Research Day NCI

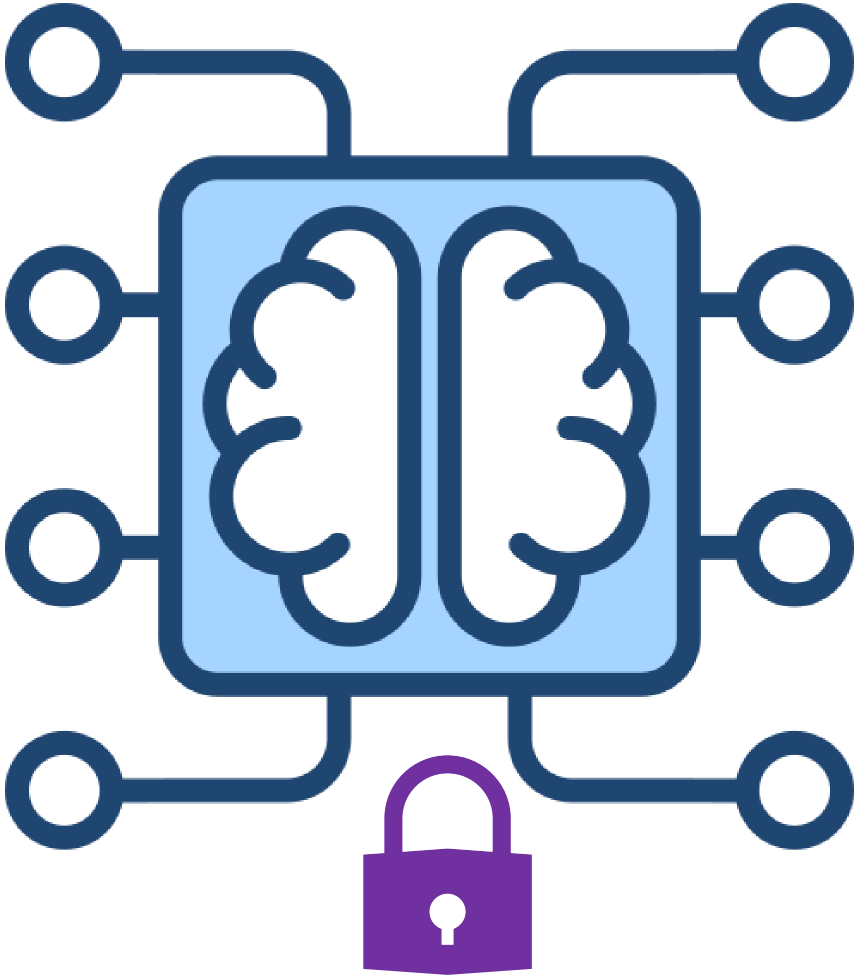
Privacy-Preserving Logistic Regression for Federated Learning Environments with a Policy to Reduce the Training Time

Jorge M. Cortés-Mendoza
Postdoctoral researcher
Cloud Competency Centre, NCI

Dublin, June 2025

Content

- Motivation
- Logistic Regression (LR)
- Related Work
- Training Policy
- Experimental Evaluation
- Conclusions



Motivation

Federated Learning (FL) and ***Homomorphic Encryption (HE)*** are two main directions ***to provide security*** and ***privacy preservation*** by addressing vulnerabilities in ***data processing***

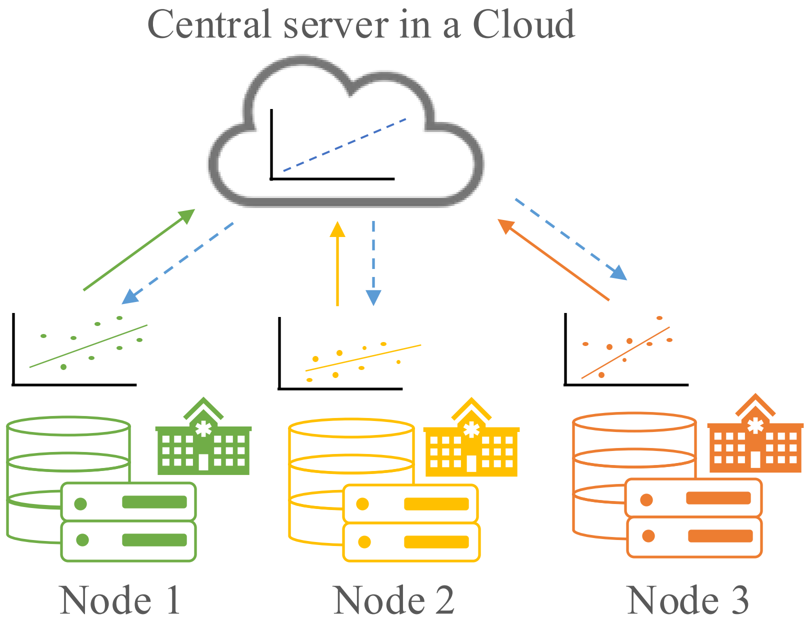


Fig 1. FL system with a central node in a cloud environment and three nodes

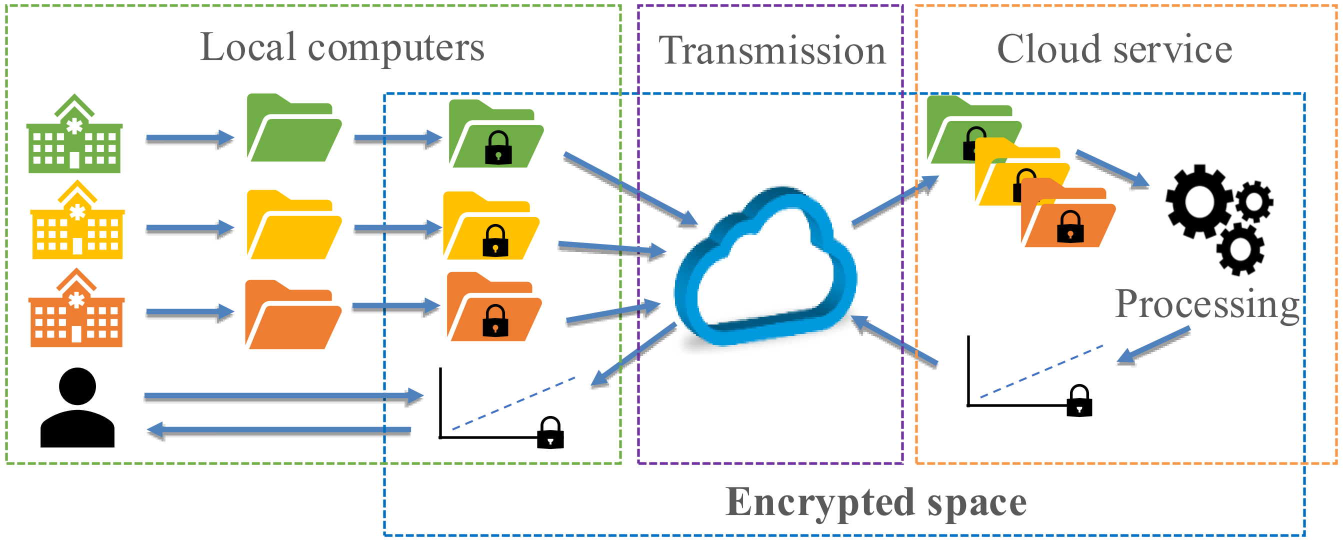


Fig 2. Cloud environment with HE can protect the entire data lifecycle (transmission, storage, and processing)

Logistic Regression (LR)

Logistic Regression (LR) is a statistical method for analyzing information where:

- A dataset $X \in \mathbb{R}^d$ and their labels $Y \in \{0,1\}$ are used to ***model a binary dependent variable***
- The prediction of a binary outcome considers the ***logistic function***

The inference of LR considers the hypothesis $h_\theta(x^{(i)}) = g(\theta^T x^{(i)})$ where

- Logistic function: $g(z) = \frac{1}{1+e^{-z}}$
- Weights: $\theta^T = [\theta_0, \theta_1, \dots, \theta_d]^T$
- Data: $x^{(i)} = [1, x_1{}^{(i)}, x_2{}^{(i)}, \dots, x_d{}^{(i)}]^T$

The training phase of LR focuses on ***finding*** $\boldsymbol{\theta^*}$, the values of θ that ***minimizes the number of errors*** in the prediction

- θ^* is used to estimate the ***binary classification*** of new data

Logistic Regression (LR)

For $x' = [1, x_1, \dots, x_d] \in \mathbb{R}^{d+1}$ is possible to guess its binary value $y' \in \{0,1\}$ by

$$y' = \begin{cases} 1 & if\ h_{\theta^*}(x') \geq \tau \\ 0 & if\ h_{\theta^*}(x') < \tau \end{cases}$$

- τ defines a variable threshold in $0 < \tau < 1$, typically with a value equal to 0.5

Gradient Descent (GD) is the optimization process to find θ^* according to the partial derivate of the cost function $J(\theta)$, represented by $\nabla_\theta J(\theta)$

Algorithm 1. Batch Gradient Descent
Input: $X, Y, \theta, \alpha, and\ nIter$
Output: θ^* (the best θ)
1 For $i \leftarrow 1$ to $nIter$
2 $\theta \leftarrow \theta - \alpha \times \nabla_\theta J(\theta, X, Y)$
3 Return θ

Related Work

There are several limitations with respect to the required ***compute availability*** and ***privacy*** of the models

Several studies have proposed ***innovations*** and ***new approaches*** to overcome the disadvantages of LR with FL and HE

Table 1. Main characteristics of FL and HE approaches for privacy-preserving LR in the literature

HE	FL	Name	Metric	Dataset	Ref
-	*	VFLR	Accuracy (A), Area under the ROC Curve (AUC)	Pima, BCWD, BDM	[1]
-	*	SecureLR	Time	MNIST	[2]
-	*	VANE	Mean Absolute Error (MAE)	BCD, Diabetes dataset (DD), UCID	[3]
-	*	VPPLR	Precision (P), Recall (R)	DD, WIBC, DD, ACAD	[4]
*	-	-	A	MNIST, notMNIST, CIFAR-10	[5]
*	-	-	AUC	iDASH (Genomic), financial	[6]
*	-	Modified GWAS	p-values, F1-score (F1)	iDASH	[7]
*	-	-	A, AUC, K-S values	Korea Credit Bureau (KCB), MNIST	[8]
*	-	-	A, AUC	iDASH, Lbw, Mi, Nhanes3, Pcs, Uis	[9]
*	-	N-LHAE	Overhead	Not described	[10]
*	-	P2OLR, P2VCLR, CECLLR	A, AUC, F1, P, R	Mi, Nhanes3, Uis	[11-13]
*	-	-	A	Digits (scikit-learn library)	[14]

Training Policy

We introduce ***a new training policy*** for FL that progressively reduces the amount of ***training data*** for each iteration

- This reduction allows to perform the ***learning process faster***, effectively reducing the ***training time*** without significant ***accuracy degradation***

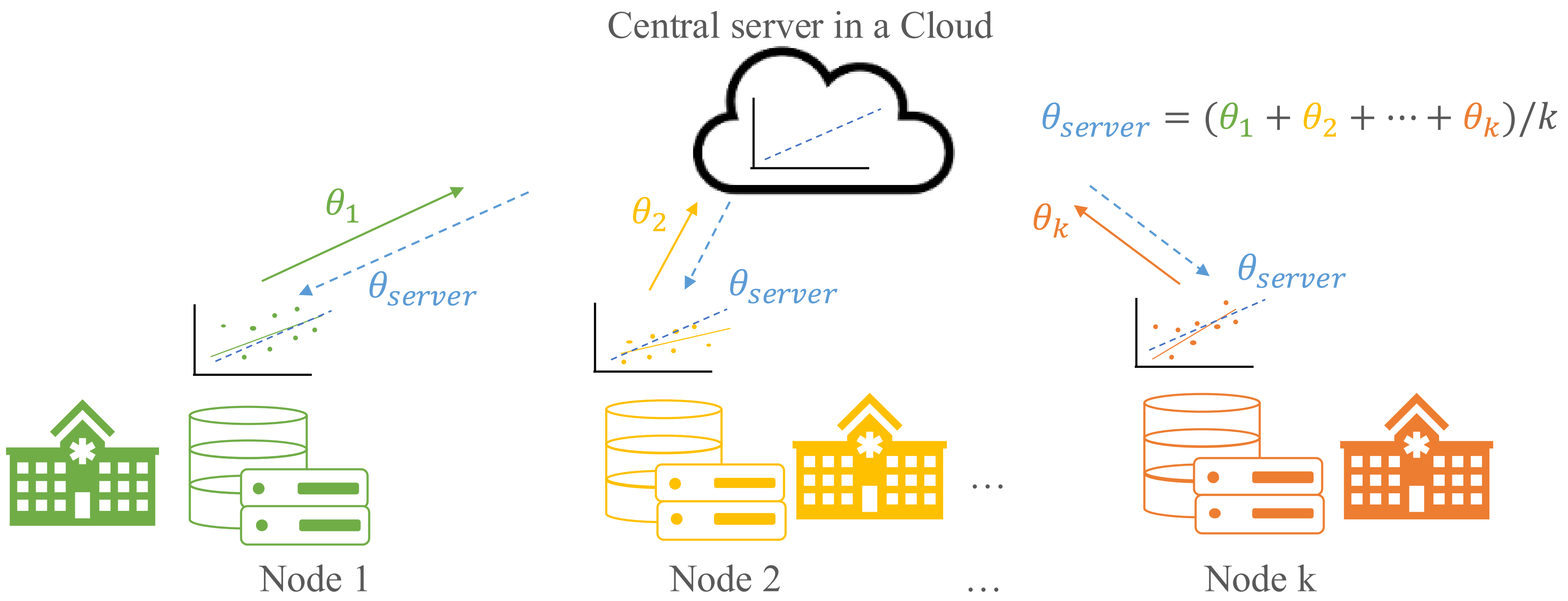
Training Policies:

1. FL Logistic Regression with reduction policy (LR_{FLn})
2. FL Logistic Regression with reduction policy and weights (LR_{FLnw})
3. FL ensemble Logistic Regression (LR_{FLe})
4. FL ensemble Logistic Regression with reduction policy (LR_{FLen})

Training Policy

FL Logistic Regression (LR_{FL})

- Datasets are evenly distributed among the system nodes
- Each local node uses ***all available local data*** to train the model in each iteration

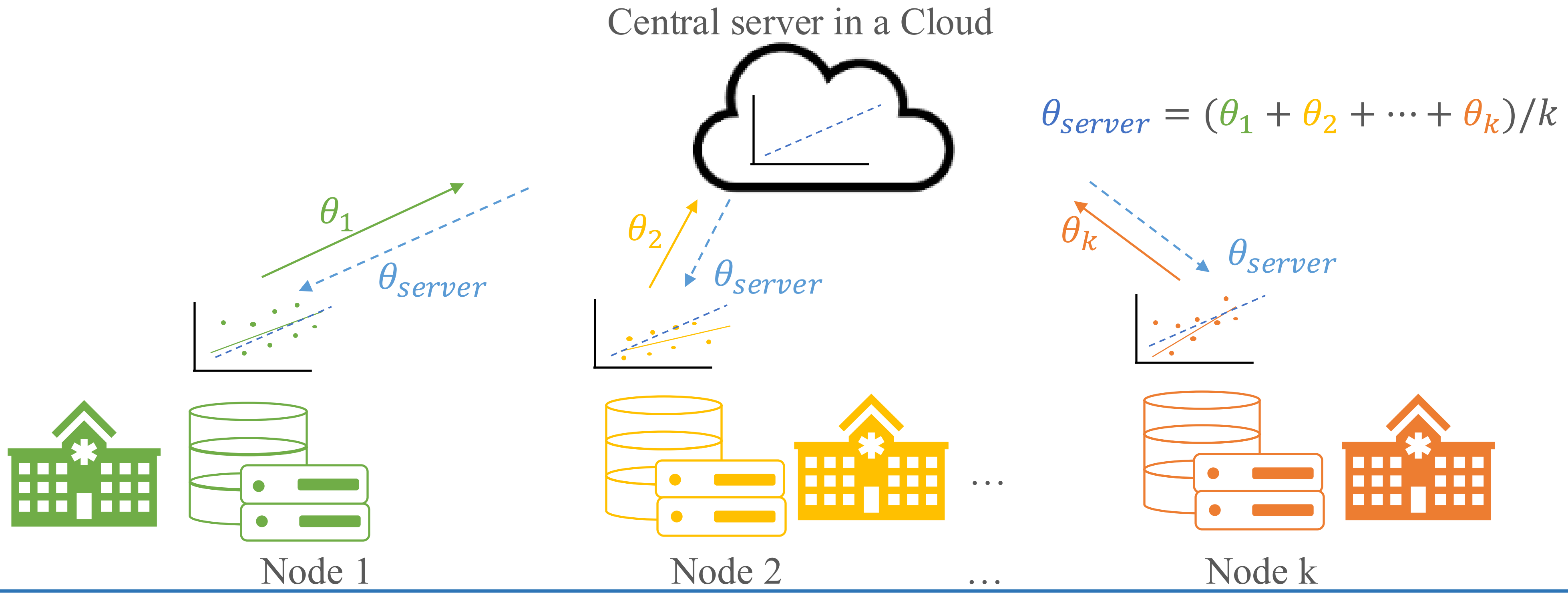


We can obtain θ^*_{server} with $\theta^*_1, \theta^*_2, \dots, \theta^*_k$ after several iterations

Training Policy

LR_{FLn} decreases ***the number of training instances*** in local nodes according to $\mathbf{1/}\boldsymbol{i}$ ***ratio*** where i defines the iteration number

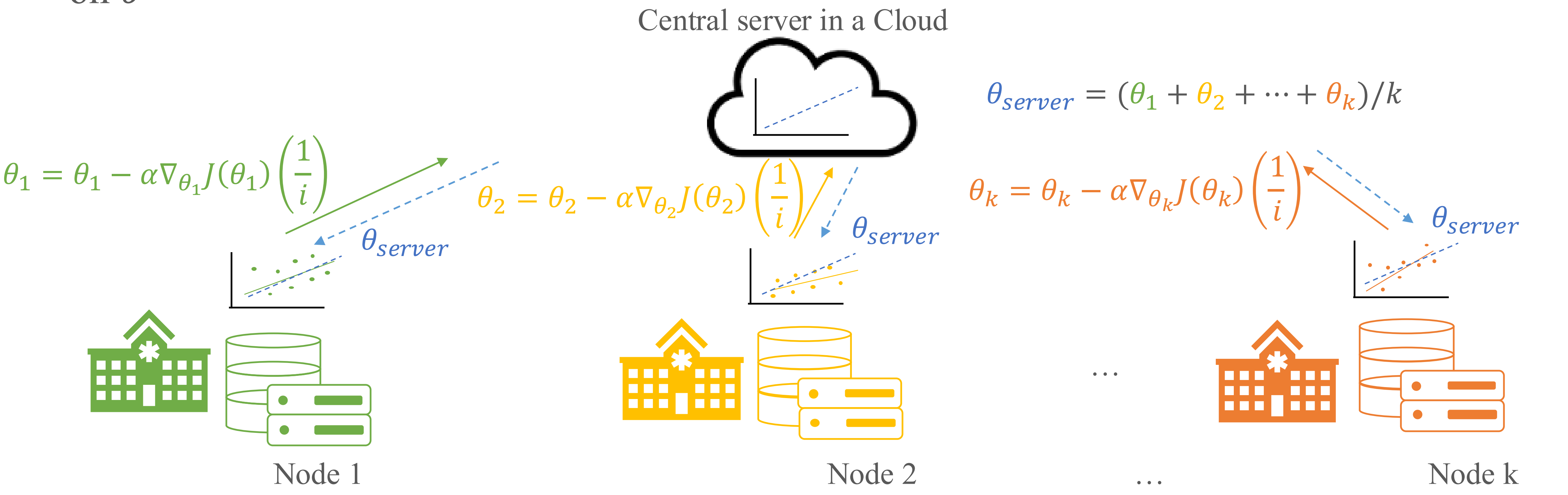
- Reduction: 0% the first iteration, 50% the second time, 66% the third iteration, and so on
- The ***subset of training*** instances is chosen ***randomly*** for each iteration



Training Policy

LR_{FLnw} updates the central node ***proportionally*** with the number of instances from the local models

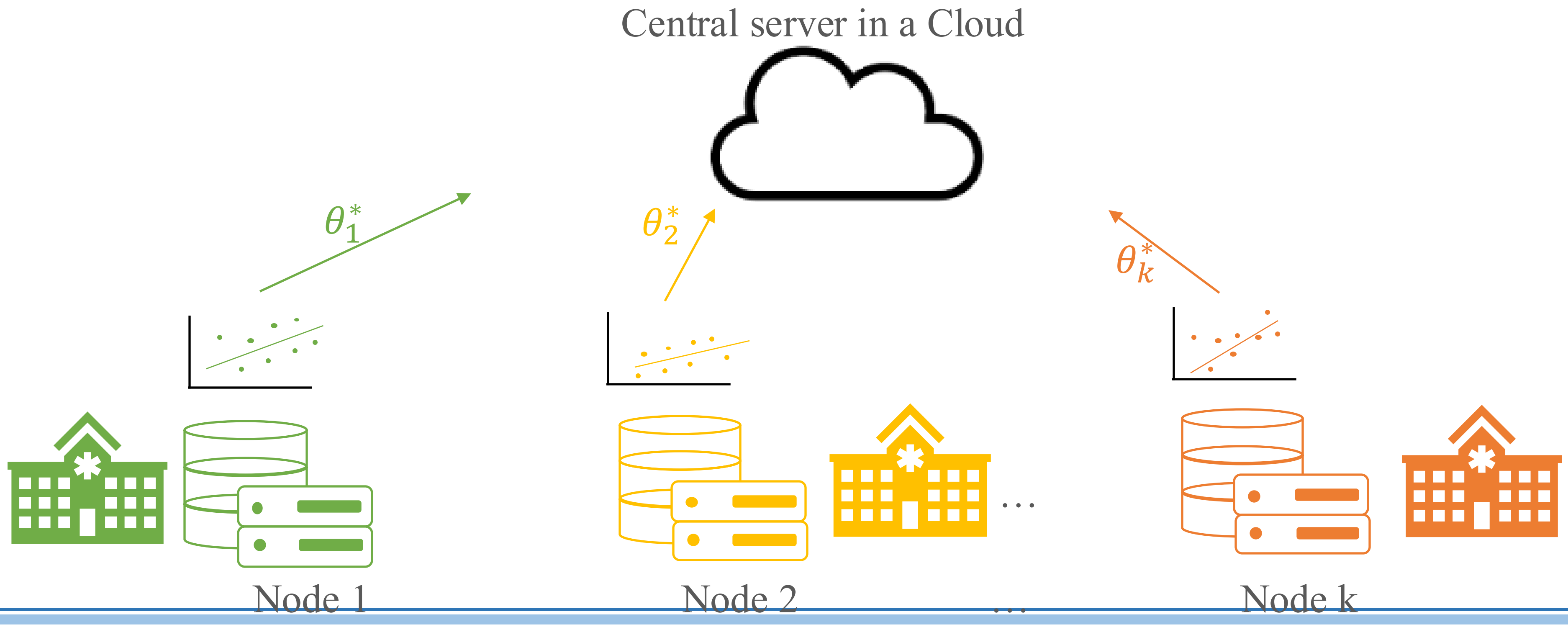
- Then, reducing the number of instances on the training dataset implies a reduced update on θ



Training Policy

In LR_{FLe}, the nodes use ***their data to train local LR models***, and then these models are sent to ***the central server*** to obtain a ***better predictive model***

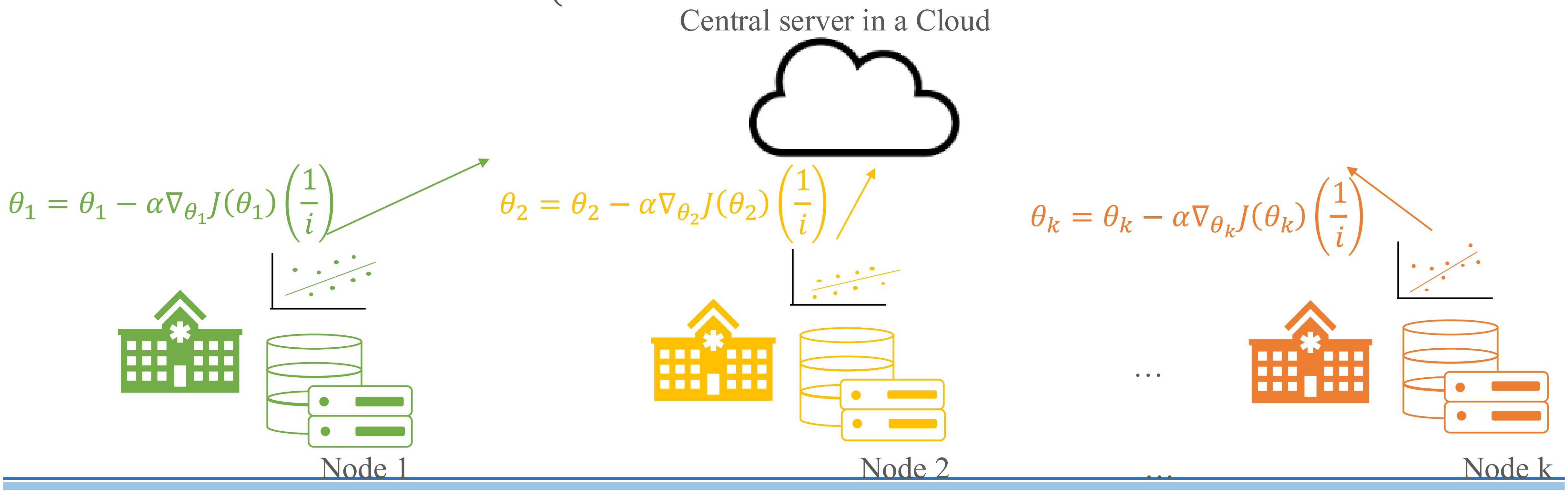
$$y = \begin{cases} 1 \ \ if \ \ (h_{\theta_1^*}(x) + h_{\theta_2^*}(x) + \cdots + h_{\theta_k^*}(x))/k \geq \tau \\ 0 \ \ if \ \ (h_{\theta_1^*}(x) + h_{\theta_2^*}(x) + \cdots + h_{\theta_k^*}(x))/k < \tau \end{cases}$$



Training Policy

LR_{FLen} the nodes use their data to train local LR models, and then these models are sent to the central server to obtain a better predictive model

$$y = \begin{cases} 1 \quad if \ \left(h_{\theta_1^*}(x) + h_{\theta_2^*}(x) + \cdots + h_{\theta_k^*}(x)\right)/k \geq \tau \\ 0 \quad if \ \left(h_{\theta_1^*}(x) + h_{\theta_2^*}(x) + \cdots + h_{\theta_k^*}(x)\right)/k < \tau \end{cases}$$



Experimental Evaluation

We compare the performance of $LR, LR_{FL}, LR_{FLn}, LR_{FLnw}$, LR_{FLe} , and LR_{FLen} considering their ***accuracy*** (A) and ***speedup***

- Horizontal FL system
- Constant number of nodes
- Federated Averaging

Six standard datasets widely used in the literature

- Min-Max normalization
- Simple Split technique

Initial configuration of LR

- 10 learning rates (α)
- 10 values of iterations $(nIter)$
- 30 initial solutions (θ)

Table 2. Characteristics and size of the datasets

Dataset	Features	Instances		
		Total (N)	n-Training	n-Testing
Low Birth Weight Study (Lbw)	9	189	151	38
Myocardial Infarction (Mi)	9	1,253	1,002	251
National Health and Nutrition Examination (Nhanes3)	15	15,649	12,519	3,130
Prostate Cancer Study (Pcs)	9	379	303	76
Indian's diabetes (Pima)	8	768	614	154
Umaru Impact Study (Uis)	8	575	460	115

Experimental Evaluation

Table 3. Average accuracy after 30 executions with the best LR configuration for different FL environment configurations

Dataset	LR	LR_{FL}	LR_{FLn}	LR_{FLnw}	LR_{FLe}	LR_{FLen}	$dif_A(LR_{FLn})$	$dif_A(LR_{FLnw})$	$dif_A(LR_{FLe})$	$dif_A(LR_{FLen})$	$Nodes$
Lbw	0.6965	0.6965	0.6833	0.6965	0.6965	0.6842	1.32	0.0	0.0	1.23	2
			0.6877	0.6965	0.6965	0.6965	0.88	0.0	0.0	0.0	3
			0.6491	0.6965	0.6965	0.6684	4.74	0.0	0.0	2.81	4
			0.6518	0.6965	0.6965	0.6211	4.47	0.0	0.0	7.54	5
			0.6684	0.6965	0.6965	0.6544	2.81	0.0	0.0	4.21	6
Mi	0.9053	0.9057	0.8965	0.7849	0.9046	0.8954	0.92	12.08	0.11	1.04	2
			0.8938	0.7851	0.9042	0.8918	1.20	12.06	0.15	1.39	3
			0.8960	0.7858	0.9042	0.8926	0.97	11.99	0.15	1.31	4
			0.8908	0.7839	0.9028	0.8956	1.49	12.18	0.29	1.01	5
			0.8963	0.7851	0.9023	0.8919	0.94	12.06	0.35	1.38	6
Nhanes3	0.7916	0.7915	0.7915	0.7914	0.7915	0.7915	0.0	0.01	0.0	0.0	2
			0.7914	0.7915	0.7915	0.7914	0.01	0.0	0.0	0.01	3
			0.7915	0.7915	0.7915	0.7915	0.0	0.0	0.0	0.0	4
			0.7916	0.7915	0.7915	0.7916	0.0	0.01	0.0	-0.01	5
			0.7915	0.7915	0.7915	0.7915	0.0	0.01	0.0	0.0	6
Pcs	0.6667	0.6658	0.6246	0.6237	0.6654	0.6211	4.12	4.21	0.04	4.47	2
			0.5908	0.6263	0.6654	0.5829	7.50	3.95	0.04	8.29	3
		0.6654	0.5728	0.6232	0.6636	0.5776	9.25	4.21	0.18	8.77	4
			0.5917	0.6189	0.6658	0.5618	7.37	4.65	-0.04	10.35	5
		0.6658	0.6114	0.6215	0.6654	0.6228	5.44	4.43	0.04	4.30	6
Pima	0.6543	0.6537	0.6476	0.3463	0.6537	0.6474	0.61	30.74	0.0	0.63	2
			0.6543	0.3506	0.6537	0.6543	-0.06	30.30	0.0	-0.06	3
			0.6548	0.3500	0.6537	0.6545	-0.11	30.37	0.0	-0.09	4
			0.6535	0.3461	0.6537	0.6535	0.02	30.76	0.0	0.02	5
			0.6545	0.3496	0.6537	0.6543	-0.09	30.41	0.0	-0.06	6
Uis	0.7365	0.7365	0.7365	0.7365	0.7365	0.7365	0.0	0.0	0.0	0.0	2
			0.7365	0.7365	0.7365	0.7365	0.0	0.0	0.0	0.0	3
			0.7365	0.7365	0.7365	0.7365	0.0	0.0	0.0	0.0	4
			0.7365	0.7365	0.7365	0.7365	0.0	0.0	0.0	0.0	5
			0.7365	0.7365	0.7365	0.7365	0.0	0.0	0.0	0.0	6

Experimental Evaluation

The speedup measures consider the ***worst time*** of all nodes in the FL environment per iteration

Table 4. Speedup of FL environments with respect to LR for all the datasets

Dataset	LR_{FL}	LR_{FLn}	LR_{FLnw}	LR_{FLe}	LR_{FLen}	Nodes
Lbw	0.91	0.91	0.92	1.20	0.99	2
	0.94	0.93	0.94	1.24	1.01	3
	0.96	0.93	0.94	1.28	1.00	4
	0.96	0.94	0.96	1.28	1.01	5
	0.98	0.95	0.96	$\boxed{1.31}$	1.01	6
Mi	2.04	2.46	2.45	2.44	3.24	2
	2.40	2.56	2.58	2.85	3.36	3
	2.54	2.61	2.62	3.15	3.50	4
	2.83	2.63	2.66	3.50	3.55	5
	2.95	2.66	2.66	$\boxed{3.60}$	3.54	6
Nhanes3	1.28	2.51	2.54	1.74	4.87	2
	1.25	2.85	2.91	2.31	5.97	3
	1.48	3.09	3.14	2.83	6.69	4
	1.69	3.29	3.30	3.12	7.15	5
	1.87	3.42	3.45	3.40	$\boxed{7.56}$	6

Dataset	LR_{FL}	LR_{FLn}	LR_{FLnw}	LR_{FLe}	LR_{FLen}	Nodes
Pcs	1.00	1.05	1.08	1.13	1.14	2
	1.09	1.09	1.10	1.38	1.16	3
	1.13	1.08	1.10	1.43	1.16	4
	1.14	1.09	1.11	$\boxed{1.47}$	1.20	5
	1.15	1.10	1.12	1.46	1.21	6
Pima	1.04	1.07	1.08	1.32	1.44	2
	1.22	1.13	1.11	1.49	1.44	3
	1.28	1.14	1.14	1.59	1.50	4
	1.38	1.14	1.13	1.77	1.52	5
	1.41	1.16	1.14	$\boxed{1.83}$	1.53	6
Uis	1.08	1.00	1.00	1.33	1.32	2
	1.17	1.01	1.02	1.45	1.36	3
	1.29	1.03	1.04	1.63	1.37	4
	1.32	1.03	1.03	1.67	1.38	5
	1.34	1.04	1.04	$\boxed{1.72}$	1.38	6

Experimental Evaluation

The normalized values of accuracy and time

- LR_{FL} provides the maximum execution time and no reduction in accuracy
- LR_{FLe} reduces the accuracy very little with respect to the worst strategy and provides an acceleration of 50% with respect to the maximum time reduction
- LR_{FLen} has the lowest execution time and an accuracy reduction of about 25%

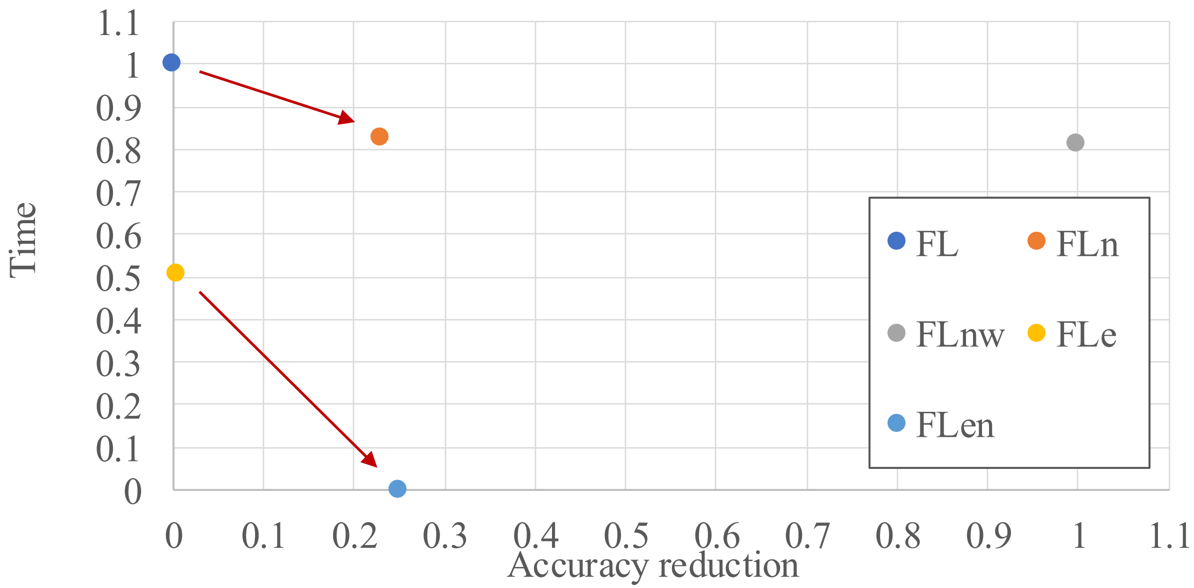


Fig 3. Normalized accuracy and time for all the FL models, datasets, and node configuration

Experimental Evaluation

We also present the time to ***encrypt***, ***decrypt***, and calculate the ***aggregation*** of the ciphertexts with the values θ_i and θ_{server}

- CKKS scheme with a security level of 128 bits, a polynomial modulo degree at most $2^{13}-1$, and a moduli chain equal to {31, 26, 26, 26, 26, 26, 26, 31}[14]

Table 5. Average time of HE operations (sec)

	Encrypt	Average	Decrypt
Lbw	0.02409	0.00845	0.00916
Mi	0.02415	0.00838	0.00930
Nhanes3	0.03171	0.01085	0.01230
Pcs	0.02415	0.00796	0.00959
Pima	0.02469	0.00839	0.00937
Uis	0.02612	0.00906	0.00986

Conclusions

- We analyze the latest advances in privacy-preserving LR solutions for processing confidential data using FL and HE
- We present the characteristics of the most recent approaches in the field: algorithms, evaluation metrics, used datasets, implementation characteristics, etc.
- We proposed one policy to reduce the training time of the federated model and conduct a comprehensive simulation analysis on the six datasets from medicine (diabetes, cancer, drugs, etc.) and genomics
- The results show that the proposed policies can reduce the training time with a slight reduction in the final accuracy of the model

Questions ?


thank you
danke
рахмат
謝謝
ngiyabonga
شكراً جزيلاً
teşekkür ederim
Баярлалаа
спасибо
faafetai lava
vinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
gracias
tapadh leat
mahalo
хвала
asante
manana
obrigada
tenki
mersi
kia ora
barka
welalin
tack
spas
kiitos
dankie
dhanyavad
nanni
nandri
bedankt
enkosi
bayarlalaa
gracie
hvala
mauruuru
köszönöm
dziękuje
dankon
akun
djiere dieuf
tau
mochchakkeram
murakoze
дякую
mamnun
chokrane
chnorakaloutioun
gratias ago
gràcies
sulpáy
go raibh maith agat
sobodi
děkuji
obrigado
mèsi
sagolun
sukriya
kop khun krap
taiku
arigatô
takk
dakujem
trugarez
didi madloba
najis tuke
ありがとう
tanemirt
rahmet
grazie
diolch
dhanyavadagalu
shukriya
merce
мерси
kam sah hamnida
rahmat
terima kasih
xièxie
ευχαριστώ
merci
감사합니다
তোমাকে ধন্যবাদ

References

1. D. He, R. Du, S. Zhu, M. Zhang, K. Liang, and S. Chan, "Secure logistic regression for vertical federated learning," IEEE Internet Computing, 26(2), 61-68, 2021
2. F. Wang, H. Zhu, R. Lu, Y. Zheng, and H. Li, "A privacy-preserving and non-interactive federated learning scheme for regression training with gradient descent," Information Sciences, 552, 183-200, 2021
3. Y. Zhang, and M. Tang, "VPPLR: Privacy-preserving logistic regression on vertically partitioned data using vectorization sharing," Journal of Information Security and Applications, 82, 2024
4. C. Bonte, and F. Vercauteren, "Privacy-preserving logistic regression training," BMC medical genomics, vol. 11, pp. 13–21, 2018
5. K. Edemacu and J. W. Kim, "Multi-party privacy-preserving logistic regression with poor quality data filtering for IoT contributors," Electronics, vol. 10, no. 17, p. 2049, 2021
6. D. Kim, Y. Son, D. Kim, A. Kim, S. Hong, and J. H. Cheon, "Privacy-preserving approximate GWAS computation based on homomorphic encryption," BMC Medical Genomics, vol. 13, no. 7, pp. 1–12, 2020
7. K. Han, S. Hong, J. H. Cheon, and D. Park, "Logistic regression on homomorphic encrypted data at scale," in AAAI-19, vol. 33, (Honolulu), pp. 9466–9471, Feb. 2019
8. J. Chiang, "Privacy-preserving logistic regression training with a faster gradient variant," arXiv preprint arXiv:2201.10838, 2022
9. Y. Zhou, L. Song, Y. Liu, P. Vijayakumar, B. B. Gupta, W. Alhalabi, and H. Alsharif, "A privacy-preserving logistic regression-based diagnosis scheme for digital healthcare," Future Generation Computer Systems, vol. 144, pp. 63–73, 2023
10. X. Yu, W. Zhao, Y. Huang, J. Ren, and D., Tang, "Privacy-preserving outsourced logistic regression on encrypted data from homomorphic encryption," Security and Communication Networks, no. Article ID 1321198, 2022
11. X. Yu, W. Zhao, D. Tang, and K. Liang, "Privacy-preserving vertical collaborative logistic regression without trusted third-party coordinator," Security and Communication Networks, no. Article ID 5094830, 2022

References

12. X. Yu, D. Tang, and W. Zhao, “Privacy-preserving cloud-edge collaborative learning without trusted third-party coordinator,” Journal of Cloud Computing, vol. 12, no. 1, pp. 1–11, 2023
13. C. Liu, Z. L. Jiang, X. Zhao, et al., “Efficient and privacy-preserving logistic regression scheme based on leveled fully homomorphic encryption,” in INFOCOM 2022, (New York), pp. 1–6, IEEE, May 2022
14. M. Albrecht, M. Chase, H. Chen, et al., “Homomorphic encryption standard,” in Protecting Privacy through Homomorphic Encryption, (K. Lauter, W. Dai, and K. Laine, eds.), ch. Part II, pp. 31–62, Cham: Springer, 2021. ISBN 978-3-030-77286-4, DOI: 10.1007/978-3-030-77287-1_2