COLLEGE DOCTORAL BRETAGNE SCIENCES DE LA VIE ET DE LA SANTE

Université de Rennes

Titre : Réutilisation de données de santé massives sécurisée par IA : Désidentification par apprentissage automatique pour des systèmes d'information apprenants fédérés.

Mots clés : Dossiers Médicaux Électroniques (DME), Désidentification Automatique, Traitement Automatique du Langage Naturel (TALN), Apprentissage Fédéré.

Résumé : Cette thèse aborde les enjeux et solutions liés à la réutilisation des données de santé, en mettant l'accent sur la protection de la vie privée des patients tout en permettant l'exploitation des données des dossiers médicaux électroniques (DME) pour la recherche clinique et l'amélioration des services de santé. Dans un premier temps, nous explorons le contexte général de la réutilisation des données clinique, en soulignant leur potentiel pour la recherche, tout en identifiant les principaux défis : la protection de la confidentialité, les contraintes réglementaires et les obstacles techniques. Ensuite, nous proposons une approche innovante pour désidentifier automatiquement les DME en français, en conformité avec le RGPD et les directives de la CNIL. En exploitant des techniques avancées d'apprentissage profond et des méthodes de supervision distante, nous avons démontré une solution économiquement viable pour rendre ces données réutilisables en toute sécurité. Les modèles développés, basés sur des représentations linguistiques avancées, montrent des performances prometteuses pour la reconnaissance des entités sensibles dans le texte médical. Dans une autre phase de nos travaux, nous avons étudié l'application de l'apprentissage fédéré (FL) pour l'extraction sécurisée d'informations personnelles à partir des DME. FL permet d'entrainer des modèles collaboratifs entre plusieurs institutions sans centraliser les données sensibles, préservant ainsi la confidentialité des patients.

Nos résultats montrent que les modèles fédérés atteignent des performances proches des modèles centralisés tout en maintenant une protection des données. Par exemple, en utilisant le modèle BERT multilingue dans un environnement fédéré simulant 20 hôpitaux, notre modèle fédéré a obtenu un score F1 de 75,7 %, proche des 78,5 % de l'approche centralisée, mettant en évidence le potentiel de FL pour l'analyse de données de santé tout en réduisant les risques liés à la confidentialité. Enfin, nous avons exploré les vulnérabilités de l'apprentissage fédéré, notamment les attaques exploitant les gradients partagés pour extraire des informations sensibles. Nous avons simulé l'attaque "Decepticons", révélant que des données personnelles telles que les identifiants des patients et des observations médicales peuvent être récupérées avec des taux alarmants allant jusqu'à 90 %. En réponse, nous discutons de contre-mesures telles que la confidentialité différentielle et l'agrégation sécurisée, tout en insistant sur la nécessité d'améliorer ces défenses face à des menaces de plus en plus sophistiquées. Ces travaux ouvrent la voie à des avancées futures pour renforcer la sécurité des systèmes d'apprentissage fédéré, en développant des mécanismes adaptatifs et spécifiquement adaptés aux données médicales.

Title: Secure Reuse of Massive Health Data with AI: De-identification through Machine Learning for Federated Learning Information Systems.

Keywords: Electronic Health Records (EHR), Automatic De-identification, Natural Language Processing (NLP), Federated Learning.

Abstract: This thesis addresses the challenges and solutions related to the reuse of health data, focusing on protecting patients' privacy while enabling the exploitation of electronic health record (EHR) data for clinical research and improving healthcare services. First, we explore the general context of health data reuse, emphasizing its potential for clinical research while identifying key challenges: confidentiality protection, regulatory constraints, and technical obstacles. Next, we propose an innovative approach for the automatic de-identification of French EHRs, in compliance with GDPR and CNIL guidelines. By leveraging advanced deep learning techniques and distant supervision methods, we demonstrated a cost-effective solution to securely render these data reusable. The developed models, based on advanced linguistic representations, show promising performance in recognizing sensitive entities within medical texts. In another phase of our work, we studied the application of federated learning (FL) for the secure extraction of personal information from EHRs. FL allows the training of collaborative models across multiple institutions without centralizing sensitive data, thereby preserving patient confidentiality. Our results show that federated models achieve performance levels close to centralized models while maintaining data protection. For instance, using the multilingual BERT model in an FL environment simulating 20 hospitals, our federated model achieved an F1 score of 75.7%, close to the 78.5% of the centralized approach, highlighting the potential of FL for health data analysis while mitigating privacy risks. Finally, we explored the vulnerabilities of federated learning, particularly attacks exploiting shared gradients to extract sensitive information. We simulated the "Decepticons" attack, revealing that personal data, such as patient identifiers and medical observations, could be retrieved at alarming rates of up to 90%. In response, we discuss countermeasures such as differential privacy and secure aggregation, emphasizing the need to enhance these defenses against increasingly sophisticated threats. This work paves the way for future advancements in strengthening the security of federated learning systems by developing adaptive mechanisms specifically tailored to medical data.