

Multi-Site Clinical Federated Learning using Recursive and Attentive Models and NVFlare

Won Joon Yun¹, Samuel Kim², and Joongheon Kim¹

¹ School of Electrical Engineering, Korea University, Seoul, Republic of Korea

² CIPHEROME Inc., San Jose, California, USA

Abstract—The prodigious growth of digital health data has precipitated a mounting interest in harnessing machine learning methodologies, such as natural language processing (NLP), to scrutinize medical records, clinical notes, and other text-based health information. Although NLP techniques have exhibited substantial potential in augmenting patient care and informing clinical decision-making, data privacy and adherence to regulations persist as critical concerns. Federated learning (FL) emerges as a viable solution, empowering multiple organizations to train machine learning models collaboratively without disseminating raw data. This paper proffers a pragmatic approach to medical NLP by amalgamating FL, NLP models, and the NVFlare framework, developed by NVIDIA. We introduce two exemplary NLP models, the Long-Short Term Memory (LSTM)-based model and Bidirectional Encoder Representations from Transformers (BERT), which have demonstrated exceptional performance in comprehending context and semantics within medical data. This paper encompasses the development of an integrated framework that addresses data privacy and regulatory compliance challenges while maintaining elevated accuracy and performance, incorporating BERT pretraining, and comprehensively substantiating the efficacy of the proposed approach.

Index Terms—Federated Learning, Clinical Data, Language Model

I. INTRODUCTION

The burgeoning expansion of digital health data has incited a mounting interest in employing machine learning methodologies, such as natural language processing (NLP), to scrutinize medical records, clinical notes, and other text-based health information. NLP techniques have exhibited considerable potential in augmenting patient care and informing clinical decision-making [1]. However, the sensitive nature of health data and the imperative for adherence to regulations, such as the Health Insurance Portability and Accountability Act (HIPAA), pose considerable challenges in terms of data privacy and security. Federated learning (FL) constitutes a distributed machine learning paradigm that empowers multiple organizations to collaboratively train a model without sharing raw data, thereby ensuring data privacy and legal compliance [2], [3], [4]. FL facilitates the cooperative training of a shared machine learning model across multiple clinics while safeguarding patient data privacy and complying with regulations. Furthermore, the collaborative training using data from diverse clinics promotes the development of more robust and accurate models, which may potentially generalize better to unseen data, culminating in enhanced diagnostics.

Additionally, FL obviates data silos that frequently arise in multi-site clinics, as it allows institutions to learn from one another’s data without infringing upon privacy regulations. This collaborative approach can foster increased knowledge sharing and improved patient outcomes. As a result, FL can address real-world data discrepancies, such as varying data quality, data distribution, and data labeling practices across different clinics, rendering the shared model more applicable to heterogeneous clinical settings and populations.

Motivated by the indispensability of FL, this paper proffers a pragmatic approach to medical NLP by amalgamating FL, NLP models, and NVFlare. We introduce two exemplary NLP models: the Long-Short Term Memory (LSTM)-based model and Bidirectional Encoder Representations from Transformers (BERT) [5]. NVFlare, devised by NVIDIA, is a versatile and scalable framework for FL that delivers system reliability, privacy preservation, and optimal resource allocation [6]. By integrating these components, the proposed framework tackles the challenges of data privacy and regulatory compliance whilst maintaining elevated accuracy and performance.

The salient contributions are tri-folded: First, this paper integrates the FL framework that addresses the challenges of data privacy and regulatory compliance while maintaining high accuracy and performance in medical NLP; Second, this paper incorporates not only a general training method but also BERT pretraining, which broadens the applicability of the proposed framework; Lastly, this paper comprehensively demonstrates the efficacy of the proposed approach, which transitions from traditional LSTM-based models to BERT, thereby establishing the practicality of the reference framework.

This paper is structured as follows: Sec. II delineates the related work encompassing FL, NLP models, and NVFlare; Sec. III expounds upon the methodology for integrating these technologies; Sec. IV deliberates the results and performance of the proposed approach; and Sec. V concludes the paper with a discourse on future research directions.

II. RELATED WORK

A. Federated Learning with Medical NLP Models

FL is a distributed machine learning paradigm that allows multiple organizations to collaboratively train a model without sharing raw data [2]. In this approach, each organization trains a local model on its data and submits the model updates to a central server. The server aggregates these updates to

produce a global model, which is then disseminated back to the participating organizations for further local updates. This iterative process continues until the desired level of accuracy and convergence is achieved. FL’s key advantage lies in preserving data privacy by retaining sensitive information within the boundaries of each participating organization. This approach has garnered significant attention in healthcare, where data privacy concerns and regulatory compliance are of paramount importance [7].

In the context of medical NLP, FL enables the development of robust models that leverage the expertise of multiple healthcare institutions while adhering to privacy regulations. By combining FL with advanced NLP models, such as LSTM and BERT, it is possible to create powerful solutions for tasks like medical entity extraction, relation extraction, clinical document classification, generating medical reports, and predicting patient outcomes from clinical notes, all while preserving data privacy.

LSTM-based models are often preferred in medical NLP tasks due to their capacity to handle sequential data and capture long-range dependencies [8]. On the other hand, BERT is an advanced language model that has demonstrated exceptional performance in various NLP tasks, surpassing traditional LSTM-based models [5]. BERT is pre-trained on large text corpora and fine-tuned on specific tasks, enabling it to discern intricate relationships and contextual information within the text. BERT’s bidirectional nature permits it to apprehend the context of words from both left and right directions, which is particularly advantageous for analyzing complex and domain-specific language found in medical data.

B. Related Federated Learning Frameworks

This section presents an overview of the prominent FL frameworks extensively employed in both industrial and academic settings. These frameworks include:

- *TFF* [9]: TFF, an open-source framework developed by Google, facilitates machine learning and other computations on decentralized data. It enables developers to implement federated learning algorithms using the high-level APIs of TensorFlow.
- *PySyft* [10]: The OpenMined community developed PySyft, a Python library providing federated learning, secure multi-party computation, and differential privacy capabilities. PySyft extends widely-used deep learning libraries such as PyTorch and TensorFlow to support secure and privacy-preserving machine learning.
- *FATE* [11]: WeBank’s open-source FL framework, FATE, is designed to facilitate secure computations on distributed data. It encompasses various federated learning algorithms, including vertical and horizontal federated learning.
- *LEAF* [12]: LEAF, an open-source benchmarking framework for federated learning, was developed by researchers at Carnegie Mellon University. It offers a collection of datasets, pre-processing tools, and evaluation metrics to enable fair comparisons among different federated learning algorithms.

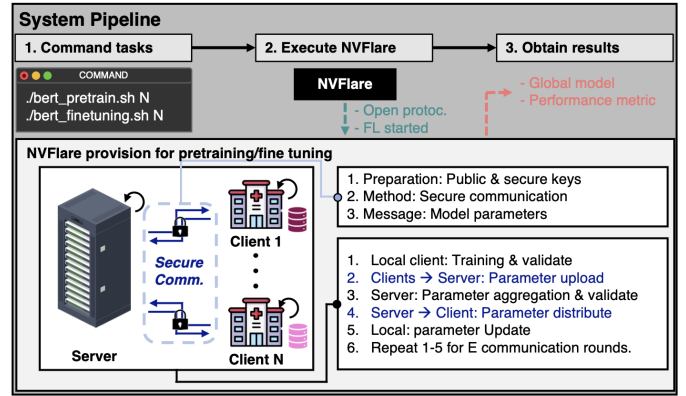


Fig. 1. System pipeline.

In comparison to these frameworks, *NVFlare* delivers superior system reliability, privacy preservation, and optimal resource allocation for distributed machine learning [6]. The framework is designed to be flexible and scalable, promoting collaboration among multiple healthcare institutions in training deep learning models without sharing raw data. While BERT has not been previously implemented on *NVFlare*, this paper demonstrates the integration of medical NLP models and *NVFlare*, thereby rendering it an ideal platform for the development of privacy-preserving medical NLP solutions.

III. METHODOLOGY

A. NVFlare integration

Fig. 1 represents the referencing system pipeline. The system pipeline comprises three main stages: i) tasks allocation, ii) *NVFlare* execution, and iii) obtaining results. The task allocation stage involves the allocation of pretraining and finetuning processes. Upon executing *NVFlare* execution, the system generates *NVFlare* provision, which establishes server-client protocols, followed by federated learning. The *NVFlare* provision involves secure communication between the server and clients, involving a process that includes the preparation of public and secure keys, secure communication methods, and transmission of model parameters. The procedure of *NVFlare* operations encompasses local client training and validation, client-to-server parameter upload, server parameter aggregation and validation, server-to-client parameter distribution, and local parameter update, which repeats for E communication rounds. Finally, the system obtains optimal global models and performance metrics.

B. Model training

The weights of NLP models are updated to detect patient diagnosis. To detect the patient diagnosis, this paper adopts binary classification as a task. The task is to detect patients with adverse drug reactions (ADR). We have collected electronic health records of 8,638 patients with clopidogrel prescriptions (1,824 patients were identified as treatment failure cases) [13].

We also adopt a widely-used pre-training objective for transformers, namely the masked-language-model (MLM) technique, as introduced in the BERT model [5]. The MLM

TABLE I
PARAMETERS USED IN THIS PAPER.

Description	Values
Number of clients	8
Hardware spec.	<ul style="list-style-type: none"> • Machine 1 (Local server) * OS: Ubuntu 20.04 LTS, * CPU: Intel Xeon E5-2638 (2ea), * GPU: NVIDIA RTX 2080 Ti (4ea), * RAM: 128 GB • Machine 2 (AWS server, p3.8xlarge)
Software info.	PyTorch v1.1.3, CUDA v11.7, NVFlare v2.2, MLM-Pytorch , X-Transformers
Data info.	<ul style="list-style-type: none"> • # of train data (pretraining): 453,377, • # of valid. data (pretraining): 8,683, • # of train data (validation): 6,927, • # of valid. data (validation): 1,732,
Learning info.	• Optimizer/learning rate: Adam, 10^{-2}

TABLE II
MEDICAL NLP MODELS USED IN THIS PAPER.

Specification/Model	BERT	BERT-mini	LSTM
Hidden dimension	128	50	128
# of attention heads	6	2	-
# of hidden layers	12	6	3

objective aims to predict the original tokens of a sentence from their masked versions to enable the model to learn to comprehend the context and semantics of the language. In this study, a masking probability of $p = 0.15$ is utilized for MLM, which involves masking 15% of the tokens in each sequence. To regulate the BERT model, 10% of the tokens were not masked but were included in the loss calculation. During the MLM training process, the model generates probabilities for each token in the vocabulary at the masked positions, which are compared with the ground truth token for the masked positions using the cross-entropy loss. Through minimizing the loss function during training, the model learns to produce contextually accurate token predictions, thereby improving its understanding of the semantics and structure of the language.

IV. DEMONSTRATION

A. Experiment Setup

To investigate the performance of the proposed framework, a feasibility testing on BERT pretraining and a comparison of three models (*e.g.*, LSTM, BERT, BERT-mini) in centralized, FL, and standalone training modes are studied. These studies are conducted with eight clients and two Linux machines. The rest of the simulation parameters are listed in Tables I and II.

B. Results

1) *Feasibility study on BERT pretraining*: In this feasibility study on BERT pretraining, the paper aims to investigate four distinct training schemes: 1) *BERT using centralized data*, 2) *BERT utilizing a small dataset*, 3) *BERT trained on imbalanced data*, and 4) *BERT using balanced data*. It is noteworthy that BERT employing centralized data is considered the upper bound of performance metric, while BERT utilizing a small dataset is regarded as the lower bound of performance metric. On the other hand, BERT trained on imbalanced data and balanced data represent the main focus of the FL schemes. The data imbalance is implemented by splitting the data into ratios of $\{0.29, 0.22, 0.17, 0.14, 0.09, 0.04, 0.03, 0.02\}$, with

TABLE III
TOP-1 ACCURACY [%] OF VARIOUS NLP MODELS.

Schemes/Model	BERT	BERT-mini	LSTM
Centralized	80.1	72.7	87.9
Standalone	72.2	68.5	67.3
FL	80.1	72.3	87.5

the number of data points identical for each client in the BERT using balanced data scheme. The MLM loss is found to be comparatively lower in BERT using centralized data, BERT using balanced data, and BERT trained on imbalanced data, as well as in BERT employing a small dataset. The MLM loss begins at 10.7 and ultimately reaches 3.5 for BERT utilizing centralized data, BERT trained on imbalanced data, and BERT using balanced data. In contrast, BERT utilizing a small dataset attains only an MLM loss of 4.4, indicating that BERT with a decentralized approach is insufficient in generating medical knowledge effectively.

2) *Feasibility study on recursive model in FL*: This study investigates the feasibility of FL through the utilization of three models: LSTM, BERT, and BERT-mini. Table III presents the performance of these models in centralized, FL, and standalone approaches, respectively. The LSTM-based model exhibits the highest performance, with values of 87.9%, 87.5%, 67.3% for centralized, FL, and standalone approaches, respectively. Similarly, BERT and BERT-mini demonstrate similar performance tendencies across the schemes. However, the LSTM-based model outperforms both BERT and BERT-mini.

3) *Limitation on BERT*: Although BERT is generally recognized for its superior performance in comparison to LSTM across a wide range of natural language processing tasks, there are certain circumstances in which LSTM may outperform BERT. These situations can be attributed to the following factors: Firstly, task characteristics play a significant role, as LSTM may possess a more appropriate structure for specific tasks relative to BERT. For example, LSTM may excel in sequence modeling or time series data processing tasks. Secondly, dataset size is a crucial determinant, since LSTM can be effectively trained with relatively smaller amounts of data. Consequently, in scenarios with limited datasets, LSTM

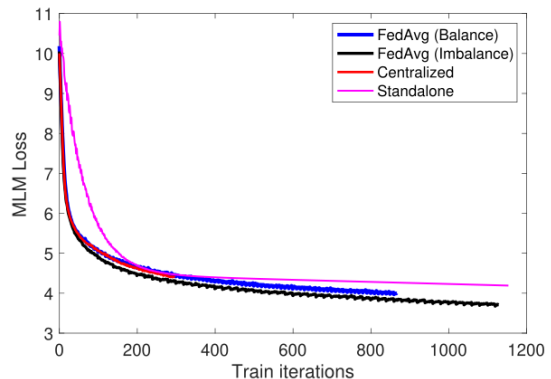


Fig. 2. MLM loss.

Initialize server and client

```
SimulatorRunner - INFO - Create the simulate clients.  
ClientManager - INFO - Client: New client site-1@127.0.0.1 joined. Sent token: 2c15ddc6-d8d3-4a98-8243-d850f27ac052. Total clients: 1  
FederatedClient - INFO - Successfully registered client:site-1 for project simulator_server. Token:2c15ddc6-d8d3-4a98-8243-d850f27ac052  
:  
FederatedClient - INFO - Successfully registered client:site-8 for project simulator_server. Token:64245db0-6e5d-4fff-8ee4-e35c6966bc24  
SimulatorRunner - INFO - Set the client status ready.  
SimulatorRunner - INFO - Deploy and start the Server App.
```

Token & SSH Protocols

Local train

```
2023-04-07 06:33:32,614 - CiBertLearner - INFO: Local epoch site-2: 1/10 (lr=0.01), train_loss=0.761, valid_acc=0.481  
2023-04-07 06:33:33,911 - CiBertLearner - INFO: Local epoch site-7: 1/10 (lr=0.01), train_loss=0.875, valid_acc=0.492  
2023-04-07 06:33:34,176 - CiBertLearner - INFO: Local epoch site-5: 1/10 (lr=0.01), train_loss=0.919, valid_acc=0.496  
2023-04-07 06:33:34,278 - CiBertLearner - INFO: Local epoch site-8: 1/10 (lr=0.01), train_loss=1.056, valid_acc=0.552  
2023-04-07 06:33:37,312 - CiBertLearner - INFO: Local epoch site-4: 1/10 (lr=0.01), train_loss=1.082, valid_acc=0.523  
2023-04-07 06:33:39,917 - CiBertLearner - INFO: Local epoch site-3: 1/10 (lr=0.01), train_loss=1.010, valid_acc=0.456  
2023-04-07 06:33:40,242 - CiBertLearner - INFO: Local epoch site-1: 1/10 (lr=0.01), train_loss=0.839, valid_acc=0.570  
2023-04-07 06:33:45,291 - CiBertLearner - INFO: Local epoch site-2: 2/10 (lr=0.01), train_loss=0.570, valid_acc=0.559
```

Training cost: 12.7 sec/local epoch

Aggregation

```
2023-04-07 07:17:34,285 - ScatterAndGather - INFO Contribution from site-4 ACCEPTED by the aggregator.
```

Server received site-4's model parameters

Federated loop

```
2023-04-07 07:07:12,885 - DX0Aggregator - aggregating 8 update(s) at round 9  
2023-04-07 07:07:12,899 - ScatterAndGather End aggregation.  
2023-04-07 07:07:12,916 - ScatterAndGather Start persist model on server.  
2023-04-07 07:07:13,012 - ScatterAndGather End persist model on server.  
2023-04-07 07:07:13,013 - ScatterAndGather Round 9 finished.  
2023-04-07 07:07:13,013 - ScatterAndGather Round 10 started.  
2023-04-07 07:07:13,013 - ScatterAndGather scheduled task train
```

Ready for the next round

Fig. 3. Demonstraion example of BERT fine-tuning.

may surpass BERT's performance. Furthermore, overfitting is an issue that plagues large models like BERT, making them susceptible to poor generalization when applied to small datasets. In contrast, LSTM can achieve superior generalization due to its fewer parameters. Finally, the performance disparity may emerge from differences in optimization methods employed during model training, including hyperparameter settings, learning rate, loss functions, and others. Under certain conditions, LSTM may learn more efficiently than BERT.

4) *Demonstrations*: Fig. 3 illustrates the implementation of the NVFlare integrated framework, showcasing a BERT fine-tuning example. Initially, the server and clients are established, with token and SSH-based protocols employed to ensure secure communication channels. Subsequently, each client proceeds to train its model locally, taking an average of 12.7 seconds per local epoch to complete the process. Upon completion, the aggregator gathers the local model parameters from each client. Once all model parameters have been collected, the subsequent federated round commences.

V. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

This paper proposes a practical approach to medical NLP by integrating FL with NVFlare. The proposed framework addresses the challenges of data privacy and regulatory compliance while maintaining high accuracy and performance. Furthermore, this study compares the performance of LSTM-based models to BERT and BERT-mini across various training settings: centralized, FL, and standalone. While LSTM outperforms BERT in the experiments, this outcome may be due to task characteristics, dataset size, overfitting, or optimization-related factors.

Future research directions includes investigating the impact of different tasks and dataset sizes on the performance of LSTM and BERT in medical NLP applications.

REFERENCES

- [1] R. Xu, N. Baracaldo, Y. Zhou, A. Anwar, and H. Ludwig, "Hybridalpha: An efficient approach for privacy-preserving federated learning," in *Proc. of ACM Workshops on Artificial Intelligence and Security*, 2019, pp. 13–23.
- [2] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.
- [3] J. Park, S. Samarakoon, A. Elgabli, J. Kim, M. Bennis, S.-L. Kim, and M. Debbah, "Communication-efficient and distributed learning over wireless networks: Principles and applications," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 796–819, 2021.
- [4] H. Saffri, M. M. Kandi, Y. Miloudi, C. Bortolaso, D. Trystram, and F. Desprez, "Towards developing a global federated learning platform for IoT," in *Proc. IEEE International Conference on Distributed Computing Systems (ICDCS)*, Bologna, Italy, July 2022, pp. 1312–1315.
- [5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [6] NVIDIA, "NvFlare: A framework for federated learning," <https://developer.nvidia.com/nvflare>, 2021.
- [7] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, and S. Bakas, "The future of digital health with federated learning," *npj Digital Medicine*, vol. 3, no. 1, pp. 1–7, 2020.
- [8] Z. C. Lipton, D. C. Kale, C. Elkan, and R. C. Wetzel, "Learning to diagnose with LSTM recurrent neural networks," in *Proc. of International Conference on Learning Representations (ICLR)*, San Juan, Puerto Rico, May 2016.
- [9] "TensorFlow federated tutorial," <https://www.tensorflow.org/federated/>, 2023.
- [10] OpenMined, "PySyft," <https://github.com/OpenMined/PySyft>, 2023.
- [11] FederatedAI, "FATE," <https://github.com/FederatedAI/FATE>, 2023.
- [12] S. Caldas, S. M. K. Duddu, P. Wu, T. Li, J. Konečný, H. B. McMahan, V. Smith, and A. Talwalkar, "LEAF: A benchmark for federated settings," *arXiv preprint arXiv:1812.01097*, 2019.
- [13] I. G. Lee, S. Kim, M. Ban, M. Kim, and J. Chiang, "Predictive models for clopidogrel outcome using prescription records and diagnosis codes," in *Machine Learning for Health Care (MLHC)*, August 2022.