# An integrated shuffler optimizes the privacy of personal genomic data used for machine learning

February 15 2024

KAUST researchers developed a machine-learning approach aimed at preserving privacy while analyzing omics data for medical research. Credit: 2024 KAUST; Heno Hwang

By integrating an ensemble of privacy-preserving algorithms, a KAUST research team has developed a machine-learning approach that addresses

a significant challenge in medical research: How to use the power of artificial intelligence (AI) to accelerate discovery from genomic data while protecting the privacy of individuals.

The study is published in the journal *Science Advances*.

"Omics data usually contains a lot of private information, such as gene expression and cell composition, which could often be related to a person's disease or health status," says KAUST's Xin Gao. "AI models trained on this data—particularly deep learning models—have the potential to retain private details about individuals. Our primary focus is finding an improved balance between preserving privacy and optimizing model performance."

The traditional approach to preserving privacy is to encrypt the data. However, this requires the data to be decrypted for training, which introduces a heavy computational overhead. The trained model also still retains private information and so can only be used in secure environments.

Another way to preserve privacy is to break the data into smaller packets and train the model separately on each packet using a team of local training algorithms, an approach known as local training or federated learning. However, on its own, this approach still has the potential to leak private information into the trained model.

A method called differential privacy can be used to break up the data in a way that guarantees privacy, but this results in a "noisy" model that limits its utility for precise gene-based research.

"Using the differential privacy framework, adding a shuffler can achieve better model performance while keeping the same level of privacy protection; but the previous approach of using a centralized third-party

shuffler that introduces a critical security flaw in that the shuffler could be dishonest," says Juexiao Zhou, lead author of the paper and a Ph.D. student in Gao's group. "The key advance of our approach is the integration of a decentralized shuffling algorithm."

He explains that the shuffler not only resolves this trust issue but achieves a better trade-off between privacy preservation and model capability while ensuring perfect privacy protection.

The team demonstrated their privacy-preserving machine-learning approach (called PPML-Omics) by training three representative deep-learning models on three challenging multi-omics tasks. Not only did PPML-Omics produce optimized models with greater efficiency than other approaches, it also proved to be robust against state-of-the-art cyberattacks.

"It is important to be aware that proficiently trained deep-learning models possess the ability to retain significant amounts of private information from the training data, such as patients' characteristic genes," says Gao. "As deep learning is being increasingly applied to analyze biological and biomedical data, the importance of privacy protection is greater than ever."

**More information:** Juexiao Zhou et al, PPML-Omics: A privacy-preserving federated machine learning method protects patients' privacy in omic data, *Science Advances* (2024). DOI: 10.1126/sciadv.adh8601

Provided by King Abdullah University of Science and Technology

Citation: An integrated shuffler optimizes the privacy of personal genomic data used for machine learning (2024, February 15) retrieved 26 June 2024 from

https://techxplore.com/news/2024-02-shuffler-optimizes-privacy-personal-genomic.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.