# Secure Extraction of Personal Information from EHR by Federated Machine Learning

Mohamed El Azzouzi, Reda Bellafqira, Gouenou Coatrieux, Marc Cuggia, Guillaume Bouzillé

*Digital Health and Informatics Innovations for Sustainable Health Care Systems*
*J. Mantas et al. (Eds.)*


*doi:10.3233/SHTI240488*

# Secure Extraction of Personal Information from EHR by Federated Machine Learning

Mohamed EL AZZOUZI [a,1], Reda BELLAFQIRA [b], Gouenou COATRIEUX [b], Marc CUGGIA[a] and Guillaume BOUZILLE [a]
[a] *Univ Rennes, CHU Rennes, INSERM, LTSI-UMR 1099, F-35000, Rennes, France*
[b] *IMT Atlantique, INSERM, LATIM-UMR 1101, F-29200, Brest, France*
ORCiD ID: EL AZZOUZI https://orcid.org/my-orcid?orcid=0009-0006-4741-9820

**Abstract.** Secure extraction of Personally Identifiable Information (PII) from Electronic Health Records (EHRs) presents significant privacy and security challenges. This study explores the application of Federated Learning (FL) to overcome these challenges within the context of French EHRs. By utilizing a multilingual BERT model in an FL simulation involving 20 hospitals, each represented by a unique medical department or pole, we compared the performance of two setups: individual models, where each hospital uses only its own training and validation data without engaging in the FL process, and federated models, where multiple hospitals collaborate to train a global FL model. Our findings demonstrate that FL models not only preserve data confidentiality but also outperform the individual models. In fact, the Global FL model achieved an F1 score of 75,7%, slightly comparable to that of the Centralized approach at 78,5%. This research underscores the potential of FL in extracting PIIs from EHRs, encouraging its broader adoption in health data analysis.

**Keywords.** Federated Learning, Named Entity Recognition, EHRs, NLP

## 1. Introduction

The automatic extraction of Personally Identifiable Information (PII) from Electronic Health Records (EHRs) is generally seen as a Named Entity Recognition (NER) task [1]. The development of NER models suitable for PII extraction requires the creation of large, labeled datasets. However, this task raises significant concerns about data security and confidentiality during the data annotation and learning phases [2].

Furthermore, the effectiveness of these methods depends on the availability of labeled data sets. Unfortunately, annotated data may not be available in a single healthcare institution. Annotating data sufficient for medical NER is costly and time-consuming and requires in-depth knowledge of the medical domain. In addition, relying solely on data from a single healthcare facility limits the applicability and generalizability of NER models. Although many medical institutions may have annotated datasets, they cannot be directly shared, as medical data contains much personal information and is highly sensitive to confidentiality.

[1] Corresponding Author: Mohamed El azzouzi PhD student, Univ Rennes, INSERM, LTSI-UMR 1099, Rennes, France; E-mail: mohamed.elazzouzi@univ-rennes.fr.

Considering these challenges, this article investigates Federated Learning (FL) as a solution [3]. Research shows FL can nearly match centralized models' performance in NER tasks using Bidirectional Long Short Term Memory-Conditional Random Field (Bi-LSTM-CRF) models and the CoNLL2003 dataset [4]. Another study used FL to identify vaccine-related adverse events from the Vaccine Adverse Event Reporting System (VAERS) dataset, highlighting FL's advantages and the negative impact of local differential privacy on accuracy [5]. Historical data usage for creating NER models for PII recognition suggested automatic annotation via sentence matching and explored performance trade-offs in federated training [6]. Further research used pre-trained and fine-tuned Bidirectional Encoder Representations from Transformers (BERT) models in federated settings for clinical texts, noting performance drops due to federated communication loss [7]. A systematic evaluation using six language models and eight corpora showed FL models consistently outperformed individual client data-trained models and performed comparably to centralized training [8]. These findings emphasize the effectiveness and challenges of FL in NER tasks, highlighting the need for ongoing research on data distribution and privacy preservation.

## 2. Methods

### *2.1. Dataset*

We used our de-identification dataset presented in our previous work [2], annotated following the Beginning Inside Outside (BIO) formatting scheme and using data exclusively from the university hospital center of Rennes to extract eight specific entities (Patient, Date, Doctor, Email, Phone, Str: Postal address, Zip codes, and City). It contains two datasets: one automatically annotated and a second manually annotated.

To simulate a federated network of distinct hospitals, we organized the automatically annotated dataset into sub-datasets representing different specialized departments or "poles," such as cardiopulmonary, neurology, and pediatrics. Each department, encompassing several hospital wards, represented an individual hospital in the simulation. We segmented the dataset by department and then further divided each segment into training and validation sets. The manually annotated subset was reserved for testing the NER models.

The federated dataset was diverse, with the number of documents varying significantly across different medical specialties. For instance, the cardiopulmonary department had the highest volume with 162,287 documents, whereas the Service of Physical Medicine and Rehabilitation had only 7,381 documents. Other departments, such as pediatrics and neurology, had intermediate volumes of 72,400 and 49,296 documents respectively, illustrating the varied document distribution across medical specialties. For more detailed information on the distribution of total documents across medical poles for the federated learning dataset, please refer to [9].

### *2.2. Model and framework*

We developed a NER model utilizing the multilingual BERT model to extract PIIs from the clinical text. To ensure data privacy and facilitate collaborative learning across multiple institutions, we implemented a FL framework. This framework utilized the NVIDIA Federated Learning Application Runtime Environment (NVIDIA FLARE) to

simulate a federated environment with 20 clients, employing the Federated Averaging (FedAvg) algorithm for aggregating client model updates.
For our initial experiment, we trained specific models for each hospital, referred to as Individual Models (IND_Model). In this setup, we assume that each hospital exclusively uses its training and validation datasets without participating in the FL process. Subsequently, we implemented a federated learning model, which was collaboratively trained by 20 sites using the FedAvg algorithm. Each site uses its local data to train a Local Federated Model (Local_FL_Model). These Local_FL_Models, unique to each site, then contribute to updating a collective Global Federated Model (Global_FL_Model). After aggregation, the updated Global_FL_Model is redistributed back to each site, serving as a foundation for further training of the Local_FL_Models. This cycle of local training, global aggregation, and redistribution continues until the Global_FL_Model converges to an optimal state. For comparison purposes, a Centralized Model (CENT_Model) was trained using a combined dataset from all 20 sites. This centralized model serves as a baseline to evaluate the performance improvements achieved through federated learning. To provide a clear understanding of the different experiments involved in our study, we present an overview figure in [9] that illustrates the individual learning, federated learning, and centralized learning approaches.

The training parameters for all these models were carefully optimized. The Centralized and individual models were trained for ten epochs using the Adam optimizer. For the FL model, we used a learning rate of $10^{-5}$ and trained the FL model for ten rounds of FedAvg, each local model undergoing training for a predetermined number of aggregation epochs set to 1 for each round of communication. Experiments were performed on a secure server with 112 CPU cores: Intel(R) Xeon(R) Gold 6258R and an NVIDIA A100 40 GB graphics card. We evaluated the models using precision, recall, and F1 score, reporting the micro Avg F1 score. All models were tested on a manually annotated test set of 1,000 documents from Rennes's clinical data warehouse.

## 3. Results and Discussion

In this section, we present and discuss the findings from our experiments using the multilingual BERT model and trained under different setups.

### *3.1. Individual Models (IND_Model) vs. Local Federated Models (Local_FL_Model)*

The analysis of the performance of Individual Models (*IND_Model*) and those trained through FL (*Local_FL_Model*) across various Hospitals reveals interesting results (see Figure 1). Generally, local federated models display superior performance compared to individual models, highlighting the advantage of the collaborative approach for NER tasks. For example, in the pole of pediatric surgery “CHIRURGIE PEDIATRIQUE,” the local federated model achieves an F1 score of 0.7321 versus 0.6240 for the individual model, and similarly in “NEUROLOGY," where the local federated model presents an F1 score of 0.7692 against 0.4944 for the individual model. This trend is consistent across almost all the departments studied, indicating that FL can significantly improve the performance of NER models. It is also important to note that some departments, despite a lower volume of documents, also benefit from this approach. For instance, "NEPHROLOGY SERVICE" and "GERIATRICS," with 12822 and 15881 documents,

respectively, show a notable improvement in the performance of federated models. This demonstrates the effectiveness of FL regardless of the dataset size.

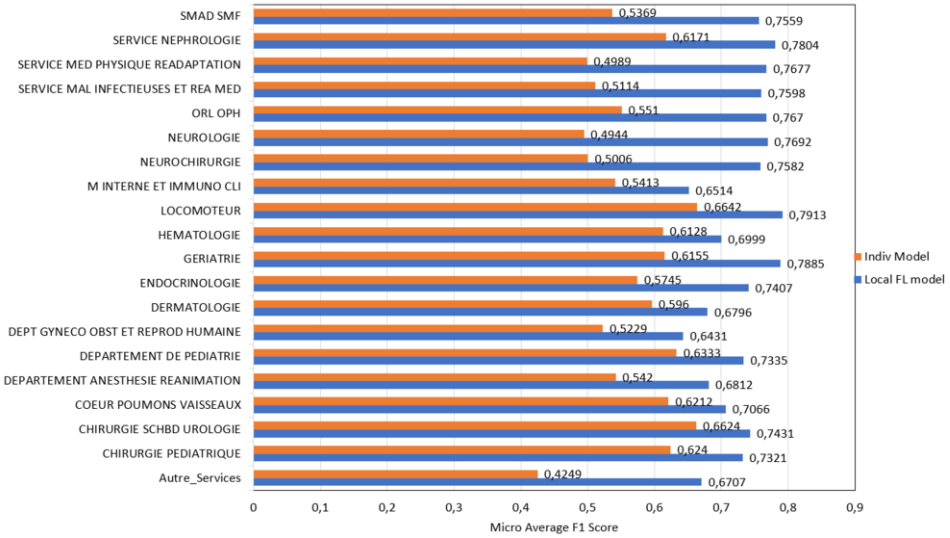


**Figure 1.** Average Micro F1 Score of Individual and Local Federated Models by Site.

### *3.2. Global Federated Learning Model vs. Centralized Model*

Comparative performance measurements of the global FL model versus the centralized model for different labels show that the centralized model generally outperforms the global FL model (see Table 1), with a margin that varies according to the different labels. The global FL model achieves a competitive F1 score of 75.7% versus 78.5% for the centralized approach, which is remarkable given the challenges and constraints associated with secure PII extraction from French EHRs.

**Table 1.** Performance metrics of FL Global Model vs. Centralized Model across various labels.

| | **FL Global Model** | | | **Centralized Model** | | | |
|---|---|---|---|---|---|---|---|
| **Entity** | Precision | Recall | F1-Score | Precision | Recall | F1-Score | **Support** |
| DATE | 0.814 | 0.905 | 0.857 | 0.837 | 0.930 | 0.881 | 1962 |
| DOCTOR | 0.848 | 0.900 | 0.873 | 0.871 | 0.924 | 0.897 | 1199 |
| EMAIL | 0.841 | 0.990 | 0.909 | 0.865 | 1.000 | 0.935 | 96 |
| PATIENT | 0.257 | 0.868 | 0.397 | 0.274 | 0.925 | 0.423 | 507 |
| PHONE | 0.654 | 0.756 | 0.701 | 0.678 | 0.784 | 0.727 | 540 |
| STR | 0.728 | 0.607 | 0.662 | 0.757 | 0.631 | 0.688 | 234 |
| VILLE | 0.708 | 0.945 | 0.810 | 0.729 | 0.973 | 0.833 | 751 |
| ZIP | 0.885 | 0.913 | 0.899 | 0.911 | 0.939 | 0.925 | 287 |
| **micro avg** | 0.664 | 0.881 | **0.757** | 0.690 | 0.909 | **0.785** | 5576 |
| macro avg | 0.717 | 0.860 | 0.764 | 0.740 | 0.888 | 0.789 | 5576 |
| weighted avg | 0.742 | 0.881 | 0.792 | 0.764 | 0.909 | 0.817 | 5576 |

Comparing our results with state-of-the-art approaches highlights the strengths and areas for improvement in our methodology. The systematic evaluation of FL in biomedical NER tasks [8] confirmed the superior performance of FL models over individual client models and comparable results to centralized training, particularly with transformer-based models. Our approach demonstrates the resilience of the BERT-based model in

federated settings, achieving performance comparable to centralized models. Both studies highlight the resilience of transformer models in FL, with our research further validating these findings in the context of French EHRs.

Our study, while pertinent, recognizes some limitations and areas for improvement. Initially, we focused primarily on a BERT-based NER model. Future iterations could benefit from exploring a diverse range of NER models and incorporating a BiLSTM-CRF architecture to potentially improve the performance and generalizability of the FL model. In addition, the issue of data confidentiality in our FL framework remains a concern, as the risk of indirect data leakage during model updates cannot be neglected [3]. Our next objective is to improve the privacy aspect of our FL system by integrating advanced privacy protection techniques [5].

## 4. Conclusions

This study highlights FL's promising potential in enhancing the privacy-preserving extraction of PIIs from French EHRs. By leveraging an FL framework and the multilingual BERT model, we demonstrated that the federated learning model could achieve performance comparable to that of a traditional centralized model while significantly reducing medical data privacy issues. Our results encourage the adoption of FL in clinical environments, enabling more Privacy-Preserving approaches to medical data analysis.

## References

[1] Durango MC, Torres-Silva EA, Orozco-Duque A. Named Entity Recognition in Electronic Health Records: A Methodological Review. Healthc Inform Res. 2023;29:286–300, doi: 10.4258/hir.2023.29.4.286.

[2] Azzouzi ME, Coatrieux G, Bellafqira R, Delamarre D, Riou C, Oubenali N, Cabon S, Cuggia M, Bouzillé G. Automatic de-identification of French electronic health records: a cost-effective approach exploiting distant supervision and deep learning models. BMC Medical Informatics and Decision Making. 2024;24:54, doi: 10.1186/s12911-024-02422-5.

[3] Lansari M, Bellafqira R, Kapusta K, Thouvenot V, Bettan O, Coatrieux G. When Federated Learning Meets Watermarking: A Comprehensive Overview of Techniques for Intellectual Property Protection. Machine Learning and Knowledge Extraction. 2023;5:1382–1406, doi: 10.3390/make5040070.

[4] Mathew J, Stripelis D, Ambite JL. Federated Named Entity Recognition [Internet]. arXiv; 2022 [cited 2023 Mar 20]. Available from: http://arxiv.org/abs/2203.15101.

[5] Kanani P, Marathe VJ, Peterson D, Harpaz R, Bright S. Private Cross-Silo Federated Learning for Extracting Vaccine Adverse Event Mentions. In: Kamp M, Koprinska I, Bibal A, Bouadi T, Frénay B, Galárraga L, Oramas J, Adilova L, Krishnamurthy Y, Kang B, et al., editors. Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Cham: Springer International Publishing; 2021. p. 490–505.

[6] Hathurusinghe R, Nejadgholi I, Bolic M. A Privacy-Preserving Approach to Extraction of Personal Information through Automatic Annotation and Federated Learning [Internet]. arXiv; 2021 [cited 2023 Apr 11]. Available from: http://arxiv.org/abs/2105.09198.

[7] Liu D, Miller T. Federated pretraining and fine tuning of BERT using clinical notes from multiple silos [Internet]. arXiv; 2020 [cited 2023 Mar 20]. Available from: http://arxiv.org/abs/2002.08562.

[8] Peng L, Luo G, Zhou S, Chen J, Xu Z, Sun J, Zhang R. An in-depth evaluation of federated learning on biomedical natural language processing for information extraction. npj Digit Med. 2024;7:1–9, doi: 10.1038/s41746-024-01126-4.

[9] Files · main · Mohamed_Elazzouzi / Annexes-mie-2024 · GitLab [Internet]. GitLab. 2024 [cited 2024 May 23]. Available from: https://gitlab.com/mohamed.elazzouzi/annexes-mie-2024/-/tree/main.