Machine Unlearning in Digital Healthcare: Addressing Technical and Ethical Challenges

Shahnewaz Karim Sakib, Mengjun Xie[1]

[1]University of Tennessee at Chattanooga
{shahnewazkarim-sakib, mengjun-xie}@utc.edu

Abstract

The "Right to be Forgotten," as outlined in regulatory frameworks such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), allows individuals to request the deletion of their personal data from deployed machine learning models. This provision ensures that individuals can maintain control over their personal information. In the digital health era, this right has become a critical concern for both patients and healthcare providers. To facilitate the effective removal of personal data from machine learning models, the concept of "machine unlearning" has been introduced. This position paper highlights the technical and ethical challenges associated with machine unlearning in digital healthcare. By examining current unlearning methodologies and their limitations, we propose a roadmap for future research and development in this field.

Introduction

The integration of machine learning technologies in healthcare has significantly transformed the field, enabling precise diagnostics, personalized treatments, and improved patient outcomes (Esteva et al. 2017; Rafiei et al. 2023). However, ensuring the confidentiality and security of these patient data is paramount because of the sensitive nature of these data. Therefore, different privacy preserving techniques, such as differential privacy and federated learning, have been deployed to ensure the usability of data without compromising the patient confidentiality (Choudhury et al. 2019; Chen et al. 2023; Khalid et al. 2023; Rani et al. 2023; Fu et al. 2024). Nevertheless, the challenge of balancing the utility of machine learning models with robust privacy protections remains a critical area of ongoing research (Parks, Wigand, and Benjamin Lowry 2023; Seeman and Susser 2024).

In recent years, the "Right to be Forgotten" has been introduced in various legislative frameworks, such as the General Data Protection Regulation (GDPR) in the European Union (Mantelero 2013) and the California Consumer Privacy Act (CCPA) in the United States (de la Torre 2018). This right ensures that individuals can request companies to remove their private data, which was previously collected for the training of machine learning model. Machine Unlearning (Cao and Yang 2015) was introduced to facilitate the removal of data in compliance with the "Right to be Forgotten." In the context of Machine Unlearning, two critical terminologies are frequently used: *the forget set* and *the retain set*. The forget set refers to the subset of data that needs to be removed to comply with data deletion requests. Conversely, the retain set encompasses the remaining data that continues to influence the model after the unlearning process.

The unlearning process involves two steps: first, the removal of a subset of data that includes the user's data (i.e., the forget set), and second, the erasure of the influence of this data from the trained model. A straightforward approach to achieve the unlearning is to retrain the model from scratch without the forget set. However, given the large-scale datasets used today, this method incurs substantial computational and storage costs.

As an emerging field, machine unlearning in healthcare remains relatively unexplored, presenting unique challenges that need to be addressed. These challenges include maintaining model accuracy and ensuring data security. To effectively implement machine unlearning in healthcare, it is crucial to address the following issues:

1. How do we address the technical challenges associated with machine unlearning in digital healthcare?
2. What are the ethical considerations in implementing machine unlearning in digital healthcare?
3. How can we develop a comprehensive roadmap for future research in machine unlearning for digital healthcare?

Background

Machine Unlearning techniques can be broadly divided into two classes: *approximate machine unlearning* and *exact machine unlearning*.

Approximate Machine Unlearning

Approximate unlearning techniques focus on mitigating the impact of deleted instances by approximating the model parameters as if the deleted data had been absent from the initial training process. There are several approaches to achieving approximate unlearning. Some techniques quantify the influence of an instance and subsequently employ gradient ascent to achieve unlearning (Gupta et al. 2021; Liu et al. 2024; Suriyakumar and Wilson 2022). Other works utilize

methods similar to differential privacy to *approximate* the unlearning (Sekhari et al. 2021; Neel, Roth, and Sharifi-Malvajerdi 2021).

Exact Machine Unlearning

This technique involves creating a modular machine learning system, wherein each component is trained on separate subsets of the data. Consequently, if a deletion request is received, only the specific component needs to be retrained, rather than the entire model. One of the most prominent exact machine unlearning techniques is Sharded, Isolated, Sliced, and Aggregated (SISA) training (Bourtoule et al. 2021). SISA utilizes an ensemble of models, each trained on a distinct shard of the dataset. To further minimize retraining costs, each shard is divided into slices, and training is done on each slice and their checkpoint are stored sequentially. However, recent studies have shown that SISA can increase disparity (Zhang et al. 2024) and reduce fairness by leaking information about minority classes (Koch and Soll 2023).

Unlearning in Digital Healthcare

Unlearning techniques have recently been explored in the context of digital healthcare to adhere to data privacy regulations, particularly in medical imaging (Deng, Luo, and Chen 2024; Nasirigerdeh et al. 2024; Alvandi 2024; Ge 2024). In their study, Nasirigerdeh et al. (Nasirigerdeh et al. 2024) evaluate the performance of various unlearning algorithms within the medical imaging domain. Their findings suggest that although approximate unlearning algorithms perform well for specific retain and forget sets, they lack generalizability. In a different approach, Deng et al. (Deng, Luo, and Chen 2024) introduce Federated Client Unlearning (FCU), which employs model-contrastive unlearning and frequency-guided memory preservation to efficiently remove specific client data while maintaining the overall performance of the model.

Technical Challenges

There are several technical challenges associated with machine unlearning in the context of digital healthcare. The most prominent ones are listed below:

1. Exact unlearning, which involves retraining only on the retain set, is highly effective but incurs significant computational costs. If a patient requests their data to be forgotten, performing exact unlearning would require substantial computational resources, leading to delays in providing diagnostic support and increased operational costs. Additionally, exact unlearning has been shown to leak information regarding minority classes (Koch and Soll 2023), which can be disastrous in the context of healthcare.
2. Existing unlearning algorithms can negatively impact the generalization of the model, particularly for larger forget sets. For example, when a machine learning model is asked to forget certain patients' data, it might underperform for those with similar characteristics (e.g., certain age groups or medical conditions), thereby affecting the overall accuracy and fairness of the model.
3. Unlearning mechanisms can obscure the rationale behind model updates, making it challenging for healthcare professionals to interpret and trust the decisions made by the model. This lack of transparency can hinder adoption and raise concerns about accountability and patient safety.
4. Integrating machine unlearning algorithms into existing clinical workflows poses logistical and organizational challenges. Consider an electronic health record (EHR) system that utilizes AI models to predict disease outbreaks. Implementing machine unlearning in this system requires careful coordination to ensure that data removal requests are processed without disrupting ongoing analyses or data flow. This integration challenge can strain resources and affect the overall functionality of the EHR system.

Ethical Considerations

When implementing machine unlearning in digital healthcare, it is crucial to address several ethical considerations to ensure responsible and fair adoption. These considerations include, but are not limited to the following:

1. It is critical to ensure that patient data remains protected against adversarial attacks during the unlearning process. Care must be taken to prevent the inadvertent exposure of sensitive information.
2. Maintaining transparency in the unlearning process and holding entities accountable for its correct implementation is essential. Patients should be fully informed about the unlearning procedures and their implications.
3. The unlearning process must be designed to avoid introducing or perpetuating biases within the remaining data or models. Its impact on fairness in healthcare outcomes should be thoroughly assessed.
4. The unlearning process must adhere to all relevant laws and regulations, including data protection laws such as GDPR and HIPAA. Ensuring legal compliance is critical to avoid legal repercussions and maintain trust.
5. Continuous monitoring and evaluation of the unlearning process and its outcomes are necessary to identify any unintended consequences or emerging ethical issues.

Recommendations for Future Directions

Based on the identified technical challenges and ethical considerations, we recommend the following future directions for exploring unlearning in digital healthcare:

Algorithm Development and Efficiency

- Develop unlearning algorithms that minimize computational costs, ensuring efficient data removal without excessive resource consumption.
- Focus on methods that enhance the generalization capabilities of models post-unlearning, addressing potential declines in accuracy and fairness.

Enhancing Transparency and Trust

- Design unlearning mechanisms that enhance the transparency of model updates to ensure healthcare professionals can trust and interpret the decisions made by the model.
- Provide patients with detailed information about the unlearning process, its benefits, and its implications for their data and healthcare outcomes.
- Establish clear guidelines and accountability measures for entities implementing unlearning procedures to ensure ethical compliance.

Integration and Implementation

- Develop tools for integrating unlearning algorithms into existing clinical systems and ensure that data removal requests are processed efficiently, maintaining the continuous flow of clinical data and analytics.
- Implement regular audits and robust security measures to prevent inadvertent exposure of sensitive information during the unlearning process, including protecting patient data from adversarial attacks.

Validation and Verification

- Develop rigorous testing and validation protocols to confirm the complete removal of specified data.
- Use advanced verification techniques to ensure no residual traces of unlearned data remain.

Acknowledgments

This work was supported in part by the National Science Foundation (awards 1663105 and 2234910).

References

Alvandi, A. O. 2024. Machine Unlearning in Digitalized Healthcare Arena A Comprehensive Exploration.

Bourtoule, L.; Chandrasekaran, V.; Choquette-Choo, C. A.; Jia, H.; Travers, A.; Zhang, B.; Lie, D.; and Papernot, N. 2021. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, 141–159. IEEE.

Cao, Y.; and Yang, J. 2015. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, 463–480. IEEE.

Chen, R. J.; Wang, J. J.; Williamson, D. F.; Chen, T. Y.; Lipkova, J.; Lu, M. Y.; Sahai, S.; and Mahmood, F. 2023. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature biomedical engineering*, 7(6): 719–742.

Choudhury, O.; Gkoulalas-Divanis, A.; Salonidis, T.; Sylla, I.; Park, Y.; Hsu, G.; and Das, A. 2019. Differential privacy-enabled federated learning for sensitive health data. *arXiv preprint arXiv:1910.02578*.

de la Torre, L. 2018. A Guide to the California Consumer Privacy Act of 2018.

Deng, Z.; Luo, L.; and Chen, H. 2024. Enable the Right to be Forgotten with Federated Client Unlearning in Medical Imaging. *arXiv preprint arXiv:2407.02356*.

Esteva, A.; Kuprel, B.; Novoa, R. A.; Ko, J.; Swetter, S. M.; Blau, H. M.; and Thrun, S. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639): 115–118.

Fu, J.; Hong, Y.; Ling, X.; Wang, L.; Ran, X.; Sun, Z.; Wang, W. H.; Chen, Z.; and Cao, Y. 2024. Differentially private federated learning: A systematic review. *arXiv preprint arXiv:2405.08299*.

Ge, L. 2024. Erasing memories: implementing client unlearning in medical image analysis. In *International Conference on Image Processing and Artificial Intelligence (ICIPAI 2024)*, volume 13213, 955–960. SPIE.

Gupta, V.; Jung, C.; Neel, S.; Roth, A.; Sharifi-Malvajerdi, S.; and Waites, C. 2021. Adaptive machine unlearning. *Advances in Neural Information Processing Systems*, 34: 16319–16330.

Khalid, N.; Qayyum, A.; Bilal, M.; Al-Fuqaha, A.; and Qadir, J. 2023. Privacy-preserving artificial intelligence in healthcare: Techniques and applications. *Computers in Biology and Medicine*, 158: 106848.

Koch, K.; and Soll, M. 2023. No matter how you slice it: Machine unlearning with sisa comes at the expense of minority classes. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 622–637. IEEE.

Liu, J.; Lou, J.; Qin, Z.; and Ren, K. 2024. Certified minimax unlearning with generalization rates and deletion capacity. *Advances in Neural Information Processing Systems*, 36.

Mantelero, A. 2013. The EU proposal for a general data protection regulation and the roots of the 'right to be forgotten'. *Computer Law & Security Review*, 29(3): 229–235.

Nasirigerdeh, R.; Razmi, N.; Schnabel, J. A.; Rueckert, D.; and Kaissis, G. 2024. Machine Unlearning for Medical Imaging. *arXiv preprint arXiv:2407.07539*.

Neel, S.; Roth, A.; and Sharifi-Malvajerdi, S. 2021. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic Learning Theory*, 931–962. PMLR.

Parks, R. F.; Wigand, R. T.; and Benjamin Lowry, P. 2023. Balancing information privacy and operational utility in healthcare: proposing a privacy impact assessment (PIA) framework. *European Journal of Information Systems*, 32(6): 1052–1069.

Rafiei, A.; Moore, R.; Jahromi, S.; Hajati, F.; and Kamaleswaran, R. 2023. Meta-learning in healthcare: A survey. *arXiv preprint arXiv:2308.02877*.

Rani, S.; Kataria, A.; Kumar, S.; and Tiwari, P. 2023. Federated learning for secure IoMT-applications in smart healthcare systems: A comprehensive review. *Knowledge-based systems*, 274: 110658.

Seeman, J.; and Susser, D. 2024. Between privacy and utility: On differential privacy in theory and practice. *ACM Journal on Responsible Computing*, 1(1): 1–18.

Sekhari, A.; Acharya, J.; Kamath, G.; and Suresh, A. T. 2021. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34: 18075–18086.

Suriyakumar, V.; and Wilson, A. C. 2022. Algorithms that approximate data removal: New results and limitations. *Advances in Neural Information Processing Systems*, 35: 18892–18903.

Zhang, D.; Pan, S.; Hoang, T.; Xing, Z.; Staples, M.; Xu, X.; Yao, L.; Lu, Q.; and Zhu, L. 2024. To be forgotten or to be fair: Unveiling fairness implications of machine unlearning methods. *AI and Ethics*, 4(1): 83–93.