SN Computer Science (2021) 2:239
https://doi.org/10.1007/s42979-021-00564-1

ORIGINAL RESEARCH

Decentralized Learning with Virtual Patients for Medical Diagnosis of Diabetes

Yuta Takahashi[2] · Han-ten Chang[3] · Akie Nakai[3] · Rina Kagawa[4] · Hiroyasu Ando[1,5] · Akira Imakura[1,2] · Yukihiko Okada[1,2] · Hideo Tsurushima[4] · Kenji Suzuki[1,2] · Tetsuya Sakurai[1,2]

Received: 8 September 2020 / Accepted: 4 March 2021 / Published online: 25 April 2021

Abstract

Machine learning, applied to medical data, can uncover new knowledge and support medical practices. However, analyzing medical data by machine learning methods presents a trade-off between accuracy and privacy. To overcome the trade-off, we apply the data collaboration analysis method to medical data. This method using artificial dummy data enables analysis to compare distributed information without using the original data. The purpose of our experiment is to identify patients diagnosed with diabetes mellitus (DM), using 29,802 instances of real data obtained from the University of Tsukuba Hospital between 01/03/2013 and 30/09/2018. The whole data is divided into a number of datasets to simulate different hospitals. We propose the following improvements for the data collaboration analysis. (1) Making the dummy data which has a reality and (2) using non-linear reconverting functions into the comparable space. Both can be realized using the generative adversarial network (GAN) and Node2Vec, respectively. The improvement effects of dummy data with GAN scores more than 10% over the effects of dummy data with random numbers. Furthermore, the improvement effect of the re-conversion by Node2Vec with GAN anchor data scores about 20% higher than the linear method with random dummy data. Our results reveal that the data collaboration method with appropriate modifications, depending on data type, improves analysis performance.

Keywords Medical data · Machine learning · Data collaboration · Generative adversarial network

Introduction

Medical big data is increasingly used for improving healthcare quality and clinical research, such as clinical decision support systems [3, 5, 44, 50], identifying patients for clinical trials [36], and post-marketing surveillance of drugs [31, 51]. While machine learning is one of the critical techniques for analyzing medical data [5, 6, 34, 40, 43, 45], patients' privacy must be protected in the learning process. As machine-learning methods for privacy protection, encryption [8, 15, 29, 32], differential privacy [1, 12, 30], and federated learning [33, 38] are well known. However, encryption requires a huge computational cost [7], and the accuracy of analysis for differential privacy tends to be low as protection becomes strong [35]. On the other hand, federated learning using accumulated encrypted models from institutions not only provides good analysis accuracy, it is also difficult to estimate the original data. However, the risk of information leakage tends to be high when it contains personal information.

Recently, a method of data collaboration analysis for distributed data among institutions has been proposed as a secured method absent data encryption [4, 26–28, 54]. The method converts raw data to "intermediate representations (IRs)" at each institution by feature extraction, namely some information in raw data is reduced. Therefore, it is impossible to estimate the original information from the IRs. The

Yuta Takahashi and Han-ten Chang contributed equally to this work.

✉ Tetsuya Sakurai
sakurai@cs.tsukuba.ac.jp

1 Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba, Japan

2 Center for Artificial Intelligence Research, University of Tsukuba, Tsukuba, Japan

3 Master's Program in Service Engineering, University of Tsukuba, Tsukuba, Japan

4 Faculty of Medicine, University of Tsukuba, Tsukuba, Japan

5 Advanced Institute for Materials Research, Tohoku University, Sendai, Japan

IRs gathered from all institutions are able to be integrally analyzed without reconverting to the original data. Since each institution generates the IR independently, it is impossible to compare the representation as the original data is compared. Therefore, the data collaboration method learns further transformations to make the IRs comparable. This comparable transformation can be derived by sharing common dummy data among institutions. The data collaboration method should be appropriate for analyzing medical data since small size of data distributed among hospitals (e.g., those with regional characteristics or rare diseases) cannot be shared as in the original form. However, it is possible to integrate and analyze the distributed data by using the method. Thus, the hospitals can make a diagnosis based on information obtained from other institutions.

In this study, the data collaboration analysis method is applied to medical data for identifying patients diagnosed with diabetes mellitus. One of the most useful pieces of information for the secondary use of medical data is diagnosis. Since structured data on diagnoses are limited in terms of accuracy and completeness [14, 37, 53], automated techniques for identifying patients diagnosed with a particular disease based on medical data have increased [11, 21, 22, 41, 52]. Our aim is to show that the data collaboration method identifies patients diagnosed with a particular disease accurately while protecting privacy. The main contributions of this paper include the following items. (1) Application of the data collaboration method to real medical data. (2) Clarification of the influences of anchor data similar to raw data on the classification of disease. (3) Improvement of the classification accuracy by non-linearity of transformation.

The remainder of this paper is organized as follows. Section 2 introduces the data collaboration method in detail and how to apply it to the medical data is illustrated in Sect. 3. In Sect. 4, an explanation of medical data and experimental settings are given. The results of data collaboration analysis are shown in Sect. 5. Finally, we discuss the results and conclude this study in Sect. 6.

Data Collaboration Analysis

For analyzing distributed data remaining the original datasets, data collaboration analysis method was originally proposed by Imakura and Sakurai (2019) [24–27] as non-model share-type federated learning systems and was developed for classification and regression problems [28] and feature selection [54]. The performance comparison between model share-type and non-model share-type federated learnings was also reported in [4]. The data collaboration method only centralizes so-called intermediate representations constructed individually instead of the original datasets. The algorithm of the data collaboration method comprises the following three-step algorithm. (1) Each institution constructs intermediate representations from raw data individually and send them to an analyst, called data collaborator. (2) From the gathered intermediate representations, the collaboration representations are constructed. (3) Collaboration representations integrated from individual original datasets are analyzed as one dataset.

Here, we briefly introduce the practical algorithm. The m-dimensional data of d institutions, X_i, are described as follows:

$$X_i = [x_{i1}, x_{i2}, ..., x_{in_i}] \in \mathbb{R}^{m \times n_i} \quad (1 \leq i \leq d), \tag{1}$$

where n_i indicates the amount of data in the i th institution, and

$$\sum_i n_i = n. \tag{2}$$

Each institution independently constructs the intermediate representation, $\tilde{X}_i$, by a map, f_i, such that

$$\tilde{X}_i = f_i(X_i) \in \mathbb{R}^{\tilde{m}_i \times n_i}. \tag{3}$$

Note that $\tilde{m}_i$ are not require to be the same. As the map function, f_i, the dimensionality reduction method, including principal component analysis, independent component analysis [23], local linear embedding [46], and locality preserving protections (LPP) [20] are considered.

Here, because f_i depends on the institution, i, the intermediate representation of the data differs $f_i(x) \neq f_j(x)$ $(i \neq j)$. In this case, we cannot combine the intermediate representations to analyze one dataset. To overcome this difficulty, the intermediate representations are transformed again to the collaboration representation, $\hat{X}_i = g_i(\tilde{X}_i) \in \mathbb{R}^{\hat{m} \times n_i}$, with the function, g_i, satisfying

$$g_i(f_i(x)) \approx g_j(f_j(x)) \quad (i \neq j). \tag{4}$$

Note that $\hat{X}_i$ is not an approximation of X_i. The dimensions of X_i and $\tilde{X}_i$ can differ. Instead of the intermediate representation, one can analyze the collaboration representation as one dataset, as follows:

$$\hat{X} = [\hat{X}_1, \hat{X}_2, ..., \hat{X}_d] \in \mathbb{R}^{\hat{m} \times n}, \tag{5}$$

To construct the map, g_i, we introduce shareable data, referred to as an anchor dataset, comprising public data or pseudo-data constructed randomly as follows:

$$X^{\mathrm{anc}} = [x_1^{\mathrm{anc}}, x_2^{\mathrm{anc}}, ..., x_r^{\mathrm{anc}}] \in \mathbb{R}^{m \times r}, \tag{6}$$

where r indicates the amount of anchor data. Applying each map, f_i, to the anchor data, we have the ith intermediate representation of the anchor dataset,

$$\tilde{X}_i^{\text{anc}} = f_i(X^{\text{anc}}) \in \mathbb{R}^{\tilde{m}_i \times r}. \tag{7}$$

Then, we share $\tilde{X}_i^{\text{anc}}$ and construct g_i, satisfying

$$\hat{X}_i^{\text{anc}} \approx \hat{X}_j^{\text{anc}}, \hat{X}_i^{\text{anc}} = g_i(\tilde{X}_i^{\text{anc}}). \tag{8}$$

Imakura and Sakurai (2019) [26] introduce a practical method for constructing g_i when g_i is linear. In this situation, the function, g_i, can be computed by solving the minimization problem,

$$\min_{g_1, g_2, \ldots, g_d} \sum_{i=1}^{d} \|Z - g_i(\tilde{X}_i^{\text{anc}})\|_{\text{F}}^2 \tag{9}$$

where $Z = [z_1, z_2, \ldots, z_r] \in \mathbb{R}^{\hat{m} \times r}$ is a target for the collaboration representations, $\hat{X}_i^{\text{anc}}$. For the details, we refer to [26, 27].

Data Collaboration for Medical Data

Overview

Figure 1 provides an overview of the proposed method. First, the data collaborator constructs the virtual data generator denoted by G and distributes it to each hospital or medical institution. By using the generator G, institution i obtains the virtual patient data, corresponding to the anchor data, X_i^{anc} (Step1). For privacy protection, the institutions only share a random number seed to generate the virtual data. Second, each institution can arbitrarily select a map by which raw data X_i and virtual data X_i^{anc} with dimension M, are converted to an intermediate representation (Step2). The intermediate representation is in the form of extracted feature, so that the dimension is usually reduced from the original data. Thus, the privacy problem is resolved, since it is impossible to estimate the original data from the representation. Regarding the dimension of intermediate representation, each hospital has its own dimension (denoted by $\tilde{M}_1$ and $\tilde{M}_2$ in Fig. 1) due to the difference of the strategy and regulation for sharing medical data. Next, the data collaborator gathers the intermediate representations and constructs the reconverting function, g_i, based on the intermediate representations of anchor data, $\tilde{X}^{\text{anc}}$. Finally, the collaboration representations with dimension K, denoted by $\hat{X}_i$, are obtained via the reconverting function g_i, and they are applied to the machine learning method as input values (Step4). In this study, the classification of patients diagnosed with diabetes mellitus (DM) was carried out in terms of social importance. That is to say, more than 425 million people worldwide were estimated to have DM [13], and the problem can cause other critical diseases [50].

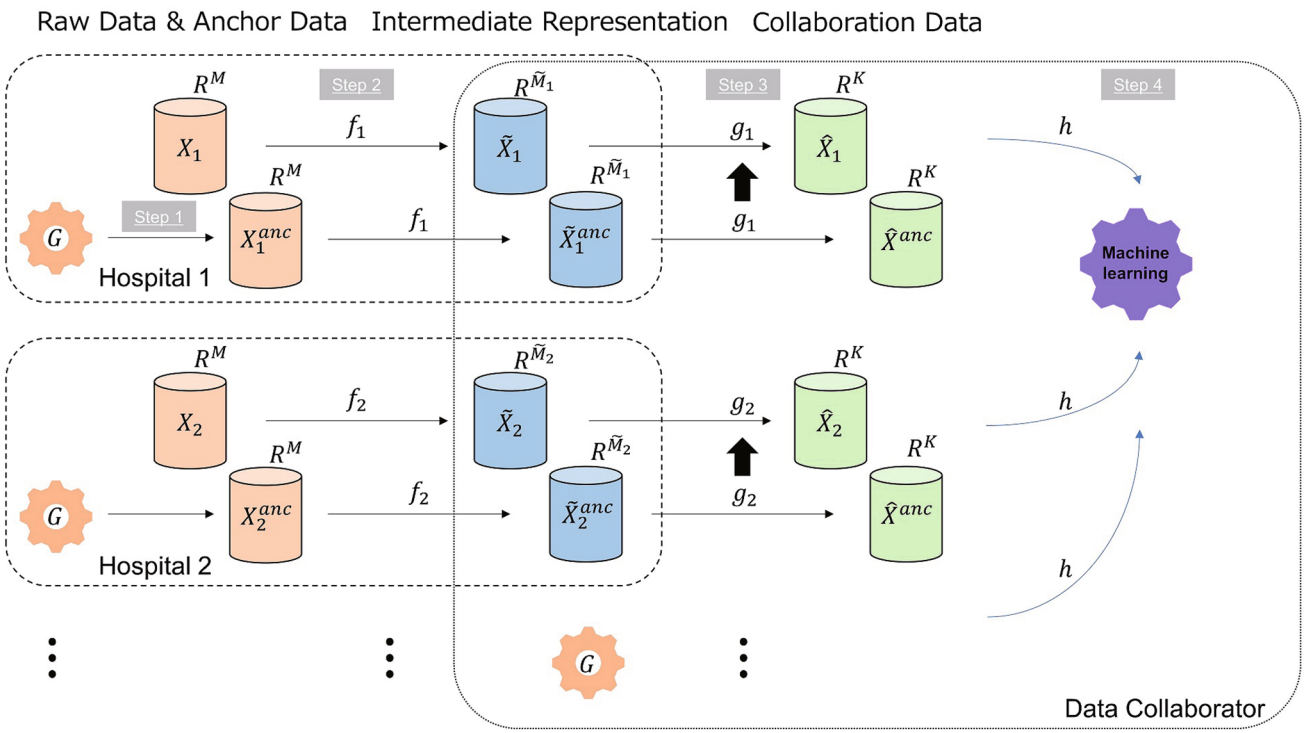

Fig. 1 Overview of this study

Generating the Intermediate Representation by LPP

Regarding the map f, this study uses LPP [20], which presents low computational costs. LPP is a linear approximation of the nonlinear Laplacian eigenmap. The algorithm has three steps: (1) Constructing the adjacency matrix by k-nearest neighbor (KNN) [2]. In this study, k=5. (2) Choosing the symmetric $m \times m$ weight matrix by calculating the weight as:

$$W_{ij} = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{t}}, \tag{10}$$

where W_{ij} is the weight between node i and j. $t \in \mathbf{R}$ is a parameter. (3) Calculating Eigenmaps generalized eigenvector problem, the eigenvectors and eigenvalues are calculated as follows:

$$XLX^T\mathbf{a} = \lambda XDX^T\mathbf{a}, \tag{11}$$

where D is a diagonal matrix whose entries are column sums of W. $L = D - W$ is the Laplacian matrix. Finally, $\mathbf{a}$ determines f.

Node2vec for Reconverting Function

The previous study [26] used SVD as a linear reconverting function g. As an alternative reconverting function, Node2Vec, which is non-linear, was exploited [19] in this paper. Node2Vec is a graph-embedding [18],[?] and network-embedding method [10], which converts the graph and network structures to a vector. Specifically, both DeepWalk [42] and Node2Vec [19] estimate the vector representation from the graph structure. The methods were based on a skip-gram model [39]. Node2Vec uses random walks, which is the sequence of nodes sampled from the edge of the graph [10]. We consider the weight matrix, $\tilde{W} \in \mathbb{R}^{r\times r}$, which is obtained by integration of the KNN adjacency matrix, $\tilde{W}_i \in \mathbb{R}^{r\times r}$. In this study, $\tilde{W}$ was the weighted summation of $\tilde{W}_i$ in order to maintain the relation of KNN after reconversion. Therefore, $\tilde{W}$ is defined as follows:

$$\tilde{W} = \sum_{i}^{d} w_i \tilde{W}_i. \tag{12}$$

The uth node for the graph related to $\tilde{W}$ is $\tilde{v}_u$. The initial node selected randomly is represented by c_0, and the jth node of the random walks is represented by c_j. Thus, c_j is sampled by the distribution, as shown below:

$$\Pr(c_j = \tilde{v}_t | c_{j-1} = \tilde{v}_s) = \begin{cases} \frac{\pi_{st}}{C} & \text{if } \tilde{W}_{st} > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

C is a normalizing constant. π_{st} is the transition probability form the node s to t, and $\pi_{st} = \alpha_{pq}(u, s)\tilde{W}_{st}$, where $u = c_{j-2}$ and α is defined by

$$\alpha_{pq}(u,t) = \begin{cases} \frac{1}{p} & \text{if } d_{ut} = 0, \\ 1 & \text{if } d_{ut} = 1, \\ \frac{1}{q} & \text{if } d_{ut} = 2, \end{cases} \tag{14}$$

where d_{ut} is the number of nodes in the shortest path from the node u to t. p is called the return parameter representing the probability to return to the original node, and q is called the in-out parameter representing the probability to leave the node u. Let l be the length of the random walk, and $c_0, c_1, c_2, ..., c_l$ is obtained. With a sufficient number of random walks applying to the skip-gram model, the vector representation of node $\tilde{v}_u$, is obtained. The integrated vectors is denoted by $Z = \mathbb{R}^{\hat{m}\times r}$, where $\hat{m}$ can be selected freely. In this study, $p = 0.5$ and $q = 1$ are used. The number of random walks is 1, 000, and the walk length is 500.

Experimental Settings

We perform numerical experiments with real medical data for verifying the following two tasks.

- Task 1: Effect of the similarity to raw data on classification score is verified.
- Task 2: Non-linearity of the reconverting function for classification scores are verified.

Task 1 is carried out with the reconverting function being linear or non-linear, and Task 2 is carried out with anchor data generated from random number or virtual patients data.

Subjects and Anchor Data

This study analyzed 29,802 patients (mean age: 59.9; gender: 50.3% female) testing HbA1c and random blood glucose. They were hospitalized at least once at the University of Tsukuba Hospital between 01/03/2013 and 30/09/2018. Based on Ethical Guidelines for Medical and Health Research Involving Human Subjects, our research is carried out with opt-out consent and is also approved by the Ethics Committee of the University of Tsukuba Hospital (Permission number: H30-187). We used only the maximum value of glucose and HbA1c for each patient. The other basic statistics are shown in Table 1.

Table 1 The statistics of raw data and 1000 virtual patients

	Mean	SD	Max	Min
(a) Statistics of raw dataset				
Age	59.9	18.5	103	0
Female (%)	50.3	/	/	/
Glucose (mg/dL)	154.1	84.0	1802	29
HbA1c (%)	6.23	1.37	19.1	3.5
(b) 1,000 virtual patient dataset				
Age	59.1	19.2	85.0	2.0
Female (%)	52.2	/	/	/
Glucose (mg/dL)	153.2	78.1	617	50
HbA1c (%)	6.48	1.72	14.25	3.69

Table 2 Mean of earth-movers' distance between raw and target data

	RANDOM	GAN	RAW
EMD	455.4	22.1	18.5

The number of anchor data are 1,000, generated by three methods. The first one was generated by a random number that was limited by the max and min value of original data. This type of anchor data was used in previous studies [26], and the data adjusted to statistics properties of real data were applied to machine learning in medical situations [49]. The second one was generated by GAN [9, 16, 17]. GAN generates similar data to the raw data in terms of statistical distribution. The 1,000 patients raw data randomly selected is used for GAN, and 1,000 virtual patients are obtained as the anchor data. The statistical value of anchor data generated by GAN is shown in Table 1. The third one is a part of the raw data that is selected randomly as anchor data for verification.

Similarity of Virtual Data (Anchor Data) to Raw Data

Anchor data similar to raw data has not yet been investigated in the context of data collaboration. This study generates several types of virtual data of patients, verifying their similarity to raw data via earth-mover's distance (EMD) [47]. EMD is calculated for the three datasets composed of a random number, the virtual patients, and raw data. The dataset has 100 data samples, which are chosen randomly five times. EMD are calculated between the raw dataset and others, as described below. The mean EMDs for five times are shown in Table 2. Thus, virtual patient data are similar to raw data, as expected, and we evaluate the performance of data collaboration with respect to these data.

Calculated data of virtual patients and their intermediate representations are shown in the additional information.

Evaluation of the Performance of the Classification Task

The classification of DM is carried out next. We designed two types of settings for collaboration: the first is increasing number of collaborative institution where each hospital has the same amount of medical data and the second is increasing number of divisions where the total size of data is fixed.

In the first setting, the raw data are divided to have 40 samples for each institution, and the number of institutions are increased until 25 considered as independent hospitals. In the second setting, the total size of data is fixed and the number of divisions are increased by 20 hospitals from 2 to 202. We used 14,401 for the training data as well as the test data. Therefore, the size of datasets per each hospitals are ranged from 1402 to 144 where the half of samples are used to test samples.

As for the individual analysis, median score among the entire institutions was adopted to avoid the effect of data selection bias. The integrated analysis that shares all raw data is called "ALL-RAW" and the individual analysis that uses only raw data at one institution is called "EACH-RAW". Further, the results of data collaboration are separated by the two types of reconverting functions and anchor data: "SVD-RANDOM" and "SVD-GAN", "Node2Vec-RANDOM"and "Node2Vec-GAN". We adopt the logistic regression with the L2 penalty for classification method and area under the receiver operating characteristic (ROC) curve are calculated as an evaluation metrics.

Result

Figure 2 shows the area under the ROC curve in the case where the number of collaborative hospitals increases and Fig. 3 shows the case where the number of divisions increases. The horizontal axis indicates the number of hospitals with the same amount of data for each hospital (Fig. 2) or the number of divisions with the decreasing amount of data for each hospital (Fig. 3). The vertical axis indicates the area under the ROC curve for the results obtained by data collaboration. The blue and light blue line indicate "ALL-RAW" and "EACH-RAW", respectively, which can be comparison criteria of scores. The green line represents non-linear reconverting function: Node2Vec, and the red line represents linear reconverting function which is originally proposed by Imakura [26]: SVD. Light-colored as well as broken lines indicate the results of using anchor data generated randomly. Finally, the lightly colored area around the line represents the standard error for 10 trials.

To verify Task 1, scores with the same reconverting functions (i.e., "Node2Vec-GAN" and "Node2Vec-RANDOM", or "SVD-GAN" and "SVD-RANDOM") are compared. In

addition, to verify Task 2, scores with the same type of anchor data (i.e., "Node2Vec-GAN" and "SVD-GAN") are compared. The final results for the scores from 25 hospitals in Fig. 2 and 202 hospitals in Fig. 3 are shown in Table 3.

As expected, the scores of data collaboration are lower than that of centralized analysis "ALL-RAW" (see Figs. 2 and 3). First, for the Task1 verification, the different types of anchor data (i.e., "SVD-RANDOM" vs "SVD-GAN" and "Node2Vec-RANDOM" vs "Node2Vec-GAN") are compared. For the both types of experimental conditions for Figs. 2 and 3, the AUC score for data collaboration method using GAN-anchors are greater than that for RANDOM-anchors. Specifically, the AUC scores with GAN-anchors improve more than 10% compared to RANDOM-anchors in any cases in Table 3 . Therefore, the result demonstrates that the similarity of the anchor data to real medical data improves performance of the data collaboration.

Next, we verify Task 2 by comparing the linear and non-linear reconverting functions: SVD and Node2Vec, respectively. As shown in Fig. 2, the data collaboration with non-linear re-conversion outperformed linear re-conversion in terms of the AUC score. In the case of Fig. 2, the score of Node2Vec-GAN with 25 collaborative hospitals results in about 10% higher than "EACH-RAW". In contrast, the score of SVD-GAN exceeds slightly by 3%. In addition, AUC scores for SVD with respect to all divisions in Fig. 3 are lower than that of the individual analysis, "EACH-RAW". These results suggest that reconverting function of SVD is insufficient for applying the data collaboration analysis to medical data. For a conceivable reason, the medical data frequently has extreme values as abnormal ones. Thus, SVD may have removed them as noise [48]. Interestingly, as shown in Fig. 3, the performance of the data collaboration with Node2Vec-GAN turns upward when the number of hospitals grows over 140 and the size of data per hospital is reduced to 100. It is suggested that Node2Vec is robust to a large number of participants but with small data sizes. From these results, it is figured out that anchor data following real data distribution and an appropriate non-linear reconverting function improve the performance of data collaboration analysis.

Table 3 Results of the analysis of 25 hospitals (unit: %)

Figure	2		3	
g_i	SVD	Node2Vec	SVD	Node2Vec
ALL-RAW	90.0	90.0	92.4	92.4
EACH-RAW	74.0	74.0	79.3	79.3
RANDOM	65.4	72.9	63.2	73.5
GAN	76.9	83.8	76.6	83.8

Summary and Discussion

In this study, the data collaboration analysis has been applied to medical data for diagnosis of diabetes mellitus (DM). The following conditions play an important role for the performance of the analysis:

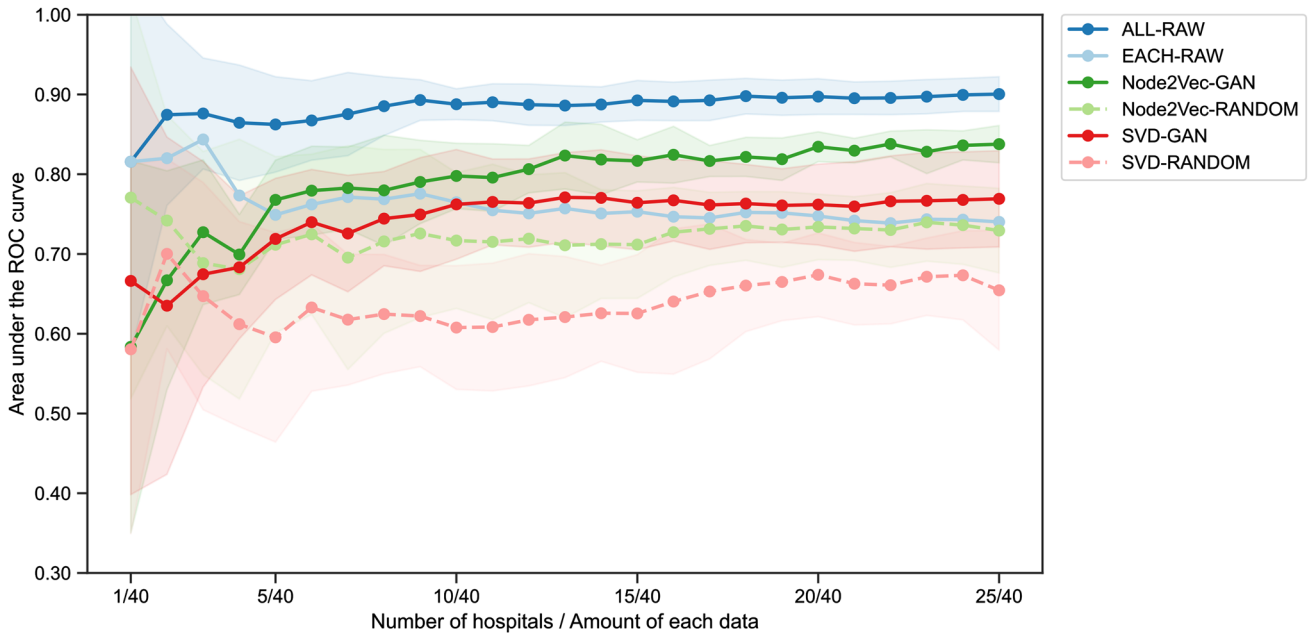

Fig. 2 The performance of the identifying DM subjects depending on SVD and Node2Vec with Ridge regression

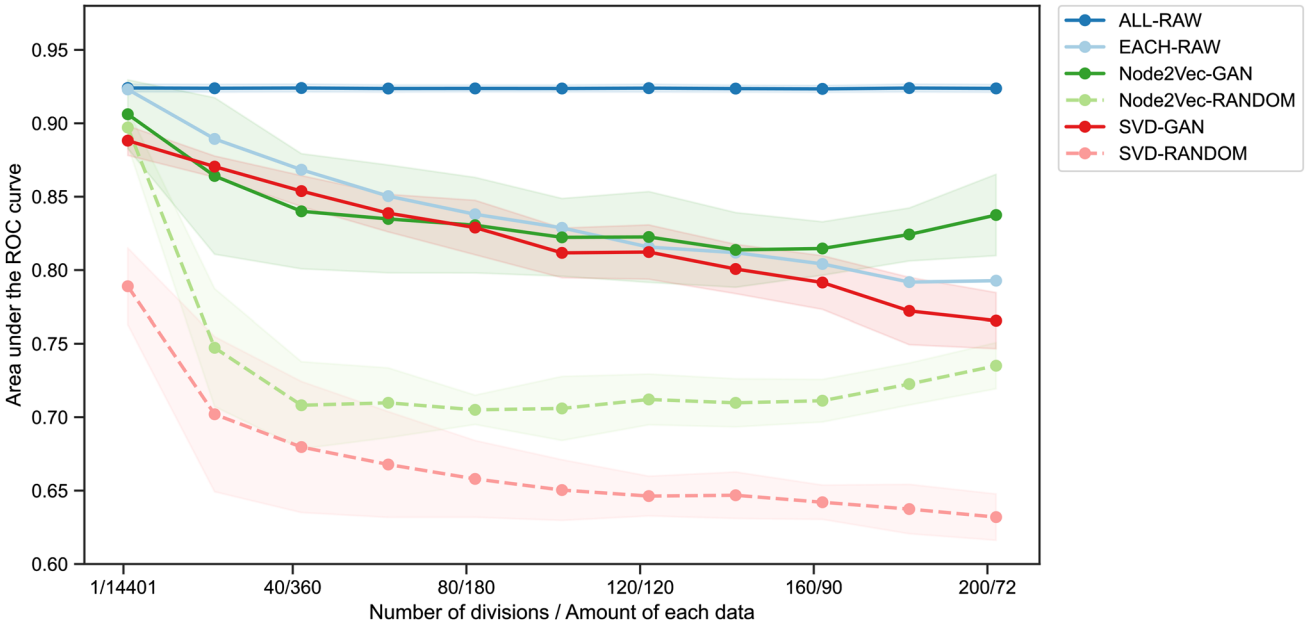

Fig. 3 The performance of the identifying DM subjects depending on SVD and Node2Vec with Ridge regression

(1) The distribution of anchor data are similar to that of raw data.
(2) The reconverting function is non-linear.

Both were satisfied in the present analysis, and the score in identification of DM by the proposed method was a maximum of 20% higher than that by the previous method. This study introduced a way of application of data collaboration analysis suitable for medical data.

Adequate analysis of medical data requires a high degree of accuracy with privacy protection. This study realised high accuracy and preserving privacy of disease classification by dimension reduction methods and appropriate non-linear transformations. The problem of proposed method is that the analysis cannot be interpreted medically. For example, the relation between the classification and non-linearly transformed features (e.g., patients with higher HbA1c being classified as DM) is uncertain. This problem should be solved with further experiments.

The limitations of this study were that the data is obtained from a single hospital and divided into pieces of dataset as virtual hospitals. Furthermore, our approach was evaluated only with numerical data, whereas distributed clinical data possibly included images and text. The applicability of the proposed method to data gathered from many independent hospitals should be verified in the next stage.

As for another future work, we should consider the case where data dimension of virtual patients generated by GAN are different among institutions. This would contribute to real medical data analyses, since the output from each medical equipment would be practically different. Additionally, virtual patient data would contain the values of a normal range for a given parameter (e.g., age, disease threshold). On the other hand, the virtual data of patients can be made to include abnormal values associated with some disease. These virtual patients can be used for educational purpose as well.

Appendix

Examples of Virtual Patients

The part of virtual patients data, generated by GAN and their intermediate representations by the LPP, are shown below (Table 4).

Table 4 The pseudo-patient data and the intermediate representation

(a) Pseudo-patient data

Age	M/F	Glucose (mg/dL)	HbA1c (%)
71	1	103	5.53
82	2	287	7.95
42	2	169	5.66
74	1	197	8.32
73	2	183	7.34
84	1	394	9.65
81	2	101	5.72
83	2	532	14.01
18	2	114	5.19
15	1	99	5.16
77	2	164	7.02
47	2	97	5.30
80	1	157	5.76
43	1	205	5.79
80	1	115	5.63
79	1	308	7.21
46	2	101	6.04
81	1	140	5.51
79	2	101	6.15
46	2	86	5.24
25	1	213	5.24
38	2	109	5.43
28	2	107	5.46
83	1	127	5.96
15	1	159	5.14
87	2	169	6.71
73	1	96	4.88
40	2	78	5.30
78	1	134	6.09
67	2	206	5.96

(b) The intermediate representation of (a)

First	Second	Third
0.30	−0.47	0.03
1.25	−0.06	0.00
1.12	0.28	−0.05
1.06	0.36	−0.05
0.42	−0.39	0.18
1.03	0.35	−0.01
0.81	−0.59	0.56
1.15	0.37	0.16
1.23	0.10	0.03
1.13	0.18	0.01
1.10	0.25	−0.08
0.40	−0.69	−0.04
1.02	0.35	−0.08
1.20	−0.03	−0.08
1.14	0.10	−0.06

Table 4 (continued)

(b) The intermediate representation of (a)

First	Second	Third
0.29	−0.45	−0.15
1.16	0.17	−0.14
1.40	−0.12	−0.13
0.36	−0.35	0.20
1.08	0.28	−0.05
1.09	0.25	−0.01
0.70	−0.60	0.41
0.40	−0.59	0.01
0.39	−0.65	0.01
0.21	−0.35	−0.01
1.06	0.47	0.10
0.32	−0.49	−0.12
0.29	−0.44	−0.04
0.42	−0.63	0.03
0.34	−0.47	−0.01

Acknowledgements The present study was supported in part by the Japan Science and Technology Agency (JST), ACT-I (No. JPMJPR16U6), Mirai Program (No. JPMJMI19B1, JPMJMI19G8), 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. JP17H03280, JP17K12690, JP18H03250, JP18H06363, JP19K12198, JP19K19347).

Author Contributions A.I., H.A., K.S., R.K., and Y.O. conceived the experiments. H.C., Y.T., A.I., H.A., and A.N. conducted the experiments. H.A., Y.O., and T.S. critically reviewed the results. H.T. and R.K. contributed to the data collection and interpretation. H.C., Y.T., A.N., and R.K. wrote the initial draft of the manuscript. All authors critically reviewed the manuscript.

Declarations

Conflict of interest The authors declare that they have no confict of interest.

References

1. Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, Zhang L. Deep learning with differential privacy. In: CCS '16 Proceedings of the 2016 ACM SIGSAC conference on computer and communications security 2016. pp. 308–318.
2. Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat. 1992;46(3):175–85.
3. Bang S, Yoo D, Kim SJ, Jhang S, Cho S, Kim H. Establishment and evaluation of prediction model for multiple disease classification based on gut microbial data. Sci Rep. 2019;9(10189):1–9.
4. Bogdanova A, Nakai A, Okada Y, Imakura A, Sakurai T. Federated learning system without model sharing through integration of dimensional reduced data representations. In: FL-IJCAI'20 Proceedings of the international workshop on federated learning for user privacy and data confidentiality in conjunction with IJCAI 2020; 2020 **(accepted)**.
5. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. N Engl J Med. 2018;378(11):1–3.
6. Chen PHC, Liu Y, Peng L. How to develop machine learning models for healthcare. Nat Mater. 2019;18:410–7.
7. Chillotti I, Gama N, Georgieva M, Izabachene M. Faster fully homomorphic encryption: Bootstrapping in less than 0.1 seconds. In: International conference on the theory and application of cryptology and information security, 2016. pp. 3–33.
8. Cho H, Wu DJ, Berger B. Secure genome-wide association analysis using multiparty computation. Nat Biotechnol. 2018;36(6):547.
9. Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B, Bharath AA. Generative adversarial networks: an overview. CoRR. 2017 arXiv:abs/1710.07035.
10. Cui P, Wang X, Pei J, Zhu W. A survey on network embedding. CoRR. 2017 arXiv:abs/1711.08752.
11. Delude CM. The details of disease. Nature. 2015;527(7576):S14.
12. Dwork C. Differential privacy. In: Bugliesi M., Preneel B., Sassone V., Wegener I, editors. Automata, languages and programming. ICALP 2006. Lecture notes in computer science, 2006, vol. 4052.
13. Federation TID. IDF DIABETES ATLAS. 8th ed. 2017. The International Diabetes Federation. 2017.
14. Fury M, John M, Schexnayder S, Molligan H, Lee O, Krause P, Dasa V. The implications of inaccuracy: comparison of coding in heterotopic ossification and associated trauma. Orthopedics. 2017;40(4):237–41.
15. Gilad-Bachrach R, Dowlin N, Laine K, Lauter K, Naehrig M, Wernsing J. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In: International conference on machine learning, 2016. pp. 201–10.
16. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: NIPS '14 advances in neural information processing systems, vol. 27, 2014. pp. 2672–80.
17. Goodfellow IJ. NIPS 2016 tutorial: Generative adversarial networks. CoRR. 2017 arXiv:abs/1701.00160.
18. Goyal P, Ferrara E. Graph embedding techniques, applications, and performance: a survey. CoRR. 2017. arxiv:abs/1705.02801.
19. Grover A, Leskovec J. Node2vec: scalable feature learning for networks. In: KDD '16 Proceedings of the 22Nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016. pp. 855–64.
20. He X. Locality preserving projections. Ph.D. thesis, University of Chicago, Chicago, IL, USA; 2005.
21. Hebbring SJ. The challenges, advantages and future of phenome-wide association studies. Immunology. 2014;141(2):157–65.
22. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. 2012;20(1):117–21.
23. Hyvärinen A, Karhunen J, Oja E. Independent component analysis, vol. 46. New York: Wiley; 2004.
24. Imakura A, Bogdanova A, Yamazoe T, Omote K, Sakurai T. Accuracy and privacy evaluations of collaborative data analysis. In: PPAI-21 Proceedings of the second AAAI workshop on privacy-preserving artificial intelligence; 2021 **(accepted)**.
25. Imakura A, Inaba H, Okada Y, Sakurai T. Interpretable collaborative data analysis on distributed data. Expert Syst Appl 2021;114891.
26. Imakura A, Sakurai T. Data collaboration analysis for distributed datasets. CoRR. 2019 arXiv:abs/1902.07535.
27. Imakura A, Sakurai T. Data collaboration analysis framework using centralization of individual intermediate representations for distributed data sets. ASCE-ASME J Risk Uncert Eng Syst Part A Civ Eng. 2020;6(2):04020018.
28. Imakura A, Ye X, Sakurai T. Collaborative data analysis: non-model sharing-type machine learning for distributed data. In: PKAW; 2020 **(accepted)**.
29. Jha S, Kruger L, McDaniel P. Privacy preserving clustering. In: European symposium on research in computer security. Springer; 2005. pp. 397–417.
30. Ji Z, Lipton ZC, Elkan C. Differential privacy and machine learning: a survey and review. CoRR. 2014. arXiv:abs/1412.7584.
31. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. NPJ Dig Med. 2018;1(40):1–3.
32. Kerschbaum F. Privacy-preserving computation. In: Privacy technologies and policy. APF 2012. 2014:41–54.
33. Konecný, J., McMahan, H.B., Yu, F.X., Richtárik, P., Suresh, A.T., Bacon, D.: Federated learning: Strategies for improving communication efficiency. CoRR. 2016. arXiv:abs/1610.05492.
34. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. Artif Intell Med. 2001;23(1):89–109.
35. Ligett, K., Neel, S., Roth, A., Waggoner, B., Wu, Z.S.: Accuracy first: Selecting a differential privacy level for accuracy-constrained ERM. CoRR. 2017. arXiv:abs/1705.10829.
36. May M. Twenty-five ways clinical trials have changed in the last 25 years. Nat Med. 2019;25:2–5.
37. McCormick N, Lacaille D, Bhole V, Avina-Zubieta JA. Validity of heart failure diagnoses in administrative databases: a systematic review and meta-analysis. PloS One. 2014;9(8):e104519.
38. McMahan, H.B., Moore, E., Ramage, D., y Arcas, B.A.: Federated learning of deep networks using model averaging. CoRR. 2016. arXiv:abs/1602.05629.
39. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: NIPS '13 Proceedings of the 26th international conference on neural information processing systems. 2013;2:3111–9.
40. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep. 2016;6(26094):1–10.
41. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. J Am Med Inf Assoc. 2013;20(e2):206–11.
42. Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations. In: KDD '14 Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining 2014. pp. 701–10.
43. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347–58.
44. Rana SP, Dey M, Tiberi G, Sani L, Vispa A, Raspa G, Duranti, Ghavami MM Dudley S. Machine learning approaches for automated lesion detection in microwave breast imaging clinical data. Sci Rep. 2019;9(10510:1–12.

45. Romagnoni A, Jégou S, Steen KV, Wainrib G, Hugot JP, (IIBDGC) IIBDGC. Comparative performances of machine learning methods for classifying Crohn disease patients using genome-wide genotyping data. Sci Rep. 2019;9(10351):1–18.
46. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science. 2000;290:2323–6.
47. Rubner Y, Tomasi C, Guibas LJ. The earth mover's distance as a metric for image retrieval. Int J Comput Vis. 2000;40(2):99–121.
48. Sadasivan PK, Dutt DN. SVD based technique for noise reduction in electroencephalographic signals. Signal Process. 1996;55(2):179–89.
49. Shaikhina T, Khovanova NA. Handling limited datasets with neural networks in medical applications: a small-data approach. Artif Intell Med. 2017;75:51–63.
50. Sohail MN, Jiadong R, Uba MM, Irshad M, Iqbal W, Arshad J, John AV. A hybrid forecast cost benefit classification of diabetes mellitus prevalence based on epidemiological study on real-life patient's data. Sci Rep. 2019;9(10103):1–10.
51. Timilsina Mohan TMdM, Yang H. Discovering links between side effects and drugs using a diffusion based method. Sci Rep. 2019;9(10436):1–9.
52. Wei WQ, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. Genome Med. 2015;7(1):41.
53. Woodfield R, Grant I, Group UBSO, Follow-Up UB, Group OW, Sudlow CLM. Accuracy of electronic health record data for identifying stroke cases in large-scale epidemiological studies: a systematic review from the uk biobank stroke outcomes group. PLoS One. 2015;10(10):e0140533.
54. Ye X, Li H, Imakura A, Sakurai T. Distributed collaborative feature selection based on intermediate representation. In: IJCAI-19 Proceedings of the 28th international joint conference on artificial intelligence; 2019. pp. 4142–4149.