n_i}$ be the ground truth for the training data. Also let $s_i$ be the number of test data samples of the $i$th institution, $s = \sum_{i=1}^{d} s_i$ and $\mathbf{Y}_i = [\mathbf{y}_{i1}, \mathbf{y}_{i2}, \ldots, \mathbf{y}_{is_i}] \in \mathbb{R}^{m \times s_i}$ be test data sets of the $i$th institution.

We do not centralize the original data set $\mathbf{X}_i$ (and $\mathbf{Y}_i$ in supervised learning). Instead, we centralize the intermediate representations constructed individually from $\mathbf{X}_i$. We also do not centralize the mapping function for the intermediate representation to reduce the risk of approximating the original data.

In the remainder of this section, we introduce a fundamental concept of the data collaboration analysis framework and propose a practical algorithm. In addition, we consider a practical operation strategy regarding privacy concerns.

### Fundamental Concept and Framework

Instead of centralizing the original data set $\mathbf{X}_i$, we consider centralizing the intermediate representation

$$\tilde{\mathbf{X}}_i = [\tilde{\mathbf{x}}_{i1}, \tilde{\mathbf{x}}_{i2}, \ldots, \tilde{\mathbf{x}}_{in_i}] = f_i(\mathbf{X}_i) \in \mathbb{R}^{\ell_i \times n_i} \tag{1}$$

constructed individually in each institution, where $f_i$ is a linear or nonlinear column-wise mapping function. Since each mapping function $f_i$ is constructed using $\mathbf{X}_i$ individually, $f_i$ and its dimensionality $\ell_i$ depend on $i$.

Examples of the mapping function include unsupervised dimensionality reductions, such as principal component analysis (PCA) (Pearson 1901; Jolliffe 1986); locality preserving projections (LPP) (He and Niyogi 2004); and supervised dimensionality reductions, such as Fisher discriminant analysis (FDA) (Fisher 1936), local FDA (LFDA) (Sugiyama 2007), semi-supervised LFDA (SELF) (Sugiyama et al. 2010), locality adaptive discriminant analysis (LADA) (Li et al. 2017); and complex moment-based supervised eigenmap (CMSE) (Imakura et al. 2019). One can also consider a partial structure of deep neural networks. The proposed framework aims to avoid difficulties of centralized analysis by achieving collaboration analysis while the original data set $\mathbf{X}_i$ and the mapping function $f_i$ remain distributed in each institution.

Because $f_i$ depends on the institution $i$, even when each institution has an identical data sample $\mathbf{x}$, the intermediate representation of the data differs; that is

$$f_i(\mathbf{x}) \neq f_j(\mathbf{x}) \quad (i \neq j) \tag{2}$$

In addition, the relationship between the original data samples $\mathbf{x}$ and $\mathbf{y}$ is generally not preserved across different institutions; that is

$$D(f_i(\mathbf{x}), f_j(\mathbf{y})) \not\approx D(\mathbf{x}, \mathbf{y}) \quad (i \neq j) \tag{3}$$

where $D(\cdot, \cdot)$ denotes a relationship between data samples, such as distance and similarity. Therefore, one cannot analyze intermediate representations as one data set, even if dimensionality is identical, $\ell_i = \ell_j$.

To overcome this difficulty, the authors transform individual intermediate representations to incorporable representations again as follows:

$$\hat{\mathbf{X}}_i = [\hat{\mathbf{x}}_{i1}, \hat{\mathbf{x}}_{i2}, \ldots, \hat{\mathbf{x}}_{in_i}] = g_i(\tilde{\mathbf{X}}_i) \in \mathbb{R}^{\ell \times n_i} \tag{4}$$

Here, $g_i$ is a column-wise mapping function such that

$$g_i(f_i(\mathbf{x})) \approx g_j(f_j(\mathbf{x})) \tag{5}$$

$$D(g_i(f_i(\mathbf{x})), g_j(f_j(\mathbf{y}))) \approx D(\mathbf{x}, \mathbf{y}) \quad (i \neq j) \tag{6}$$

Preserving the relationships of the original data set, one can analyze the obtained data $\hat{\mathbf{X}}_i$ $(i = 1, 2, \ldots, d)$ as one data set as follows:

$$\hat{\mathbf{X}} = [\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2, \ldots, \hat{\mathbf{X}}_d] \in \mathbb{R}^{\ell \times n} \tag{7}$$

Because the mapping function $f_i$ for the intermediate representation is not centralized, the function $g_i$ cannot be constructed only from the centralized intermediate representations $\tilde{\mathbf{X}}_i$. To construct the mapping function $g_i$, we introduce sharable data referred to as an anchor data set consisting of public data or dummy data constructed randomly:

$$\mathbf{X}^{\text{anc}} = [\mathbf{x}_1^{\text{anc}}, \mathbf{x}_2^{\text{anc}}, \ldots, \mathbf{x}_r^{\text{anc}}] \in \mathbb{R}^{m \times r} \tag{8}$$

where $r \geq \ell_i$. Applying each mapping function $f_i$ to the anchor data, we have the $i$th intermediate representation of the anchor data set

$$\tilde{\mathbf{X}}_i^{\text{anc}} = [\tilde{\mathbf{x}}_{i1}^{\text{anc}}, \tilde{\mathbf{x}}_{i2}^{\text{anc}}, \ldots, \tilde{\mathbf{x}}_{ir}^{\text{anc}}] = f_i(\mathbf{X}^{\text{anc}}) \in \mathbb{R}^{\ell_i \times r} \tag{9}$$

Then, we centralize $\tilde{\mathbf{X}}_i^{\text{anc}}$ and construct $g_i$ such that

$$\hat{\mathbf{X}}_i^{\text{anc}} = [\hat{\mathbf{x}}_{i1}^{\text{anc}}, \hat{\mathbf{x}}_{i2}^{\text{anc}}, \ldots, \hat{\mathbf{x}}_{in_i}^{\text{anc}}] = g_i(\tilde{\mathbf{X}}_i^{\text{anc}}) \in \mathbb{R}^{\ell \times r} \tag{10}$$

satisfies

$$\hat{\mathbf{X}}_i^{\text{anc}} \approx \hat{\mathbf{X}}_j^{\text{anc}}, D(\hat{\mathbf{x}}_{ik}^{\text{anc}}, \hat{\mathbf{x}}_{jl}^{\text{anc}}) \approx D(\mathbf{x}_k^{\text{anc}}, \mathbf{x}_l^{\text{anc}}) \quad (i \neq j) \tag{11}$$

The fundamental procedure in the proposed data collaboration analysis framework is as follows:

1. *Construction of intermediate representations*
   Each institution constructs intermediate representations individually and centralizes them.
2. *Construction of collaboration representations*
   From the centralized intermediate representations, the collaboration representations are constructed.
3. *Collaboration analysis*
   Collaboration representations obtained from individual original data sets are analyzed as one data set.

### Proposal for Practical Algorithm

A fundamental component of the proposed framework involves constructing the collaboration representations using the anchor data (Phase 2). The mapping function $g_i$ can be constructed using the following two steps.

1. *Target setting*

   We set target $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_r] \in \mathbb{R}^{\ell \times r}$ for the collaboration representations $\hat{\mathbf{X}}_i^{\text{anc}}$ of the anchor data satisfying

$$\mathbf{Z} \approx \hat{\mathbf{X}}_i^{\text{anc}} \quad (i = 1, 2, \ldots, d) \tag{12}$$

   or

$$D(\mathbf{z}_k, \mathbf{z}_l) \approx D(\mathbf{x}_k^{\text{anc}}, \mathbf{x}_l^{\text{anc}}) \quad (k, l = 1, 2, \ldots, r) \tag{13}$$

2. *Map function construction*

   We construct mapping function $g_i$ such that

$$\mathbf{Z} \approx g_i(\tilde{\mathbf{X}}_i^{\text{anc}}) \quad (i = 1, 2, \ldots, d) \tag{14}$$

There may be several ways for computing Steps 1 and 2. This paper assumes $g_i$ to be a linear map. Considering only Eq. (12) for Step 1, we propose a practical algorithm.

Because the map function $g_i$ is a linear map, using a matrix $\mathbf{G}_i \in \mathbb{R}^{\ell \times \ell_i}$, we have

$$\hat{\mathbf{X}}_i = g_i(\tilde{\mathbf{X}}_i) = \mathbf{G}_i \tilde{\mathbf{X}}_i, \qquad \hat{\mathbf{X}}_i^{\text{anc}} = g_i(\tilde{\mathbf{X}}_i^{\text{anc}}) = \mathbf{G}_i \tilde{\mathbf{X}}_i^{\text{anc}} \tag{15}$$

Then, to achieve Eq. (12), we address the following minimization problem:

$$\min_{\mathbf{G}_1', \mathbf{G}_2', \ldots, \mathbf{G}_d', \mathbf{Z}} \sum_{i=1}^{d} \|\mathbf{Z} - \mathbf{G}_i' \tilde{\mathbf{X}}_i^{\text{anc}}\|_{\text{F}}^2 \tag{16}$$

This problem is difficult to solve directly. Instead, we consider solving the following minimal perturbation problem, i.e.

$$\min_{\mathbf{E}_i, \mathbf{G}_i' (i=1,2,\ldots,d), \mathbf{Z}} \sum_{i=1}^{d} \|\mathbf{E}_i\|_{\text{F}}^2 \quad \text{s.t.}\, \mathbf{G}_i'(\tilde{\mathbf{X}}_i^{\text{anc}} + \mathbf{E}_i) = \mathbf{Z} \tag{17}$$

The minimal perturbation problem Eq. (17) with $d = 2$ is called the total least squares problem and is solved by singular value decomposition (SVD) (Ito and Murota 2016). In the same manner, one can solve Eq. (17) with $d > 2$ using SVD. Let

$$\left[(\tilde{\mathbf{X}}_1^{\text{anc}})^{\text{T}}, (\tilde{\mathbf{X}}_2^{\text{anc}})^{\text{T}}, \ldots, (\tilde{\mathbf{X}}_d^{\text{anc}})^{\text{T}}\right] = [\mathbf{U}_1, \mathbf{U}_2] \begin{bmatrix} \Sigma_1 & & \\ & \Sigma_2 & \end{bmatrix} \begin{bmatrix} \mathbf{V}_{11}^{\text{T}} & \mathbf{V}_{21}^{\text{T}} & \cdots & \mathbf{V}_{d1}^{\text{T}} \\ \mathbf{V}_{12}^{\text{T}} & \mathbf{V}_{22}^{\text{T}} & \cdots & \mathbf{V}_{d2}^{\text{T}} \end{bmatrix} \tag{18}$$

be the SVD of the matrix combining $\tilde{\mathbf{X}}_i^{\text{anc}}$, where

$$\mathbf{U}_1 \in \mathbb{R}^{r \times \ell}, \qquad \Sigma_1 \in \mathbb{R}^{\ell \times \ell}, \qquad \mathbf{V}_{i1} \in \mathbb{R}^{\ell_i \times \ell} \tag{19}$$

and $\Sigma_1$ has larger part of singular values. Then, we have

$$\mathbf{Z} = \mathbf{C}\mathbf{U}_1^{\text{T}} \tag{20}$$

where $\mathbf{C} \in \mathbb{R}^{\ell \times \ell}$ is a nonsingular matrix.

Next, setting $\mathbf{Z} = \mathbf{U}_1^{\text{T}}$, we compute $\mathbf{G}_i$ from Eq. (14). The matrix $\mathbf{G}_i$ can be computed individually by solving the following linear least squares problem:

$$\mathbf{G}_i = \arg\min_{\mathbf{G}} \|\mathbf{Z} - \mathbf{G}\tilde{\mathbf{X}}_i^{\text{anc}}\|_{\text{F}}^2 = \mathbf{U}_1^{\text{T}}(\tilde{\mathbf{X}}_i^{\text{anc}})^{\dagger} \tag{21}$$

where $(\tilde{\mathbf{X}}_i^{\text{anc}})^{\dagger}$ denotes the Moore-Penrose pseudo-inverse of the matrix $\tilde{\mathbf{X}}_i^{\text{anc}}$.

Algorithm 1 summarizes the algorithm of the proposed method for supervised learning.

One of the main computational costs of the proposed method is for SVD (18) that depends on the number of anchor data $r$ and dimensionality of the intermediate representations $\ell_i$. We can use some approximation algorithms including randomized SVD (Halko et al. 2011) for reducing the computational costs. On the other hand, the anchor data $\mathbf{X}^{\text{anc}}$ also strongly affects the recognition performance of the proposed method. A simple method is to set $\mathbf{X}^{\text{anc}}$ as a random matrix. If the anchor data has the same statistics with the original data set, it may improve the recognition performance of the proposed method. We intend to investigate practical techniques for constructing suitable anchor data in the future.

**Algorithm 1.** Proposed method

**Input:** $\mathbf{X}_i \in \mathbb{R}^{m \times n_i}$, $\mathbf{L}_i \in \mathbb{R}^{l \times n_i}$, $\mathbf{Y}_i \in \mathbb{R}^{m \times s_i}$ $(i = 1, 2, \ldots, d)$ individually.

**Output:** $\mathbf{L}_{\mathbf{Y}_i} \in \mathbb{R}^{l \times s_i}$ $(i = 1, 2, \ldots, d)$ individually.

{Phase 0. Preparation}

1: Centralize $\mathbf{X}^{\text{anc}} \in \mathbb{R}^{m \times r}$

{Phase 1. Construction of intermediate representations}

2: Construct $\tilde{\mathbf{X}}_i = f_i(\mathbf{X}_i)$ and $\mathbf{X}_i^{\text{anc}} = f_i(\mathbf{X}^{\text{anc}})$ for each $i$ individually

3: Centralize $\tilde{\mathbf{X}}_i$, $\tilde{\mathbf{X}}_i^{\text{anc}}$, $\mathbf{L}_i$ for all $i$

{Phase 2. Construction of collaboration representations}

4: Compute left singular vectors $\mathbf{U}_1$ of SVD (18)

5: Compute $\mathbf{G}_i = \mathbf{U}_1^{\text{T}}(\tilde{\mathbf{X}}_i^{\text{anc}})^{\dagger}$

6: Compute $\hat{\mathbf{X}}_i = \mathbf{G}_i \tilde{\mathbf{X}}_i$

7: Set $\hat{\mathbf{X}} = [\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2, \ldots, \hat{\mathbf{X}}_d]$ and $\mathbf{L} = [\mathbf{L}_1, \mathbf{L}_2, \ldots, \mathbf{L}_d]$

{Phase 3. Collaboration analysis}

8: Construct model $h$ by a machine learning algorithm using $\hat{\mathbf{X}}$ as training date and $\mathbf{L}$ as the ground truth, i.e., $\mathbf{L} \approx h(\hat{\mathbf{X}})$.

9: Predict test data $\hat{\mathbf{Y}}_i$ using a model $h$ and obtain $\mathbf{L}_{\mathbf{Y}_i} = h(\mathbf{G}_i f_i(\mathbf{Y}_i))$.

### Practical Operation Strategy Regarding Privacy Concerns

Here, we consider a practical operation strategy regarding privacy concerns based on the proposed framework for supervised learning. This paper uses the term privacy is preserved when each entry of corresponding data cannot be (approximately) obtained by others. Here, this paper does not consider the privacy of data set statistics.

Based on this definition, one can assert that regarding the original data $\mathbf{X}_i$ in each institution, privacy is preserved if the data collaboration analysis satisfies the following operation strategies:

1. There are two roles: users who have training and test data sets individually and an analyst who centralizes the intermediate representations and analyzes them.
   a. The users and analyst possess some of the data, as illustrated in Tables 1 and 2.

**Table 1.** Practical operation strategy: data for each role

| Role | Data |
|---|---|
| User $i$ | $\mathbf{X}_i$, $\tilde{\mathbf{X}}_i$, $\mathbf{L}_i$, $\mathbf{X}^{\text{anc}}$, $\tilde{\mathbf{X}}_i^{\text{anc}}$, $\mathbf{Y}_i$, $\mathbf{L}_{\mathbf{Y}_i}$, $f_i$, $g_i$, $h$ |
| Analyst | $\tilde{\mathbf{X}}_i$, $\mathbf{L}_i$, $\tilde{\mathbf{X}}_i^{\text{anc}}$, $g_i$ $(i = 1, 2, \ldots, d)$, $h$ |

**Table 2.** Practical operation strategy: role for each data

| Data | Role |
|---|---|
| $\mathbf{X}_i$, $\mathbf{Y}_i$, $\mathbf{L}_{\mathbf{Y}_i}$, $f_i$ | User $i$ |
| $\widetilde{\mathbf{X}}_i$, $\mathbf{L}_i$, $\widetilde{\mathbf{X}}_i^{\text{anc}}$, $g_i$ | User $i$ and analyst |
| $\mathbf{X}^{\text{anc}}$ | All users |
| $h$ | All users and analyst |

b. Each step of Algorithm 1 is executed by the corresponding role, as demonstrated in Fig. 1.

2. Each mapping function $f_i$ is constructed with the following requirements:
   a. The original data can be approximated only with an intermediate representation and the mapping function $f_i$ or its approximation.
   b. The mapping function $f_i$ can be approximated only with both the input and output data of $f_i$.
3. The analyst does not collude with user(s) to obtain the original data of other users.

In this operation strategy, each user does not possess the intermediate representations of other users and the analyst does not possess the original anchor data $\mathbf{X}^{\text{anc}}$. Therefore, the original data set $\mathbf{X}_i$ cannot be (approximately) obtained by others; that proves the privacy of the original data $\mathbf{X}_i$ is preserved in our definition.

### *Related Works*

One possibility for achieving a high-quality analysis while avoiding the difficulties of centralized analysis involves the usage of privacy-preserving computation. There are two types of typical privacy-preserving computation techniques based on cryptography (Jha et al. 2005; Kerschbaum 2012; Cho et al. 2018; Gilad-Bachrach et al. 2016) and differential privacy (Abadi et al. 2016; Ji et al. 2014; Dwork 2006).

Cryptographic privacy-preserving (or secure multi-party) computations can compute a function over distributed data while retaining the privacy of the data. Fully homomorphic encryption (FHE) (Gentry 2009) can compute any given function; however, it is impractical for large data sets with respect to computational cost even using the latest implementations (Chillotti et al. 2016). Differential privacy is another type of privacy-preserving computation that protects the privacy of the original data sets by randomization. In terms of computational cost, these computations are more efficient than cryptographic computations; however, they may have low prediction accuracy because of the noise added for protecting privacy.

Federated learning, involving centralizing a model, has also been studied in this context (Konečnỳ et al. 2016; McMahan et al. 2016). Federated learning achieves a high-quality analysis avoiding the difficulties of centralized analysis by centralizing a model function instead of using cryptography or randomization. However, it may carry a risk of exposing the original data set as a result of centralizing a model for each institution. Therefore, in practice, federated learning is used in conjunction with privacy-preserving computations (Yang 2019).

Our proposed framework differs from these existing approaches as it does not use privacy-preserving computations or a model centralization.

## Numerical Experiments

This section presents an evaluation of the recognition performance of the proposed data collaboration analysis method and compares it with that of centralized and individual analyses for classification problems. In our target situation, it should be noted that centralized analysis is just ideal because one cannot share the original data sets $\mathbf{X}_i$. The proposed data collaboration analysis must achieve a recognition performance higher than that of individual analysis and lower, but similar to, that of centralized analysis.

We used kernel ridge regression (Saunders et al. 1998) for the individual and centralized analyses and Step 8 in the proposed method (Algorithm 1). In the proposed method, each intermediate representation is constructed from $\mathbf{X}_i$ by kernel LPP (K-LPP) (He and Niyogi 2004). We note that K-LPP is an unsupervised dimensionality reduction; however, the constructed map $f_i$ depends on $i$ because it depends on data set $\mathbf{X}_i$. The anchor data set is constructed as a random matrix.

In the training phase, we use the ground truth $\mathbf{L}$ as a binary matrix whose $(i, j)$ entry is 1 if the training data $\mathbf{x}_j$ is in class $i$. This type of ground truth $\mathbf{L}$ is used for several classification algorithms including ridge regression and deep neural networks (Bishop 2006). All numerical experiments were performed using MATLAB 2018b.

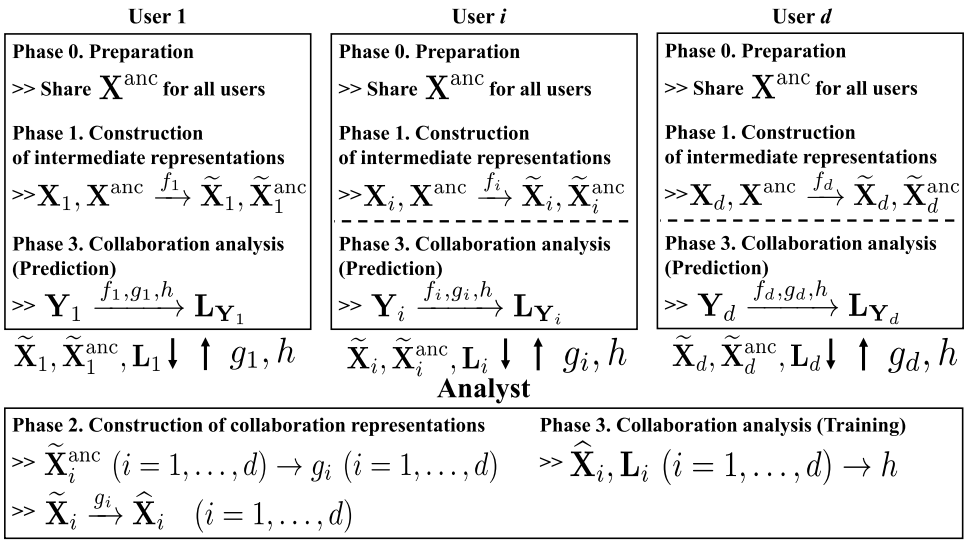


**Fig. 1.** Practical operation strategy: algorithm flow.



ASCE-ASME J. Risk Uncertainty Eng. Syst., Part A: Civ. Eng., 2020, 6(2): 04020018

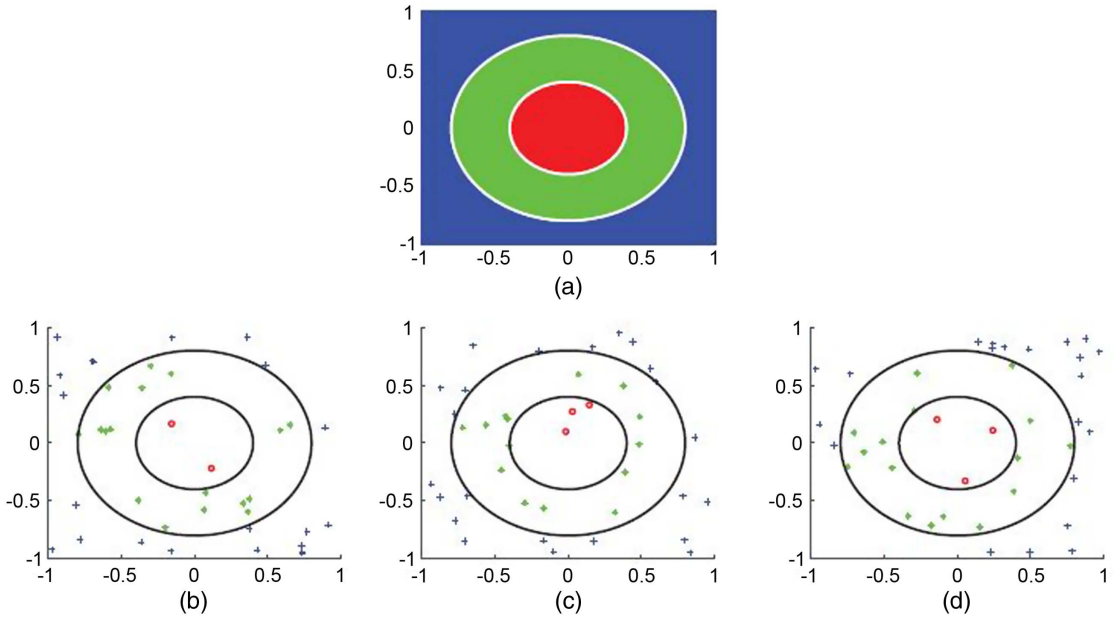


**Fig. 2.** Training data set and ground truth for artificial data: (a) training data set; (b) training data set in user 1; (c) training data set in user 2; and (d) training data set in user 3.

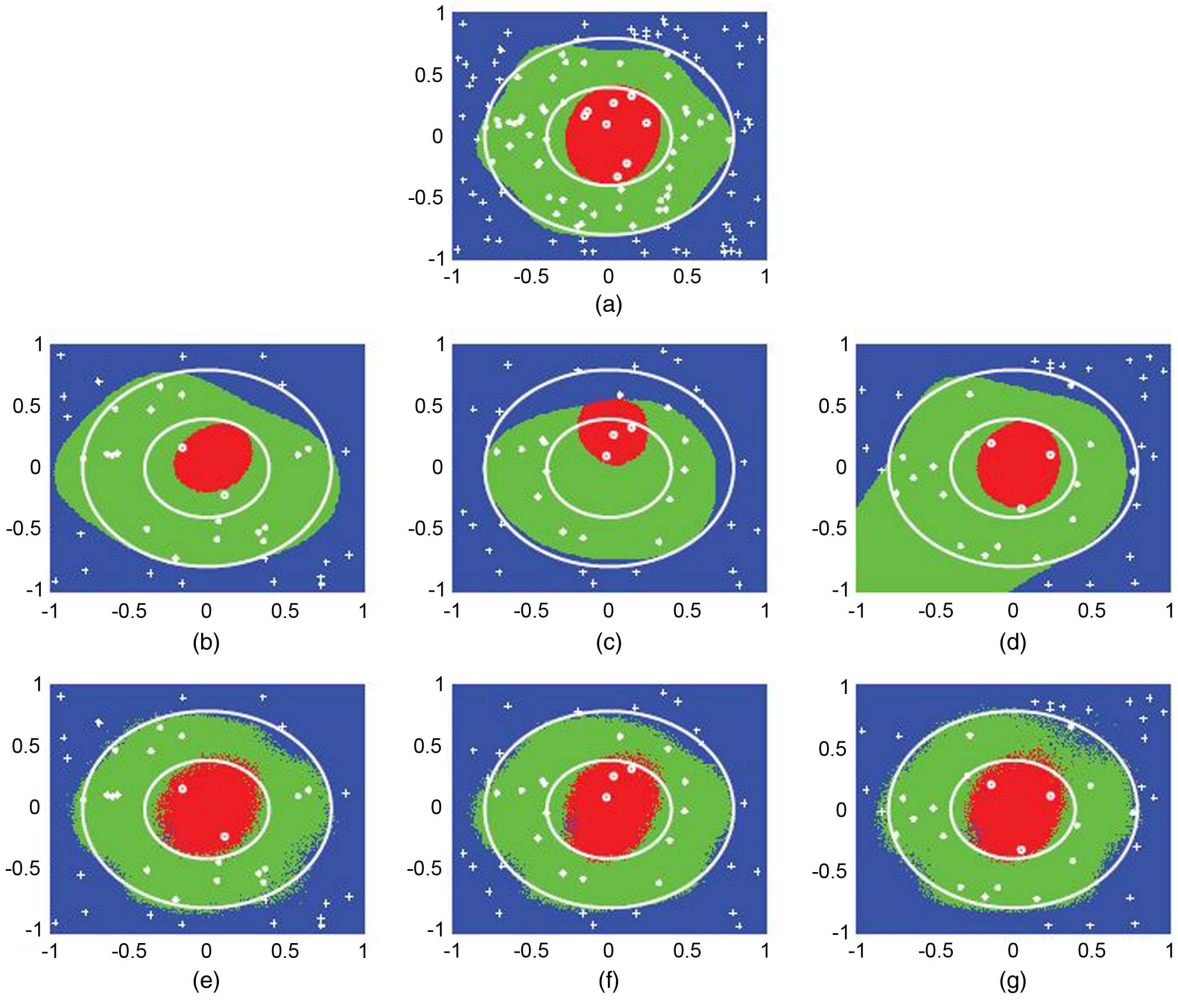


**Fig. 3.** Recognition results for artificial data: (a) centralized analysis; (b) individual analysis in user 1; (c) individual analysis in user 2; (d) individual analysis in user 3; (e) data collaboration analysis (user 1 has test data set); (f) data collaboration analysis (user 2 has test data set); and (g) data collaboration analysis (user 3 has test data set).



ASCE-ASME J. Risk Uncertainty Eng. Syst., Part A: Civ. Eng., 2020, 6(2): 04020018

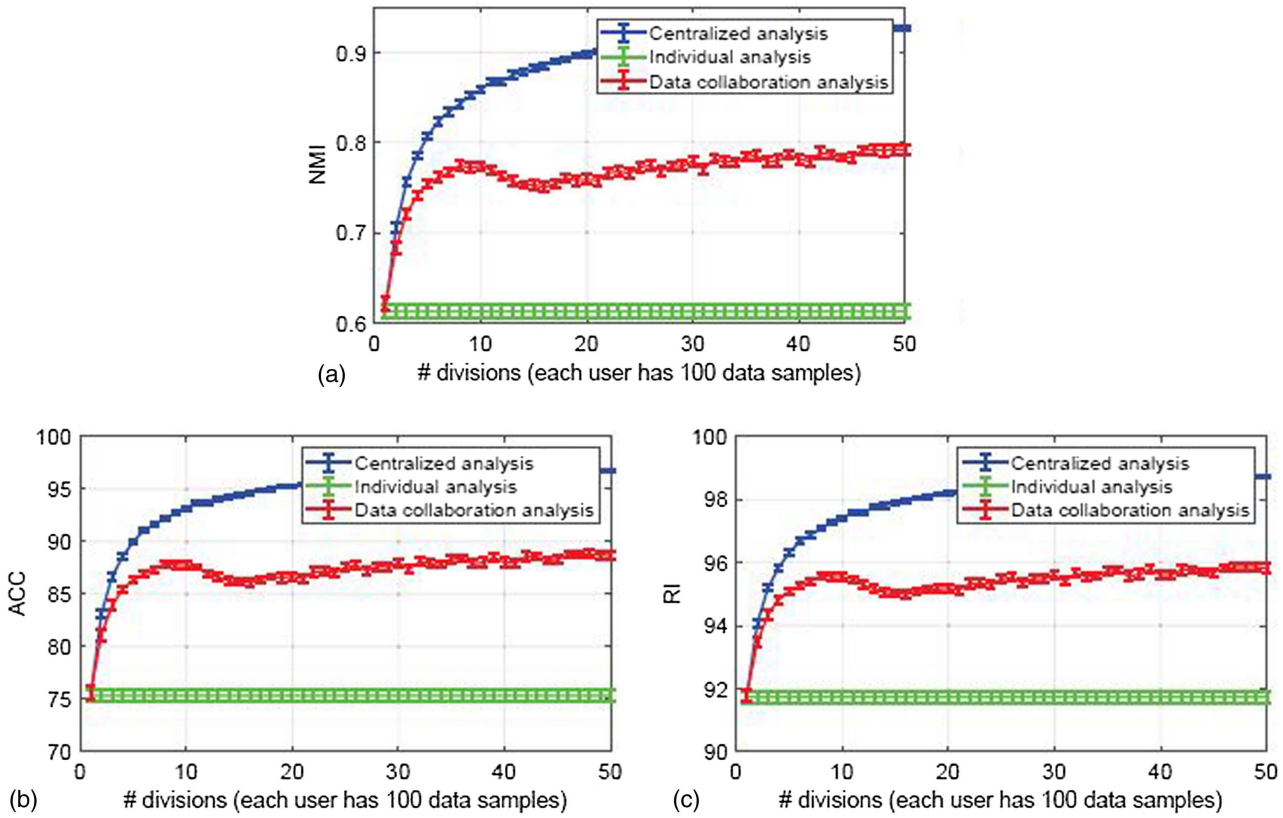


**Fig. 4.** Recognition performance for MNIST: (a) NMI; (b) ACC; and (c) RI.

### *Artificial Data*

In this experiment, we used a three-class classification of 10-dimensional artificial data. Fig. 2(a) illustrates the first two dimensions of the ground truth. Figs. 2(b–d) illustrate 40 training data points in each user of the first two dimensions with the corresponding labels: $\circ$, $\bullet$, and $+$. For the test data set, we used $201 \times 201$ data points whose first two dimensions were square grid points in $[-1, 1] \times [-1, 1]$. The remaining eight dimensions of the training and test data sets were random values in $[-0.1, 0.1]$ generated by the Mersenne Twister. The Gaussian kernel was used for all methods.

The accuracy (ACC) of centralized analysis and the average ACC of three users of individual and proposed data collaboration analyses are 92.3, 79.8, and 91.3. Fig. 3 presents the recognition results. In each subfigure, white markers: $\circ$, $\bullet$, and $+$, denote training data points. From the comparison between the results of centralized and individual analyses, we observed that the recognition results of individual analysis are significantly poorer than those of centralized analysis because of the insufficiency of data samples. In contrast, the proposed data collaboration analysis achieves results comparable to those of centralized analysis.

### *Handwritten Digits Data (MNIST)*

In this experiment, we used a 10-class classification of handwritten digits (MNIST) (LeCun 1998), where the number of features was $m = 784$. Here, we set 100 data samples for each user and evaluated the recognition performance, normalized mutual information (NMI) (Strehl and Ghosh 2002), accuracy (ACC), rand index (RI) (Rand 1971), for 1,000 test data samples, increasing the number of users from 1 to 50. We used the Gaussian kernel for all methods.

Fig. 4 presents the average and standard error of the recognition performance for 20 trials for each method. It can be seen that the recognition performance of the proposed data collaboration analysis increases with an increasing number of users and achieves a significantly higher recognition performance than individual analysis.

### *Gene Expression Data*

In this numerical experiment, we used a three-class classification problem for cancer data from a previous study (Golub et al. 1999). The data set has 38 training and 34 test data samples with $m = 7{,}129$ features. Here, we considered the case of two users and allocated 19 data samples for each user. Then, we evaluated the recognition performance for 20 trials. A linear kernel was used for all methods.

Fig. 5 presents a three-dimensional visualization of the training $+$ and test $\circ$ data samples for each method. Table 3 summarizes the recognition performance (average $\pm$ standard error). In three-dimensional visualization, three classes are well separated in low-dimensional space constructed by the proposed data collaboration analysis as well as centralized analysis. We observed that the proposed data collaboration analysis achieved higher recognition performance than individual analysis for real-world problems.

### *Remarks of Numerical Results*

The results of numerical experiments reveal that the proposed data collaboration analysis achieves higher recognition performance for artificial and real-world data sets than individual analysis. It should be noted that because centralized analysis is ideal, the recognition

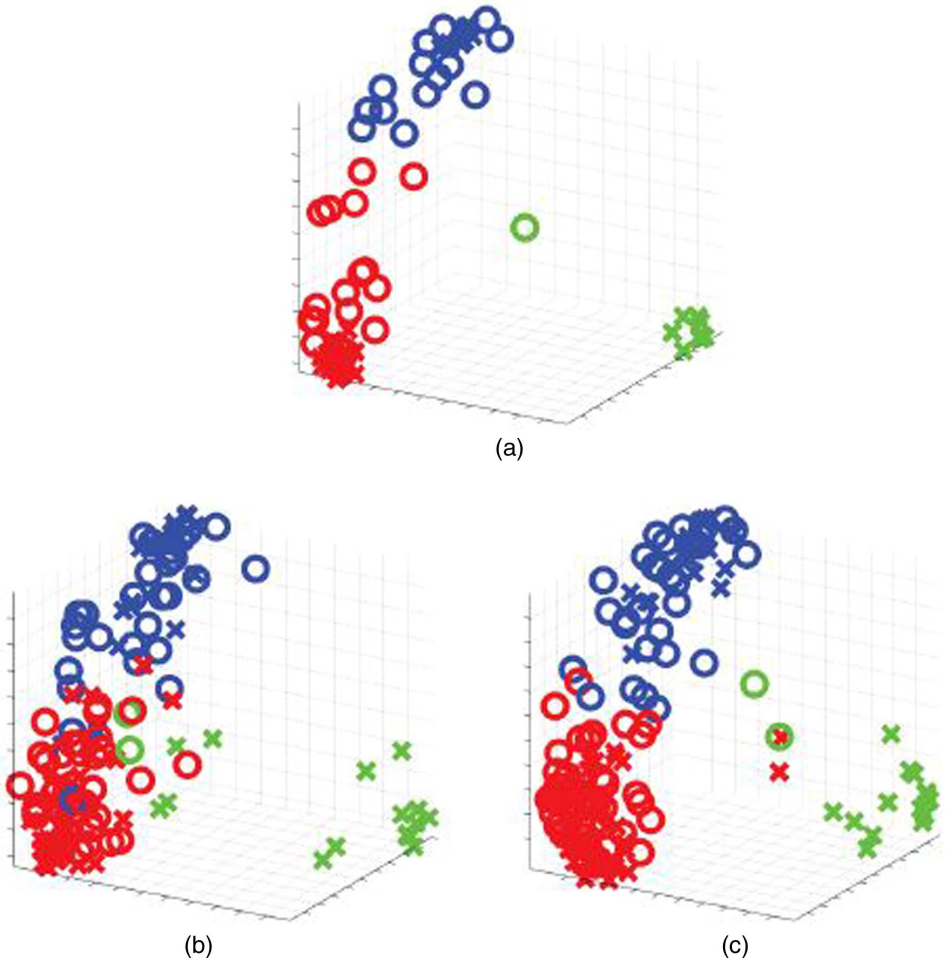


**Fig. 5.** Three-dimensional visualization for gene expression data: (a) centralized analysis; (b) individual analysis; and (c) data collaboration analysis.

**Table 3.** Recognition performance for gene expression data

| Method | NMI | ACC | RI |
|---|---|---|---|
| Centralized | $0.92 \pm 0.00$ | $97.1 \pm 0.00$ | $96.6 \pm 0.00$ |
| Individual | $0.48 \pm 0.04$ | $83.7 \pm 1.52$ | $74.2 \pm 2.29$ |
| Data collaboration | $0.76 \pm 0.03$ | $93.1 \pm 0.91$ | $89.3 \pm 1.68$ |

performance of the proposed data collaboration analysis is not required to be higher than that of centralized analysis.

## Conclusions

This paper has proposed a data collaboration analysis framework for distributed data sets based on centralizing individual intermediate representations, while the original data sets and mapping functions remain distributed. This paper has also proposed a practical algorithm within the framework and a practical operation strategy regarding privacy concerns. The proposed framework differs from existing approaches in that it does not use privacy-preserving computations and does not centralize mapping functions. Numerical experiments demonstrate that the proposed method achieves higher recognition performance for artificial and real-world data sets than individual analysis.

In future works, we will investigate the usage of a nonlinear mapping function $g_i$ and how to set anchor data to improve recognition performance for large real-world problems.

## Data Availability Statement

Some or all data, models, code-generated or used during the study are available from the corresponding author by request. Available items: program codes, data sets used in the numerical experiments.

## Acknowledgments

The present study is supported in part by the Japan Science and Technology Agency (JST), ACT-I (No. JPMJPR16U6), the New Energy and Industrial Technology Development Organization (NEDO) and the Japan Society for the Promotion of Science (JSPS), Grants-in-Aid for Scientific Research (Nos. 17K12690 and 18H03250).

## References

Abadi, M., A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. 2016. "Deep learning with differential privacy." In *Proc., 2016 ACM SIGSAC Conf. on Computer and Communications Security*, 308–318. New York: Association for Computing Machinery.

Bishop, C. M. 2006. *Pattern recognition and machine learning (Information science and statistics)*. Berlin: Springer.

Chillotti, I., N. Gama, M. Georgieva, and M. Izabachene. 2016. "Faster fully homomorphic encryption: Bootstrapping in less than 0.1 seconds." In *Proc., Int. Conf. on the Theory and Application of Cryptology and Information Security*, 3–33. Berlin: Springer.

Cho, H., D. J. Wu, and B. Berger. 2018. "Secure genome-wide association analysis using multiparty computation." *Nat. Biotechnol.* 36 (6): 547. https://doi.org/10.1038/nbt.4108.

Dwork, C. 2006. "Differential privacy." In Vol. 4052 of *Automata, Languages and Programming. ICALP 2006. Lecture Notes in Computer Science*, edited by M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener. Berlin: Springer.

Fisher, R. A. 1936. "The use of multiple measurements in taxonomic problems." *Ann. Hum. Genet.* 7 (2): 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

Gentry, C. 2009. "Fully homomorphic encryption using ideal lattices." In Vol. 9 of *Proc., 41 Annual ACM Symp. on Theory of Computing*, 169–178. New York: Association for Computing Machinery.

Gilad-Bachrach, R., N. Dowlin, K. Laine, K. Lauter, M. Naehrig, and J. Wernsing. 2016. "Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy." In *Proc., Int. Conf. on Machine Learning*, 201–210. Washington, DC: American Association for the Advancement of Science.

Golub, T. R., et al. 1999. "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring." *Science* 286 (5439): 531–537. https://doi.org/10.1126/science.286.5439.531.

Halko, N., P. G. Martinsson, and J. A. Tropp. 2011. "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions." *SIAM Rev.* 53 (2): 217–288. https://doi.org/10.1137/090771806.

He, X., and P. Niyogi. 2004. "Locality preserving projections." In *Proc., Advances in Neural Information Processing Systems*, 153–160. London: MIT Press.

Imakura, A., M. Matsuda, X. Ye, and T. Sakurai. 2019. "Complex moment-based supervised eigenmap for dimensionality reduction." In Vol. 33 of *Proc., 33rd AAAI Conf. on Artificial Intelligence (AAAI-19)*, 3910–3918. Palo Alto, CA: AAAI Press.

Ito, S., and K. Murota. 2016. "An algorithm for the generalized eigenvalue problem for nonsquare matrix pencils by minimal perturbation approach." *SIAM J. Matrix Anal. Appl.* 37 (1): 409–419. https://doi.org/10.1137/14099231X.

Jha, S., L. Kruger, and P. McDaniel. 2005. "Privacy preserving clustering." In *European Symp. on Research in Computer Security*, 397–417. Berlin: Springer.

Ji, Z., Z. C. Lipton, and C. Elkan. 2014. "Differential privacy and machine learning: A survey and review." Preprint, submitted December 24, 2014. https://arxiv.org/abs/1412.7584.

Jolliffe, I. T. 1986. "Principal component analysis and factor analysis." In *Principal component analysis*, 115–128. New York: Springer.

Jurs, P. C., G. A. Bakken, and H. E. McClelland. 2000. "Computational methods for the analysis of chemical sensor array data from volatile analytes." *Chem. Rev.* 100 (7): 2649–2678. https://doi.org/10.1021/cr9800964.

Kerschbaum, F. 2012. "Privacy-preserving computation." In *Annual Privacy Forum*, 41–54. Berlin: Springer.

Konečnỳ, J., H. B. McMahan, F. X. Yu, P. Richtarik, A. T. Suresh, and D. Bacon. 2016. "Federated learning: Strategies for improving communication efficiency." Preprint, submitted October 18, 2016. http://arxiv.org/abs/1610.05492.

Lasisi, A., and N. Attoh-Okine. 2018. "Principal components analysis and track quality index: A machine learning approach." *Transp. Res. Part C: Emerg. Technol.* 91 (Jun): 230–248. https://doi.org/10.1016/j.trc.2018.04.001.

Lasisi, A., and N. Attoh-Okine. 2020. "An unsupervised learning framework for track quality index and safety." *Transp. Infrastruct. Geotechnol.* 7 (1): 1–12. https://doi.org/10.1007/s40515-019-00087-6.

LeCun, Y. 1998. "The MNIST database of handwritten digits." Accessed January 15, 2019. http://yann.lecun.com/exdb/mnist/.

Li, X., M. Chen, F. Nie, and Q. Wang. 2017. "Locality adaptive discriminant analysis." In *Proc., 26th Int. Joint Conf. on Artificial Intelligence*, 2201–2207. Palo Alto, CA: AAAI Press.

McMahan, H. B., E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas. 2016. "Communication-efficient learning of deep networks from decentralized data." Preprint, submitted February 17, 2016. https://arxiv.org/abs/1602.05629.

Pearson, K. 1901. "LIII. On lines and planes of closest fit to systems of points in space." *London, Edinburgh, Dublin Philos. Mag. J. Sci.* 2 (11): 559–572. https://doi.org/10.1080/14786440109462720.

Rand, W. M. 1971. "Objective criteria for the evaluation of clustering methods." *J. Am. Stat. Assoc.* 66 (336): 846–850. https://doi.org/10.1080/01621459.1971.10482356.

Saunders, C., A. Gammerman, and V. Vovk. 1998. "Ridge regression learning algorithm in dual variables." In *Proc., 15th Int. Conf. on Machine Learning (ICML'98)*, 515–521. Burlington, MA: Morgan Kaufmann Publishers.

Strehl, A., and J. Ghosh. 2002. "Cluster ensembles—A knowledge reuse framework for combining multiple partitions." *J. Mach. Learn. Res.* 3: 583–617. https://doi.org/10.1162/153244303321897735.

Sugiyama, M. 2007. "Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis." *J. Mach. Learn. Res.* 8: 1027–1061.

Sugiyama, M., T. Idé, S. Nakajima, and J. Sese. 2010. "Semi-supervised local Fisher discriminant analysis for dimensionality reduction." *Mach. Learn.* 78 (1–2): 35. https://doi.org/10.1007/s10994-009-5125-7.

Tarca, A. L., R. Romero, and S. Draghici. 2006. "Analysis of microarray experiments of gene expression profiling." *Am. J. Obstetrics Gynecol.* 195 (2): 373–388. https://doi.org/10.1016/j.ajog.2006.07.001.

Tichy, N. M., M. L. Tushman, and C. Fombrun. 1979. "Social network analysis for organizations." *Acad. Manage. Rev.* 4 (4): 507–519. https://doi.org/10.5465/amr.1979.4498309.

Yang, Q. 2019. "GDPR, data shortage and AI." In *Invited Talk of the 33rd AAAI Conf. on Artificial Intelligence (AAAI-19)*. Palo Alto, CA: AAAI Press.