**FEDERATED LEARNING IN CLINICAL DATASET**

**Professor: Dr. Yanmin Gong -**Aman k Yadav

**Abstract:**

*Federated Learning (FL) has emerged as a promising paradigm for collaborative machine learning across decentralized datasets. This report explores the application of Federated Learning in the context of clinical datasets .The goal is to leverage data from multiple healthcare institutions without compromising patient privacy, enabling the development of robust and generalizable models for early detection and prediction of infections.*

**INTRODUCTION:**

Federated Learning is a machine learning approach that enables training models across multiple decentralized edge devices or servers while keeping data localized. It ensures privacy, reduces communication costs, and allows for collaborative model training. In this guide, we will explore the fundamental concepts, applications, advantages, and challenges associated with Federated Learning.

**A diagram of a medical system

Description automatically generated**

Machine learning projects are great to improve our world, to solve problems, and to take informed decisions. This project could help doctors to diagnose diseases of the urinary infection. And then doctors could take the appropriate actions to cure such diseases. In fact, this machine learning system is 100% accurate, whereas human doctors can commit mistakes when diagnosing these diseases. However, forgotten aspects of machine learning are security and privacy. This machine learning is not only very useful and accurate; but it also protects the privacy of datasets in each hospital by using federated learning.

This machine learning system also has a lot of potential for the future. The code of this ML system is a useful pattern that can be copied and extrapolated to more complex kinds of diagnoses. For example, we can change the logistic regression algorithm for a convolutional neural network capable of dealing with datasets of medical images. And the logic to protect the privacy of datasets in each hospital will be the same.

**WORKING PROCESS:**

In this project, I have initialized 4 hospitals showing how actually

federated learning works. The hospitals cannot share the cases of their

patients because they are competitors, and it is necessary to protect the privacy of patients. Hence, the ML model will be learned in a federated way. **A diagram of a network

Description automatically generated**

Federated learning is iterated 1000 times. At each iteration, a copy of the

shared model is sent to all the hospitals. Each hospital trains its own local

model with its own local dataset. Each local model improves a little bit in its own direction. Then we compute the local losses and local accuracies to keep track of them and to make graphs of them. We send the local models to the trusted aggregator that will average all the model updates. This averaged model is the shared model that is sent to all the hospitals at the beginning of each iteration.

A diagram of data processing

Description automatically generated

Figure: Disease processing expert system using FL

In this way, only the ML model will be shared. Whereas the local cases of

each hospital will be kept private, and they will be used to train model.

updates in a local way. Federated learning will protect the privacy of datasets in each hospital and at the same time, we will generate a more robust machine learning model, which will benefit all hospitals. This shared ML model preserves the privacy of individual patients and at the same time, reveals important statistics of stereotypical cases.

**BENEFITS OF USING FEDERATED LEARNING IN CLINICAL DATASET:**

**Privacy-Preserving Collaboration:**

* + **Benefit:** Clinical datasets often contain sensitive patient information. Federated Learning allows healthcare institutions to collaborate on model development without sharing raw patient data, thus preserving privacy.
  + **Example:** Multiple hospitals collaborate to develop a predictive model for patient readmission risk without sharing individual patient records. Each hospital trains the model on its own dataset, contributing only model updates.

**Localized Data Training:**

* + **Benefit:** Federated Learning enables training models on data stored locally at healthcare facilities. This is advantageous in situations where data cannot be easily centralized due to regulatory constraints or logistical challenges.
  + **Example:** In a decentralized healthcare system, rural clinics use Federated Learning to train a model for disease prediction based on their local patient demographics and health records, without the need to centralize data.

**Enhanced Generalization:**

* + **Benefit:** Clinical datasets from different institutions may have unique characteristics. Federated Learning allows models to be trained on diverse datasets, leading to improved generalization across various patient populations.
  + **Example:** Federated Learning is applied to train a diagnostic model across diverse clinical datasets from urban and suburban areas, resulting in a model that generalizes well to different populations.

**Reduced Data Transfer:**

* + **Benefit:** Transmitting raw clinical data over networks can pose security risks. Federated Learning minimizes data transfer by exchanging model updates, reducing the potential for data breaches during communication.
  + **Example:** Rather than sharing entire medical imaging datasets, Federated Learning is employed to train a radiology model collaboratively. Only model updates, such as changes in feature weights, are exchanged to minimize data transfer.

**Customized Models for Local Needs:**

* + **Benefit:** Federated Learning allows healthcare institutions to train models that are tailored to their specific patient populations and healthcare practices, accommodating variations in treatment approaches.
  + **Example:** Hospitals within a network use Federated Learning to develop personalized treatment recommendation models, considering variations in patient demographics, prevalent diseases, and treatment protocols

**DRAWBACKS OF USING FEDERATED LEARNING:**

**Security and Compliance Challenges:**

* + **Drawback:** Healthcare datasets are subject to strict regulatory requirements (e.g., HIPAA). Implementing Federated Learning requires careful consideration of security measures to ensure compliance and protect patient privacy.
  + **Example:** Federated Learning in healthcare must adhere to strict regulations like HIPAA. Ensuring the secure transmission of model updates and maintaining compliance is critical to avoid legal and ethical issues.

**Data Heterogeneity:**

* + **Drawback:** Clinical datasets may vary in terms of formats, standards, and quality. Federated Learning across heterogeneous datasets poses challenges in maintaining a standardized model across different healthcare institutions.
  + **Example:** Federated Learning across institutions with varying electronic health record formats requires addressing differences in data representations to ensure consistency and effective model training.

**Interoperability Issues:**

* + **Drawback:** Ensuring compatibility between different healthcare systems can be challenging. Federated Learning systems need to address interoperability issues to facilitate effective collaboration.
  + **Example:** Integrating Federated Learning systems with existing healthcare IT infrastructure poses challenges. Ensuring seamless communication between different electronic health record systems requires addressing interoperability issues.

**Resource Disparities:**

* + **Drawback:** Healthcare facilities may have varying computational resources and expertise. Federated Learning could be limited by the capabilities of the least resourceful participants, leading to potential inefficiencies.
  + **Example:** Federated Learning collaboration involves both large hospitals and small clinics. The limited computational resources of smaller facilities may impact the efficiency of model training and hinder the collaboration's overall progress.

**Ethical Considerations:**

* + **Drawback:** Federated Learning involves multiple stakeholders, raising ethical concerns related to the equitable distribution of benefits, access to advancements, and potential biases in models developed on diverse datasets.
  + **Example:** The development of Federated Learning models should involve transparent decision-making processes to address ethical concerns. Ensuring equitable distribution of benefits and avoiding biases in model predictions are critical considerations.

**SCREENSHOT OF DATASET:**

**A table of black and white text

Description automatically generatedA screenshot of a computer code

Description automatically generated**

The main idea of this data set is to prepare the algorithm of the expert system, which will perform the presumptive diagnosis of diseases of urinary system. It will be the example of diagnosing of the urinary infection. For better understanding of the problem, let us consider definitions of diseases given by medics.

**Problem in urinary bladder** is characterized by sudden occurrence of pains in the abdomen region and the urination in form of constant urine pushing, micturition pains and sometimes lack of urine keeping. Temperature of the body is rising, however most often not above 38C. The excreted urine is turbid and sometimes bloody. At proper treatment, symptoms decay usually within several days. However, there is inclination to returns. At persons with acute inflammation of urinary bladder, we should expect that the illness will turn into protracted form.

**CODE SNIPPET:**

1. **Define function to read dataset text file.**

First, we need to define functions to download the dataset, to read the dataset text file, and to parse the lines and fields of the dataset:

A screen shot of a computer program

Description automatically generated

1. **Download dataset.**

Then, this download the files related to this dataset: diagnosis.names and diagnosis.data. A screen shot of a computer

Description automatically generated

1. **Define function to parse the line and fields of the dataset.A screenshot of a computer program

   Description automatically generated**
2. **Split dataset into training and testing dataset.**

This defines some functions in order to randomly split this dataset in 2: Training dataset and testing dataset. As we can see, the temperature is parsed into a real number. And the Boolean values (yes or no) are parsed into real numbers as well: 1. and 0., respectively.

**A screenshot of a computer program

Description automatically generated**

**A number grid with numbers

Description automatically generated with medium confidence**

1. **Training function (train model)**

This defines some functions to train the machine learning model while keeping track of the training loss and the training accuracy.

**A screenshot of a computer program

Description automatically generated**

**A white background with black text

Description automatically generatedA graph of a line

Description automatically generated with medium confidence**

1. **Creating virtual workers.**

This start by creating the virtual workers that simulate the computers of each hospital. And then we establish communications among all of them.

**A screenshot of a computer code

Description automatically generatedA close-up of a text

Description automatically generated**

1. **Federated learning function.**

We define some functions to train the machine learning model in a federated way while keeping track of the training loss and the training accuracy, for each hospital separately. The whole process is done in a trusted aggregator, in 1000 iterations. (We can vary the number of iterations.) At each iteration, a copy of the shared model is sent to all the 4 hospitals. Each hospital trains its own local model with its own local dataset, in 5 local iterations. (We can vary the number of local iterations.) Each local model improves a little bit in its own direction. Then we compute the local losses and local accuracies to keep track of them. So, we will be able to create graphs of the learning curves: **Training Losses versus Iterations** and **Training Accuracies versus Iterations**.

**A screenshot of a computer code

Description automatically generated**

**A table of numbers and symbols

Description automatically generated with medium confidence**

**A graph showing the difference between urinary bladder and urinary bladder

Description automatically generatedA graph of a diagram

Description automatically generated with medium confidence**

In summary, the learning curves **Training Losses versus Iterations** and **Training Accuracies versus Iterations** have 4 colors for all 4 hospitals. Each graph has 4 curves of different colors: Blue, orange, green, and red. The curves are not lines; they are rather regions. Because each iteration of federated learning is complex: First, 5 local iterations in each virtual worker (each hospital) to train each local model. Each local model improves a little bit in its own direction. Then, the 4 different models are sent to the trusted aggregator that averages them. Finally, the averaged model is sent back to the 4 hospitals. Such averaged model can have lower performance in comparison to the local models, which are more locally adapted to the local datasets. That's why the progress in the learning curves goes back and forth. Moreover, the graph has 1000 iterations. That's why the curves become regions.

**Conclusion:**

The implementation of Federated Learning to predict urinary infections represents a significant stride toward leveraging collaborative efforts in healthcare data analysis. This approach allows healthcare institutions to pool insights from diverse datasets without compromising patient privacy. The privacy-preserving nature of Federated Learning enables the development of robust predictive models while adhering to stringent healthcare data protection regulations.

Through the collaborative efforts of multiple healthcare facilities, a predictive model for urinary infection has been trained on localized datasets. The model generalizes well across different patient demographics and clinical settings, providing a valuable tool for early detection and intervention in urinary infections. The reduction in data transfer and the ability to customize models for local needs contribute to the efficiency and relevance of the predictive system.

**Future work:**

As we look to the future of this Federated Learning project for predicting urinary infections, several avenues for enhancement and expansion emerge:

**Integration with Real-time Data Streams:**

Explore the integration of real-time patient data streams to continuously update and

improve the predictive model, ensuring its relevance to evolving patient conditions.

**Incorporation of Multi-Modal Data:**

* + 1. Extend the Federated Learning framework to include multiple modalities of patient data, such.
    2. as electronic health records, medical imaging, and laboratory results, to enhance the
    3. comprehensiveness of the predictive model.

**Expansion to Additional Healthcare Institutions:**

* + 1. Collaborate with a broader network of healthcare institutions to further diversify.
    2. the dataset, accommodating variations in clinical practices, patient populations, and
    3. geographic locations.

**Enhanced Security Measures:**

* + 1. Implement advanced security measures to address potential vulnerabilities in the
    2. Federated Learning process, ensuring the robust protection of sensitive patient information.

**References:**

[1] C. Zhu, Z. Xu, Y. Gu, S. Zheng, X. Sun, J. Cao, B. Song, J. Jin, Y. Liu, X.

Wen, et al., Prediction of post-stroke urinary tract infection risk in immobile

patients using machine learning: an observational cohort study, J. Hosp. Infect.

122 (2022) 96–107.

[2] S. Biswas, A. Pal, P. Chakraborty, K. Chaudhury, S. Das, Machine learning basedurinary pH sensing using polyaniline deposited paper device and integration of smart web app interface: Theory to application, Biosens. Bioelectron. 211 (2022)114332.

[3] S.S. Mahdi, G. Battineni, M. Khawaja, R. Allana, M.K. Siddiqui, D. Agha,

does artificial intelligence impact digital healthcare initiatives? a review of AI

applications in dental healthcare, Int. J. Inf. Manag. Data Insights 3 (1) (2023)

100144.

[4] Y. Chen, M. Wang, A.A. Heidari, B. Shi, Z. Hu, Q. Zhang, H. Chen, Mafa

H. Turabieh, Multi-threshold image segmentation using a multi-strategy shuffled

frog leaping algorithm, Expert Syst. Appl. 194 (2022) 116511.

[5] M. Yu, M. Han, X. Li, X. Wei, H. Jiang, H. Chen, R. Yu, Adaptive soft erasur

with edge self-attention for weakly supervised semantic segmentation: thyroid

ultrasound image case study, Comput. Biol. Med. 144 (2022) 105347.

[6] C. Zhao, H. Wang, H. Chen, W. Shi, Y. Feng, JAMSNet: A remote pulse extraction network based on joint attention and multi-scale fusion, IEEE Trans. Circuits Syst.Video Technol. (2022) 1, http://dx.doi.org/10.1109/TCSVT.2022.3227348.

[7] N. Goździkiewicz, D. Zwolińska, D. Polak-Jonkisz, The use of artificial intelli-

gence algorithms in the diagnosis of urinary tract infectionsâ a literature review,

J. Clin. Med. 11 (10) (2022) 2734.

THANK YOU