Navigating Unlearning in Medical AI: A Framework for Diabetic Retinopathy Classification

Xinghao Li, Chi Liu, *Member, IEEE*, Youyang Qu, *Senior Member, IEEE*, Shujie Cui, Cunjian Chen, *Senior Member, IEEE*, Xingliang Yuan, *Senior Member, IEEE*, Zongyuan Ge, *Senior Member, IEEE*, and Longxiang Gao, *Senior Member, IEEE*

***Abstract*—Medical Artificial Intelligence (AI) systems hold great potential for clinical and diagnostic applications. However, their deployment raises significant privacy and ethical concerns, primarily due to the reliance on large-scale datasets containing sensitive health information. The processes of collecting, storing, and using patient data bring forth widespread concerns about privacy protection and data ownership. Additionally, the dynamic nature of medical data presents risks related to outdated or inaccurate information, which can directly impact patient care. Introducing the concept of "unlearning" in medical AI is therefore crucial, not only to address patient privacy needs but also to ensure that models remain adaptable to evolving medical knowledge and data quality. Nevertheless, implementing unlearning in medical AI is considerably more complex compared to other domains, as errors in this context could lead to severe consequences for patient health. This paper aims to establish an evaluation framework specifically for unlearning in classification tasks on the Mobile Brazilian Retinal Dataset (mBRSET) in medical AI. The framework addresses key issues such as privacy, clinical safety, diagnostic accuracy, and fairness. By providing a detailed foundation for assessing unlearning practices, the framework is tailored to the unique challenges presented by the mBRSET dataset. Ultimately, this work seeks to identify the trade-offs and requirements necessary to implement responsible unlearning processes in this highly sensitive domain, thereby contributing to the safe and ethical advancement of medical AI.**

***Index Terms*—Machine unlearning, medical artificial intelligence.**

I. Introduction

With the continuous advancement of deep learning technology, research in medical Artificial Intelligence (AI) has become increasingly sophisticated [1], [2], [3]. Medical AI has the potential to improve healthcare quality and enhance patient-centered care by providing more accurate diagnoses, optimizing workflows, and simplifying administrative tasks [4]. Studies indicate that using large and diverse datasets for model training can significantly enhance the accuracy and generalizability of these models, making them more effective in various clinical settings [5], [6]. For instance, incorporating patient data from multiple hospitals and geographic locations improves model performance across diverse regions, and datasets that include a wide range of disease characteristics help the models better handle rare or uncommon conditions. Moreover, including diverse populations, such as different races, genders, and age groups, helps to reduce prediction bias and improve fairness, ensuring equitable high-quality diagnosis and treatment for all patients. However, the use of such large-scale datasets also raises significant ethical and regulatory concerns.

In medical AI, data privacy and informed consent are critical ethical considerations [7], [8]. Patients must provide informed consent for the use of their data and retain the right to withdraw that consent at any time. Additionally, data bias can lead to implicit biases within models that negatively impact specific groups, resulting in unfair diagnoses and treatment, thus raising serious ethical concerns about discrimination [9]. Moreover, AI models are vulnerable to data extraction attacks, where attackers might infer original training data by accessing the model, thereby compromising patient privacy. From a legal standpoint, the General Data Protection Regulation (GDPR) [10] in the European Union and the Health Insurance Portability and Accountability Act (HIPAA) [11] in the United States set clear guidelines on the collection and processing of medical data, ensuring privacy and safeguarding user control over personal information. These ethical and legal challenges, coupled with the risk of data extraction attacks, highlight the importance of equipping medical

Received 21 July 2025; accepted 25 September 2025. This work was supported in part by the National Science and Technology Major Project under Grant 2022ZD0116800, in part by the Qilu University of Technology (Shandong Academy of Sciences) Major Project under Grant 2025ZDZX02, in part by Qilu University of Technology (Shandong Academy of Sciences) Youth Outstanding Talent Program under Garnt 2024QZJH02, in part by Shandong Provincial Natural Science Foundation under Grant ZR2023QF152 and Grant ZR2023QF104, and in part by Shandong Provincial University Youth Innovation and Technology Support Program under Grant 2022KJ291. *(Corresponding authors: Youyang Qu; Longxiang Gao.)*

Xinghao Li, Shujie Cui, and Zongyuan Ge are with the Faculty of Information Technology (IT), Monash University, Clayton, VIC 3800, Australia (e-mail: xinghao.li@monash.edu; shujie.cui@monash.edu; zongyuan.ge@monash.edu).

Chi Liu is with the Faculty of Data Science, City University of Macau, Macao SAR, China (e-mail: chiliu@cityu.edu.mo).

Youyang Qu and Longxiang Gao are with the Shandong Academy of Sciences, Qilu University of Technology, Jinan 250353, China, also with the Shandong Computer Science Center, National Supercomputer Center in Jinan, Jinan 250101, China, and also with the Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science, Jinan 250014, China (e-mail: youyang.qu@data61.csiro.au; gaolx@sdas.org).

Cunjian Chen is with the Department of Data Science and AI, Monash University, Melbourne, VIC 3800, Australia (e-mail: Cunjian.Chen@monash.edu).

Xingliang Yuan is with the School of Computing and Information Systems, The University of Melbourne, Parkville, VIC 3052, Australia (e-mail: xingliang.yuan@unimelb.edu.au).

Recommended for acceptance by L. Feng.

Digital Object Identifier 10.1109/TETCI.2025.3641713

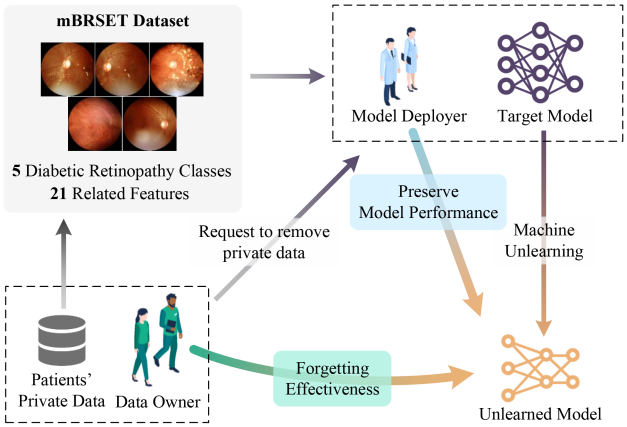


Fig. 1. Machine unlearning in medical AI scenario. The medical AI target model is trained on the Healthcare dataset. Patients, as data owners, reserve the right to request the removal of their data from the model by the model deployer. This request can be fulfilled using machine unlearning methods.

AI models with the ability to "forget" specific data. However, removing data from a model is inherently challenging, as machine learning models deeply integrate and abstract information during training, making it difficult to completely eliminate the influence of specific data once it has been learned.

To address this issue, researchers have introduced the concept of machine unlearning [12], [13], [14]. Broadly defined, machine unlearning refers to the process of removing specific data from a model so that its behavior resembles that of a model that has never been exposed to this data. The goal is to ensure that the deleted data no longer influences the model's predictions and behavior, thereby protecting data privacy and ensuring control over data while mitigating biases. Bourtoule et al. [13] were the first to propose a set of criteria for evaluating unlearning schemes, which include unlearning effectiveness, post-unlearning performance comparison, and training time efficiency, establishing a foundational framework for assessing the efficacy of unlearning.

In medical AI, the demand for unlearning is far greater than in general applications. Models used in general machine learning applications, such as text classification or recommendation systems, typically do not need to meet the high levels of precision, interpretability, or regulatory compliance required in medical systems. Fig. 1 illustrates the machine unlearning scenario in medical AI. Due to the high sensitivity of medical data, any operational error can directly impact patient health and safety [15]. Therefore, when implementing unlearning in medical AI, it is crucial to ensure that clinical accuracy is not adversely affected. Additionally, the medical model must avoid introducing new biases while eliminating the influence of specific data, ensuring fairness across all patient groups. This implies that unlearning in medical AI must strictly adhere to privacy protection standards and undergo rigorous validation to ensure that diagnostic capability and fairness remain consistent after unlearning. These higher demands make machine unlearning in the medical field significantly more complex and critical compared to other application areas. However, machine unlearning in medical AI is still in its early stages, and existing methods are primarily evaluated based on general standards for effectiveness and reliability. To address these challenges, we propose an evaluation framework specifically designed for medical AI models, focusing on the classification task on the Mobile Brazilian Retinal Dataset (mBRSET) [16], addressing their unique requirements and evaluating unlearning in terms of effectiveness, efficiency, potential risks, and clinical performance of the model after unlearning. This comprehensive evaluation aims to provide a solid theoretical and practical foundation for the safe and compliant development of medical AI.

In summary, we make the following major contributions to this paper:

- We conduct a systematic review of existing machine unlearning methods and identify the specific goals for machine unlearning in medical diabetic retinopathy classification, proposing a universal framework.
- We develop standardized evaluation metrics for machine unlearning in medical diabetic retinopathy classification, which measure unlearning effectiveness, classification accuracy, and unlearning efficiency.
- We benchmark the most representative methods within each category of machine unlearning using the proposed evaluation metrics, validating the significance of evaluation metrics and providing insights into the strengths, weaknesses, and unlearning mechanisms.

The rest of this paper is organized as follows. The preliminaries of machine unlearning are presented in Section II. Section III introduces five unlearning methods and an unlearning performance evaluation benchmark. The unique challenges and objectives for unlearning in medical AI are shown in Section IV. The proposed framework for medical AI unlearning is in Section V, and the experiment results are in Section VI. Section VII discusses the limitations of our current framework and outlines directions for future work. Finally, Section VIII summarizes and concludes this paper.

II. Preliminaries: Machine Unlearning

Machine unlearning refers to effective algorithms designed for machine learning models to forget specific target data [17]. To forget specific data points, the model needs to effectively remove the influence of the data, aiming to erase the characteristics derived from the data. Given the increasing emphasis on data privacy from both legal and ethical perspectives, many researchers are actively exploring machine unlearning. Generally, unlearning is categorized into two types: exact unlearning and approximate unlearning [12]. Exact unlearning guarantees that the model's behavior is identical to a model that was never trained on the unlearned data, effectively removing all traces of its influence. In contrast, approximate unlearning aims to remove the impact of data to a practical extent, achieving similar behavior as if the data were not used, but without the computational burden and precision required for exact unlearning, often striking a balance between efficiency and effectiveness.

- *Exact Unlearning [18], [19]:* Let f be the training algorithm, D the full dataset, $D_f \subseteq D$ the data to be unlearned, and $D_r = D \setminus D_f$. An unlearning procedure outputs parameters θ' such that $\theta' \overset{d}{=} f(D_r)$, i.e., the post-unlearning model is distributionally indistinguishable from training

from scratch on D_r. A trivial way to satisfy this is to actually retrain on D_r, which is computationally expensive; efficient exact unlearning without full retraining is generally hard except in special cases.

- *Approximate Unlearning [20], [21]:* Given model parameters θ and the sub-dataset D_f to be unlearned, approximate unlearning involves finding a new set of parameters $\tilde{\theta}$ such that the behavior of the model is similar to θ' obtained by retraining without D_f, but with significantly less computational cost. The goal here can be described as minimizing the distance $||f(D_r) - \tilde{\theta}||$, where $||\cdot||$ is an appropriate norm (e.g., Euclidean distance). This ensures that the influence of D_f is minimized, without needing full retraining.

Approximate unlearning aims to simulate the effect of retraining without requiring full access to the original training process. Common strategies, such as influence function–based parameter adjustments or model perturbations, are employed to reduce the influence of the unlearned data while preserving predictive performance. Compared to exact unlearning, these methods offer a practical trade-off among computational efficiency, model utility, and unlearning effectiveness, making them particularly suitable for real-world machine learning systems, especially in medical applications where model updates must be both efficient and reliable.

III. Unlearning Methods and Unlearning Performance Evaluation Benchmark

With the growing emphasis on data privacy, research in machine unlearning has increasingly gained widespread attention. In this section, we introduce five widely recognized machine unlearning methods and provide a detailed analysis using existing evaluation criteria.

A. SISA Training Framework for Unlearning

The SISA (Sharded, Isolated, Sliced, and Aggregated) training framework [13] is an approach designed to facilitate efficient unlearning in machine learning models by minimizing the computational cost associated with removing dataset D_f.

The original training dataset D is divided into S disjoint shards:

$$D = \bigcup_{s=1}^{S} D_s, \tag{1}$$

where $D_s \cap D_{s'} = \emptyset$ for $s \neq s'$, and $s, s' \in S$. Each shard D_s is further divided into K slices:

$$D_s = \bigcup_{k=1}^{K} D_{s,k}, \tag{2}$$

where $D_{s,k} \cap D_{s,k'} = \emptyset$ for $k \neq k'$, and $k, k' \in K$. For each shard s, a model M_s is initialized and trained sequentially on its slices:

$$M_s^{(k)} = f(M_s^{(k-1)}, D_{s,k}), \quad k = 1, 2, \ldots, K, \tag{3}$$

where $M_s^{(k)}$ represents the model state after training on slice $D_{s,k}$. To unlearn a data point $x_i \in D_f = \{(x_i, y_i)\}_{i=1}^{F}$, identify the shard s and slice k containing x_i. Retraining starts from the saved state $M_s^{(k-1)}$, excluding x_i from $D_{s,k}$, and continues sequentially through subsequent slices:

$$M_s^{(k+1)}, M_s^{(k+2)}, \ldots, M_s^{(K)}. \tag{4}$$

This process effectively removes the influence of x_i without requiring full retraining on the entire dataset.

The final model M is obtained by aggregating all shard models $\{M_s\}_{s=1}^{S}$. Aggregation can be performed using methods such as majority voting or averaging, depending on the specific application.

This structured approach ensures that unlearning a specific data point requires retraining only a fraction of the model, significantly enhancing efficiency compared to retraining the entire model from scratch. The SISA framework is particularly beneficial for large-scale models and datasets, where full retraining would be computationally prohibitive.

B. Deltagrad: Rapid Retraining of Machine Learning Models

The DeltaGrad algorithm, as proposed by Wu et al. [22], aims to efficiently implement unlearning by leveraging the cached information from the initial training phase. This method is particularly effective for scenarios where small changes, such as the addition or deletion of a small subset of data, occur in the training dataset. The core idea of DeltaGrad is to approximate the updated model parameters after modifying the training data without retraining from scratch. This is achieved by utilizing the gradients and the Hessian matrix information computed during the original training process.

Let θ represent the model parameters trained on the original dataset D. When a small subset of data ΔD is added or removed, DeltaGrad estimates the updated parameters θ' as follows:

$$\theta' \approx \theta - H^{-1} \nabla_\theta L(\Delta D; \theta), \tag{5}$$

where H denotes the Hessian matrix of the loss function L with respect to θ, representing the second-order partial derivatives and providing information about the curvature of the loss function. $\nabla_\theta L(\Delta D; \theta)$ represents the gradient of the loss function evaluated on the subset ΔD, which is the data to be either added or removed. In practice, H is used as an *approximate* curvature preconditioner obtained from the original training, and the above Newton-style update serves as a warm start that can be followed by a few corrective gradient steps (the sign and scale depend on whether data are added or removed). By using this approximation, DeltaGrad enables the rapid retraining of machine learning models, effectively allowing for unlearning without the high computational cost associated with full retraining. This approach is particularly advantageous in cases where data privacy concerns require the efficient removal of specific data points from the training dataset.

C. Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks

Golatkar et al. [23] propose a method to selectively remove information about specific training data from a deep neural network without retraining from scratch. This is achieved by modifying the network's weights to eliminate traces of the data

that need to be unlearned. The core idea is to apply a scrubbing function $S(w)$ to the network's weights w, such that any probing function of the weights becomes indistinguishable from the same function applied to the weights of a network trained without the data being unlearned. This ensures that the influence of the unlearned data is effectively removed.

Mathematically, let w represent the weights of a network trained on the dataset D, and D_f be the subset of data to be unlearned. The goal is to find a scrubbing function $S(w)$ such that the distribution of the scrubbed weights $S(w)$ given the dataset D is indistinguishable from the distribution of weights w' trained on the dataset $D \setminus D_f$:

$$P(S(w) \mid D) \approx P(w' \mid D \setminus D_f) \tag{6}$$

To achieve this, the authors introduce an upper bound on the amount of information retained in the weights about the data to be unlearned. This bound can be estimated efficiently, even for deep neural networks, and is used to guide the scrubbing process. The approach leverages the stability properties of stochastic gradient descent (SGD) to ensure that the scrubbing effectively removes information about the unlearned data while preserving the model's performance on the remaining data. By employing these techniques, the method seeks to strike a balance between effective unlearning and preserving the model's overall generalization capability.

D. Certified Data Removal From Machine Learning Models

Guo et al. [21] propose a mechanism for certified removal of data from machine learning models, ensuring that the model's behavior becomes indistinguishable from one that was never trained on the removed data. The core idea is to define a certified removal mechanism M that, when applied to a model h trained on dataset D, produces a new model h' such that, for any subset T of the hypothesis space H, the following holds:

$$e^{-\epsilon} \leq \frac{P(M(h, D, x_i) \in T)}{P(A(D \setminus \{x_i\}) \in T)} \leq e^{\epsilon}, \tag{7}$$

where A denotes the learning algorithm; x_i is the data point to be removed; ϵ is a small positive constant that bounds the divergence between the original and updated models. For brevity, we present the guarantee in ϵ-form, but the original mechanism is (ϵ, δ)-style and assumes a randomized training algorithm for certification. This inequality ensures that the probability distributions of the models before and after data removal are close, thereby providing a strong guarantee of effective data removal. To achieve this, the authors develop a certified removal mechanism for L_2-regularized linear models trained using differentiable convex loss functions. The mechanism involves: 1) Applying a Newton step to the model parameters to remove the influence of the data point x_i. 2) Adding a random perturbation to mask any residual effects, thereby ensuring that the model's behavior aligns with one that was never trained on x_i.

The Newton step is defined mathematically as:

$$\theta' = \theta - H^{-1} \nabla_\theta L(x_i; \theta), \tag{8}$$

where θ are the current model parameters; H is the Hessian matrix of the loss function L with respect to θ, providing information about the curvature; $\nabla_\theta L(x_i; \theta)$ is the gradient of the loss function evaluated at the data point x_i. By leveraging this certified removal mechanism, the resulting model's parameters θ' effectively remove the influence of x_i, thus aligning with the intended guarantee of unlearning while retaining model performance on the remaining data. In deep settings, practical implementations rely on curvature approximations.

E. Machine Unlearning of Features and Labels

Warnecke et al. [24] introduce a novel framework for removing sensitive features and labels from machine learning models, aiming for provable guarantees in convex settings, without the need to retrain from scratch. The core of their method lies in leveraging influence functions to retroactively modify the model parameters such that the effects of specific features or labels are removed.

The following inequality is used to describe the behavior of the updated model θ':

$$e^{-\epsilon} \leq \frac{P(U(\theta, D, Z \to Z') \in T)}{P(A(D') \in T)} \leq e^{\epsilon}, \tag{9}$$

where A is the original learning algorithm, D represents the original dataset, Z denotes the features or labels to be removed, Z' is the modified dataset with the sensitive information removed, and ϵ is a small constant that bounds the divergence between the distributions of the retrained model and the updated model. These bounds hold under strongly convex losses. This certified guarantee ensures that the model's behavior after unlearning is statistically close to that of a model that never included the sensitive data.

The unlearning mechanism involves two key update strategies:

First-Order Update: The first-order update uses the gradient of the loss function to modify the model parameters, thereby removing the influence of specific features or labels. This update strategy is computationally efficient and is defined as follows:

$$\Delta(\theta) = -\tau \sum_{z \in Z} (\nabla_\theta L(z; \theta) - \nabla_\theta L(z'; \theta)), \tag{10}$$

where τ is the unlearning rate, $\nabla_\theta L(z; \theta)$ represents the gradient of the loss function with respect to the model parameters θ, and z and z' are the original and modified data points, respectively. This approach is particularly useful when an approximate correction suffices, and it is suited for high-dimensional models where computational efficiency is crucial.

Second-Order Update: The second-order update incorporates the Hessian matrix of the loss function, providing a more precise parameter adjustment that accounts for the curvature of the loss surface. The update is expressed as:

$$\Delta(\theta) = -H^{-1} \sum_{z \in Z} (\nabla_\theta L(z; \theta) - \nabla_\theta L(z'; \theta)), \tag{11}$$

where H is the Hessian matrix of the loss function L with respect to the parameters θ. The use of the Hessian allows for a more nuanced update that closely matches the effects of retraining

without the sensitive data. For models with strongly convex loss functions, this second-order approach not only achieves effective unlearning but also provides a certified bound on the difference between the updated model and a model retrained from scratch.

By introducing closed-form updates for model parameters using influence functions, Warnecke et al. provide a scalable solution to the problem of data removal in machine learning. Their approach allows for the efficient correction of specific features and labels.

F. Unlearning Performance Evaluation Benchmark

Bourtoule et al. [13] first proposed a foundational framework for evaluating machine unlearning strategies, establishing six key goals that provide a comprehensive basis for assessing unlearning approaches. These six goals are designed to evaluate the unlearning scheme from three main aspects: effectiveness of unlearning, maintenance of model performance, and efficiency of the unlearning process. The specific goals are as follows:

- *Intelligibility (G1):* The strategy should be easy to understand and implement, ensuring that even non-experts can verify the unlearning process.
- *Comparable Accuracy (G2):* The model's accuracy after unlearning should be comparable to that of the baseline retraining strategy. Specifically, the model's performance should not degrade significantly, even when a large number of data points are unlearned.
- *Reduced Unlearning Time (G3):* The strategy should have a provably lower unlearning time compared to retraining the model from scratch, thus significantly reducing the required time for unlearning.
- *Provable Guarantees (G4):* The strategy should provide provable guarantees that the influence of unlearned data has been effectively removed, similar to the behavior of a model that has never seen the unlearned data.
- *Model Agnostic (G5):* The strategy should be model-agnostic, meaning that it should be applicable to a wide variety of models with different types and levels of complexity.
- *Limited Overhead (G6):* The strategy should minimize additional computational overhead, avoiding any significant increase in the already computationally intensive training procedures.

These goals collectively provide a robust framework for evaluating unlearning strategies in general situations, ensuring that they are effective, efficient, and applicable in diverse scenarios while maintaining model accuracy and providing necessary guarantees of data removal. Although this evaluation framework has been proposed, most of the existing work does not strictly adhere to each individual goal when demonstrating the effectiveness of unlearning methods. However, these studies still primarily evaluate the proposed approaches based on the three main aspects of the framework: the effectiveness of unlearning, the maintenance of model performance, and the efficiency of the unlearning process. Table I provides a comparison of the six unlearning methods discussed earlier, based on the aforementioned framework. The result shows that the Retrain method has limitations on the training time and computational resource costs, and the other methods, except for SISA, would have a bad influence on the model performance after unlearning.

TABLE I
EVALUATION OF FOUR MACHINE UNLEARNING METHODS BASED ON FRAMEWORK OF [13]. ✓ = CRITERION SATISFIED; ↓ = DEGRADED.

	G1	G2	G3	G4	G5	G6
Retrain	✓	✓	↓	✓	↓	✓
SISA	✓	✓	✓	✓	✓	✓
DeltaGrad	✓	↓	✓	✓	✓	✓
FIM	✓	↓	✓	✓	✓	✓
CDR	✓	↓	✓	✓	✓	✓
FU	✓	↓	✓	✓	✓	✓

IV. MEDICAL AI: UNIQUE CHALLENGES AND OBJECTIVES FOR UNLEARNING

The medical field presents unique and stringent requirements for unlearning compared to other domains, primarily due to the sensitive nature of medical data and the complex relationships between stakeholders [25], [26]. These challenges arise from several aspects:

- *Feature-Specific Data Deletion:* In medical applications, there is often a need to delete specific features or attributes related to patient data, such as particular diagnoses or sensitive patient characteristics. This feature-level unlearning is more challenging because it requires removing only specific information without affecting the overall model performance. For example, deleting all data related to a certain medical condition must be done without introducing biases or disproportionately impacting model accuracy for other conditions.
- *Risk of Bias Introduction:* In medical AI, removing certain subsets of data (e.g., specific demographics or conditions) can lead to unintended biases in the model, compromising the fairness of the system [27]. This is particularly critical in healthcare, where bias can directly affect patient outcomes [28]. Thus, unlearning methods must be carefully designed to ensure that deleting data does not negatively impact the equity and reliability of the model.
- *Compliance with Regulatory Standards:* Medical AI systems must adhere to various regulatory standards that emphasize data privacy and security. Unlearning methods must not only be effective but also transparent and compliant with these regulations. This requires provable guarantees that the requested data has been successfully removed, without leaving any residual influence on the model.
- *High Stakes of Clinical Accuracy:* In medical settings, any reduction in model accuracy due to unlearning can have severe consequences for patient care. Therefore, medical AI requires unlearning methods that maintain a high level

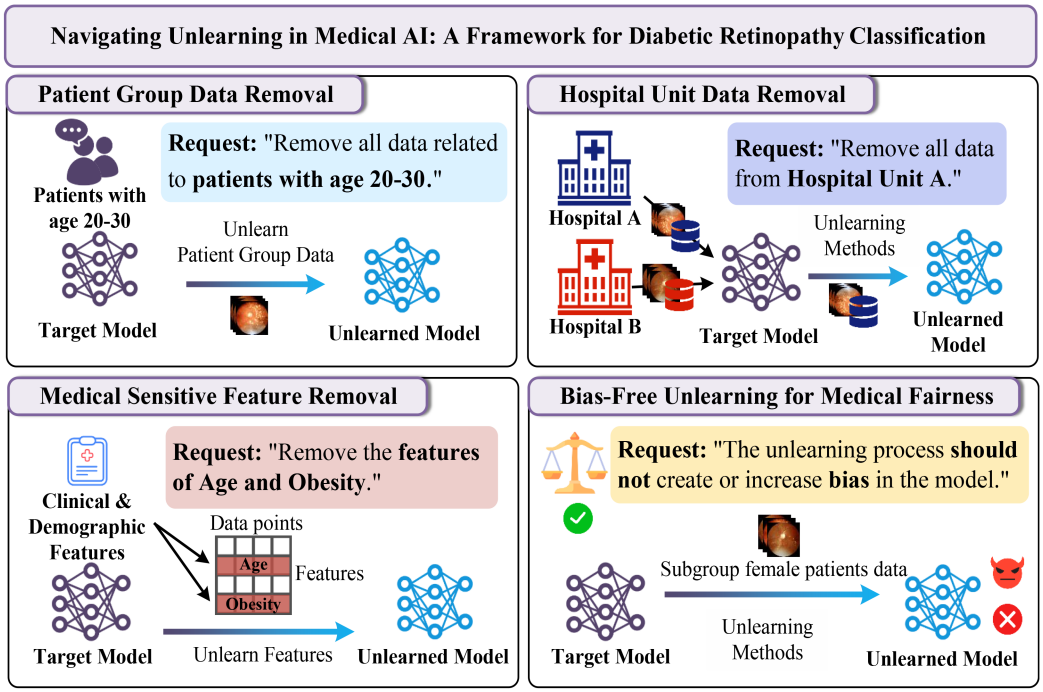


Fig. 2. Navigating Unlearning in Medical AI focus on four special goals, considering the practical tasks and sensitivity of data in medical scenarios. (1) A specific patient group requests the removal of their data from the target model to protect their privacy. (2) A hospital requests data to be unlearned from the target model as a unit to withdraw. (3) Remove sensitive features of medical data that have already been used to train the target model. (4) Consider the changes in the target model before and after unlearning from the perspective of model fairness.

of clinical accuracy, even after the deletion of specific data. Unlike in general applications, where minor accuracy trade-offs might be acceptable, medical AI models must ensure that their performance remains reliable to prevent misdiagnosis or suboptimal treatment recommendations.

These unique challenges highlight the importance of developing unlearning methods tailored to the healthcare domain, ensuring that data deletion processes are not only effective but also compliant, equitable, and capable of maintaining the high standards of clinical accuracy required for patient care.

V. Proposed Framework for Medical AI Unlearning

To address the unique challenges in medical AI, we propose a tailored unlearning evaluation framework that extends general unlearning principles to meet specific requirements of the medical field. As shown in Fig. 2, our proposed framework builds on the general goals of unlearning, adapting, and expanding them to accommodate the sensitive and high-stakes nature of healthcare. The framework encompasses a combination of general unlearning objectives and specific goals designed to cater to medical AI, ensuring compliance with stringent data privacy regulations, enhancing fairness, and maintaining high clinical accuracy.

A. Specific Goals of Unlearning in Medical AI

Medical AI unlearning must cater to certain unique requirements arising from the complexity and sensitivity of healthcare data. We identify the following specific medical goals (MGs) tailored to the medical domain:

- *Patient Group Data Removal Effectiveness (MG1):* The unlearning framework must support the removal of specific patient groups, such as patients from a particular age group, ethnicity, gender, or those with a specific health condition. This type of data removal is essential for respecting privacy requests and complying with ethical standards when individuals or entire groups refuse to provide data usage rights. Additionally, it helps mitigate biases that arise from the over-representation of certain groups, promoting fairness and balance across different populations.
- *Hospital Unit Data Removal Effectiveness (MG2):* The ability to remove data from a specific hospital or healthcare provider is crucial, particularly in scenarios where a participating hospital withdraws consent or changes its data-sharing policy. By removing data at the unit level, our framework ensures that biases or specific practices unique to a given institution do not disproportionately influence the model's predictions. This is particularly important for maintaining the model's generalizability across different healthcare environments without being overly influenced by localized factors.
- *Medical Sensitive Feature Removal Effectiveness (MG3):* The framework must also facilitate the removal of specific, sensitive features that may not be directly relevant to the model's predictive task but pose privacy risks or ethical concerns. Examples include biological age, socioeconomic status, or genetic information. Removing such features

supports the principle of data minimization, ensuring that only necessary information is used for decision-making. This is vital for preventing biased decisions based on irrelevant factors and maintaining fairness and transparency in medical AI systems.

- *Bias-Free Unlearning for Medical Fairness (MG4):* The unlearning process must ensure that removing certain groups or sensitive features does not create or increase bias in the model. It is crucial that unlearning does not negatively impact the representation or quality of care for specific patient demographics. Bias-free unlearning ensures that the model maintains or even improves fairness metrics, such as demographic parity or equal opportunity, thus promoting equitable healthcare outcomes for all groups, regardless of the modifications.

The proposed framework aims to address the general challenges of unlearning while also meeting the specific needs of the medical domain, ensuring that data removal processes are effective, ethical, and compliant with privacy regulations. By integrating both general and domain-specific goals, we provide a comprehensive approach to unlearning that is well-suited for the high-stakes and sensitive environment of healthcare.

B. Specific Unlearning Metrics for Medical AI

To evaluate the effectiveness of unlearning in medical AI, it is crucial to use appropriate metrics that capture the technical performance and the clinical relevance, fairness, and robustness of the model after unlearning. In this subsection, we introduce a set of specific metrics that are particularly applicable for assessing unlearning in the medical domain:

- *Sensitivity and Specificity:* After unlearning, it is essential to maintain high sensitivity and specificity to ensure that the model continues to identify conditions without over-predicting or under-predicting outcomes correctly. These metrics help verify that the unlearning process does not degrade the model's diagnostic accuracy. Sensitivity and specificity are critical metrics for evaluating a model's predictive performance, particularly in medical contexts where both false negatives and false positives carry significant consequences [29], i.e.,

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \tag{12}$$

$$\text{Specificity} = \frac{TN}{TN + FP}, \tag{13}$$

where TP is true positives, FN is false negatives, TN is true negatives, and FP is false positives.

- *F1-score:* F1-score, the harmonic mean of precision and recall, is a balanced metric that evaluates the model's performance, particularly in imbalanced data scenarios, i.e.,

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{14}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{15}$$

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN}. \tag{16}$$

- *Membership Inference Attack (MIA):* MIA is a technique used to determine whether a specific data point was part of the model's training set [30]. This method analyzes the model's behavior, such as its output logits, confidence levels, or loss values, on the target data to define whether it is in the training dataset, i.e.,

$$\text{MIA Accuracy} = \frac{TP_{\text{attack}} + TN_{\text{attack}}}{\text{Total Samples}}, \tag{17}$$

where TP_{attack} is correctly identified training samples as part of the training set, TN_{attack} is correctly identified samples not part of the training set, and Total Samples is the total number of samples in the attack set. As demonstrated in this work [31], the optimal defense against this type of MIA corresponds to a 50% attack accuracy, meaning the attacker's performance is no better than random guessing in determining whether a given example was part of the training set. In medical AI, this approach is particularly important for assessing whether sensitive patient data has been effectively "unlearned", thereby providing essential validation for privacy compliance.

- *Fairness Metrics:* Unlearning processes must not introduce or amplify biases against specific patient groups. After unlearning, it is critical to use these metrics to confirm that the quality of predictions remains equitable for all patient demographics, thus supporting the goal of fair and unbiased healthcare outcomes. We will use Statistical Parity Difference (SPD) [32] to measure fairness by evaluating whether the model's predictions are equally distributed across demographic groups, i.e.,

$$\text{SPD} = \left| P(\hat{Y} = 1|\text{Group A}) - P(\hat{Y} = 1|\text{Group B}) \right|, \tag{18}$$

where $P(\hat{Y} = 1|\text{Group A})$ is a proportion of positive predictions for Group A and $P(\hat{Y} = 1|\text{Group B})$ is a proportion of positive predictions for Group B. A lower SPD indicates that the model treats different demographic groups more equally, reducing the likelihood of biased predictions. Additionally, we provide Disparate Impact (DI) as a supplementary metric to offer a ratio-based perspective on fairness:

$$\text{DI} = \frac{P(\hat{Y} = 1|\text{Group A})}{P(\hat{Y} = 1|\text{Group B})}. \tag{19}$$

DI is particularly useful for understanding whether a group is underrepresented in positive outcomes. A DI value closer to 1 indicates more balanced treatment across groups. DI provides additional context and supports a more comprehensive assessment of fairness.

These metrics are used to evaluate whether the unlearning process in medical AI maintains model performance, clinical relevance, privacy guarantees, and fairness while effectively removing specified data. By employing this evaluation, we can comprehensively assess the impact of unlearning and provide

confidence in the reliability and safety of medical AI systems after machine unlearning.

VI. Experimental Results

This section presents the experimental evaluation of existing unlearning methods to determine whether their positive performance under general unlearning evaluation frameworks can translate effectively to the medical AI context. Our primary objective is to assess whether these methods meet the unique goals and requirements of unlearning in healthcare, as defined by our proposed framework.

A. Experimental Setup

The experimental setup includes details on the datasets, models, and unlearning techniques. Specifically:

Datasets: The experiments are conducted using the mBRSET dataset [16], which contains extensive medical information, clinical data, and demographic metadata. Due to its diversity and comprehensiveness, this dataset is well-suited to address several goals outlined in our Medical AI Unlearning Evaluation Framework. To systematically evaluate the unlearning methods, we designed four types of unlearning datasets, each corresponding to one of the four MGs. For each dataset type, six groups of data were selected, and the experimental results presented in the following sections represent the averaged performance across these groups.

- *Patient Group Data Removal:* The demographic metadata, including age, gender, and ethnicity, provides a basis for evaluating the removal of specific patient groups. This allows us to investigate how unlearning affects the model when certain patient demographics are excluded. We sampled three categories of unlearning datasets (UDs) from the training set to assess the performance of unlearning approaches when unlearning patient information from different subgroups. These categories included a female patient subgroup (positive: negative = 3:7), a subgroup of patients aged 20-30 years (positive: negative = 1:9), and a subgroup of patients with unhealthy habits (positive: negative = 9:1). For each category, we sampled six subsets following the specified distributions, with each subset containing 256 samples. A negative indicates that the sample is free of disease, while a positive indicates that there are four levels of disease. To simply represent the distribution of data we divide it into two types negative and positive.
- *Hospital Unit Data Removal:* The dataset does not inherently include attributes indicating data ownership by specific healthcare facilities or hospital units. However, we assume a data distribution across different hospital units and partition the dataset accordingly to simulate data from distinct hospitals. This approach allows us to evaluate how removing a particular unit's data influences the model's generalizability and fairness, especially in scenarios where a healthcare provider may revoke data usage permissions.
- *Medical Sensitive Feature Removal:* The dataset includes sensitive medical and demographic information, such as health histories and blood pressure readings. These features are ideal for evaluating the impact of removing sensitive data.
- *Bias-Free Unlearning for Medical Fairness:* The availability of comprehensive demographic information enables the evaluation of bias introduction or mitigation during unlearning. Removing data from specific groups (e.g., particular genders) allows us to monitor fairness metrics and ensure the model does not introduce or exacerbate biases during the unlearning process.

  The mBRSET dataset total included 5164 samples. We used 80% as the training dataset and 20% as the test dataset, which could also be called the unseen dataset. Except for each UD, the rest of the training dataset is denoted by the remaining dataset.

Models: In our experiments, we combine ResNet with additional fully connected layers to effectively utilize the diverse data available in the mBRSET dataset. Given its strong performance in image classification tasks, ResNet was used to extract features from retinal fundus images. In addition, we used fully connected layers to process clinical and demographic metadata, such as age, gender, and medical history. The extracted features from both the image data and structured data were then fused to form a comprehensive representation for diagnosis. This approach allowed us to leverage both visual and non-visual data to improve diagnostic accuracy, providing a more holistic assessment suitable for medical AI applications.

Unlearning Techniques: We evaluate five widely used unlearning methods that have demonstrated effectiveness in general machine learning contexts and represent diverse algorithmic paradigms identified in recent unlearning survey [12]: SISA [13], DeltaGrad [22], Selective Forgetting in Deep Networks: Eternal Sunshine of the Spotless Net (FIM) [23], Certified Data Removal from Machine Learning Models (CDR) [21], Machine Unlearning of Features and Labels (FU) [24]. This selection reflects key methodological dimensions, including exact and approximate unlearning, parameter-level and representation-level approaches. They provide a representative foundation for evaluating unlearning performance in medical AI systems.

B. MG1 Performance

In this experiment, we selected three representative patient subgroups for unlearning: female patients, patients aged 20–30, and patients with reported bad habits (such as smoking or alcohol use). These subgroups are chosen to reflect clinically relevant demographic and lifestyle characteristics that often correlate with differential health outcomes and prediction patterns in medical models. Table II highlights the F1-score and MIA success probability changes for five unlearning methods after unlearning 256 samples from the female patient subgroup (positive: negative = 3:7) in the model. Among these methods, retraining demonstrates superior unlearning performance with a significant reduction in the unlearning dataset's F1-score with no performance degradation in the remaining dataset and only a

TABLE II
PERFORMANCE OF FIVE MACHINE UNLEARNING METHODS AFTER UNLEARNING THE FEMALE PATIENT SUBGROUP (POSITIVE: NEGATIVE = 3:7). ↑ = IMPROVED; ↓ = DEGRADED.

	Unlearning Dataset	Remaining Dataset	Test Dataset	MIA
Retrain	↓ 14%	↓ 0.0%	↓1.9%	↓25%
SISA	↓ 11%	↓ 2.2%	↓2.5%	↓24%
FIM	↓ 35%	↓ 26%	↓28%	↓12%
DeltaGrad	↓ 63%	↓ 58%	↓60%	↑5.8%
CDR	↓ 43%	↓ 39%	↓40%	↓15%

minor impact on the test dataset. Most notably, retraining reduces the MIA success rate by 25%, bringing it down to approximately 50%. This indicates that after retraining, distinguishing training from non-training samples in the unlearning subset becomes close to chance, thus achieving the desired privacy preservation. In comparison, the SISA method exhibits slightly weaker unlearning performance but introduces slightly more degradation in the remaining dataset and the test dataset. This trade-off arises from SISA's design, which involves splintering and re-aggregating data during the unlearning process. While this mechanism mimics retraining and preserves the model's structure, it inherently limits the method's ability to fully capture the global data distribution, leading to marginally reduced adaptability to the overall dataset. Despite these limitations, SISA achieves a comparable reduction in the MIA success rate (24%), demonstrating strong privacy-preserving capabilities, close to the performance of retraining.

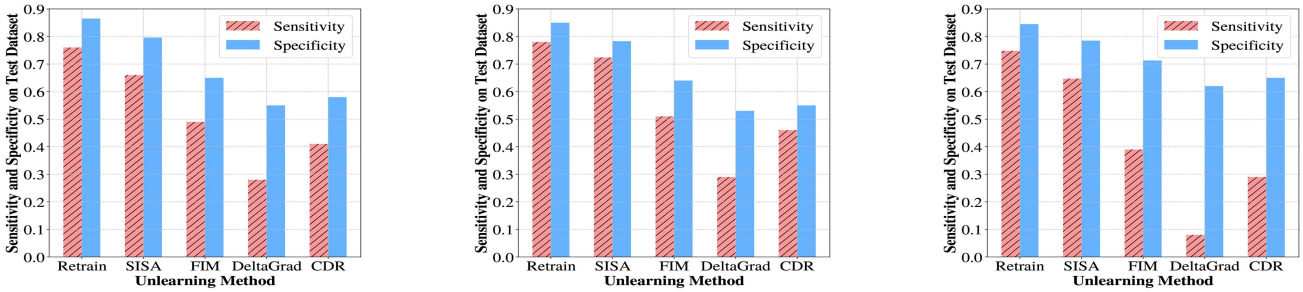
Conversely, FIM achieves moderate unlearning performance with a reduction of 35% on the unlearning dataset. However, it sacrifices substantial performance in the remaining dataset and the test dataset. Furthermore, FIM only reduces the MIA success rate by 12%, leaving it at 61%, which is far from the expectation of 50%. This indicates that FIM's unlearning approach leaves residual traces of the unlearned data, compromising its privacy guarantees. Similarly, CDR achieves better performance in unlearning effectiveness. However, it reduces the MIA success rate by 15%, leaving it at 59.18%, far from the optimal 50% achieved by Retrain and SISA. This reflects CDR's limited ability to ensure privacy from unlearning target datasets. From the perspective of the elevated MIA success rate, it appears inconsistent with the proportion of performance decline observed for FIM and CDR on the unlearning dataset. This discrepancy arises because FIM and CDR, during the unlearning process, adversely impact the overall model performance, resulting in a broader degradation across the model's capabilities.

DeltaGrad shows severe degradation on the unlearning subset, seemingly indicating aggressive unlearning. However, this comes at the cost of significant degradation in the remaining dataset and the test dataset, severely undermining the overall model utility. Additionally, its reduction in MIA is marginal and can even reverse (higher MIA). In fact, DeltaGrad increases MIA by 5.8% (to 73.48%). The reason is that under large, distribution-shifting deletions, the local approximation used by DeltaGrad becomes inaccurate; outdated curvature preconditioning amplifies errors, pushing the decision boundary in the wrong direction, which explains the severe utility drop and even higher MIA we observe.

Importantly, most unlearning methods rely on assumptions that make analysis tractable, e.g., small deletion ratios or specific training pipelines. In medical settings, multimodal feature entanglement, subgroup heterogeneity, and severe class imbalance amplify deletion-induced distribution shift. Under these conditions, local approximation or shard-based strategies struggle to jointly satisfy forgetting effectiveness, global utility, fairness, and privacy. Thus, the core challenge is the mismatch between medical data properties and the methods' original operating regime.

From the results shown in Fig. 3, the variation in the proportion of positive samples in the unlearning dataset significantly impacts the model's Sensitivity and Specificity, following a consistent pattern. When the unlearning dataset contains a high proportion of positive samples (e.g., Fig. 3(c), positive: negative = 9:1), the effect on Sensitivity is particularly pronounced. This is because positive samples are already underrepresented in the training dataset, and removing a large number of positive samples further decreases their weight in the training set, severely impairing the model's ability to detect positive cases. This phenomenon is especially evident in the DeltaGrad, FIM, and CDR methods, where Sensitivity declines sharply, indicating that these methods struggle to maintain the model's positive-class recognition capability when a high proportion of positive samples are unlearned. In particular, DeltaGrad suffers from a drastic sensitivity drop in Setting 3 (Fig. 3(c)). This can be attributed to its reliance on cached gradient and Hessian information computed on the original training distribution. When a large portion of positive samples is removed, this second-order information becomes outdated and misaligned with the updated data distribution. The resulting parameter update may inadvertently distort the decision boundary, disproportionately affecting the already underrepresented positive class, and leading to a spike in false negatives. In contrast, Retrain recomputes gradients based on the remaining data, which allows it to adaptively recalibrate the decision boundary and preserve high sensitivity.

Conversely, when the unlearning dataset contains a low proportion of positive samples (e.g., Fig. 3(b) and Fig. 3(a)), the impact on Sensitivity is much smaller. This is because negative samples are dominant in the training dataset, and the sheer magnitude of negative samples ensures that removing a subset of them has a diluted effect on the overall model performance. This effect is particularly pronounced in SISA and Retrain. SISA splintering and re-aggregation mechanism only updates the shards containing the target data, leaving the other shards unaffected. Consequently, the removal of negative samples primarily results in local adjustments, with minimal impact on the model's global ability to classify negative cases, thereby maintaining high Specificity. In other words, the dominance of negative samples in the training dataset acts as a buffer, mitigating the impact of unlearning on the model's performance. For Retrain, due to a large number of negative samples in the



(a) Female subgroup (b) Age 20–30 subgroup (c) Bad habits subgroup

Fig. 3. Performance of different unlearning methods after unlearning different groups of people.

TABLE III
MIA SUCCESS PROBABILITY AFTER UNLEARNING HOSPITAL UNIT DATA WITH DIFFERENT UNLEARNING METHODS. THIS TABLE COMPARES MIA SUCCESS RATES (%) ACROSS METHODS, WHERE VALUES CLOSER TO 50% INDICATE STRONGER UNLEARNING PERFORMANCE.

	128 samples	256 samples	512 samples
Original	69.13%	69.75%	69.89%
Retrain	51.12%	52.51%	52.78%
SISA	51.25%	52.59%	52.83%
FIM	60.38%	61.08%	61.53%
DeltaGrad	73.19%	73.45%	73.87%
CDR	58.95%	59.15%	59.58%

training dataset, only 10% of the negative samples were deleted, which would not have too much impact on the prediction of the model.

This dominance of negative samples serves as a "stabilizing factor" during the unlearning process, particularly when the unlearning mechanism causes limited disruption to the global data distribution. This explains why SISA exhibits minimal performance degradation when unlearning datasets are dominated by negative samples, while datasets with a high proportion of positive samples cause much greater disruption to the model's performance. This analysis further underscores the complexity of achieving effective unlearning in medical contexts. As illustrated in Fig. 3, the groups requiring unlearning exhibit highly diverse distributions, which demand unlearning methods to adapt to various requirements effectively. Balancing unlearning effectiveness with maintaining model performance is a critical challenge, necessitating unlearning approaches with greater flexibility and robustness to address the unique demands of medical applications.

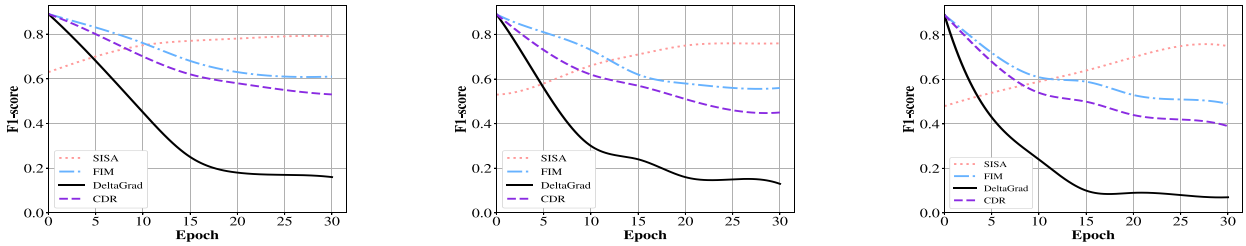
C. MG2 Performance

From Table III, the MIA success probability illustrates the effectiveness of each method in unlearning hospital unit data at varying scales (128, 256, and 512 samples). Retrain and SISA methods demonstrate comparable MIA success rates as the number of unlearning samples increases, maintaining great performance on unlearning. The FIM and CDR methods, however, show increasing MIA probabilities as the deletion size grows, reflecting difficulties in achieving thorough unlearning due to the partial parameter adjustment nature of these methods. The DeltaGrad method demonstrates relatively high MIA values, which increase slightly as the number of unlearning samples grows. This increases because DeltaGrad's overly aggressive unlearning performance introduces significant distortions in the model, causing the unlearned samples to deviate from the original data distribution. Consequently, DeltaGrad not only fails to effectively erase the influence of the unlearned data but also amplifies the risk of membership information leakage, undermining both privacy preservation and overall model robustness.

The performance of different unlearning methods after unlearning hospital unit data is illustrated in Figs. 4 and 5. These experiments simulate variations in data contributions from different hospital units by deleting 128, 256, and 512 samples, respectively. The data distribution of each hospital unit is assumed to be consistent with the overall training dataset (with a positive-to-negative sample ratio of 3:7). The experiments simulate varying contributions from hospital units to analyze the adaptability and limitations of different unlearning methods in handling datasets of different sizes.

From the unlearning dataset performance in Figs. 4 and 5, it is evident that increasing the number of deleted samples significantly impacts the unlearning effectiveness of each method. The F1-score of the SISA method on the unlearning dataset exhibits a gradual recovery trend. Particularly, when 128 and 256 samples are deleted, SISA adapts relatively quickly to the new data distribution through shard retraining, showing a faster recovery and eventually reaching an F1 score close to the initial level. However, when the number of deleted samples increases to 512, the recovery slows considerably, and the initial performance is not fully restored. This indicates that SISA demonstrates strong global adaptability in small-scale data deletion tasks but faces limitations in large-scale deletion scenarios, where the shard aggregation mechanism struggles to compensate for global feature loss. Additionally, the shard mechanism may exacerbate distributional imbalances between shards in large-scale deletions, further limiting the model's adaptability.

Retrain requires significantly more epochs to achieve convergence, making it impractical to include alongside other methods



(a) F1-score on the unlearning dataset with 128 samples after unlearning

(b) F1-score on the unlearning dataset with 256 samples after unlearning

(c) F1-score on the unlearning dataset with 512 samples after unlearning

Fig. 4. Performance of different unlearning methods after unlearning different hospital unit data.

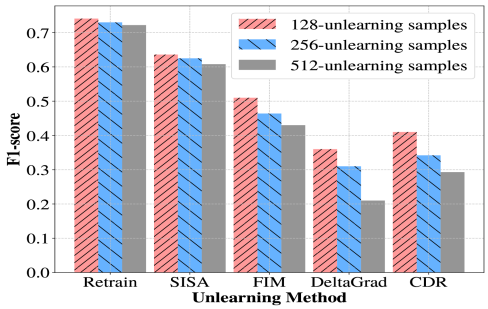


Fig. 5. F1-score on test dataset after unlearning different hospital unit data.

in the same plot. Retrain involves removing the target unlearning data from the training dataset and retraining the model until it converges. SISA, on the other hand, achieves unlearning by removing the target data from its corresponding slices and aggregating the remaining slices. As a result, the performance on the unlearning dataset is similar to that on the test dataset. Similarly, in Retrain, since the proportion of removed data relative to the total dataset is not particularly large, the overall model performance does not degrade significantly. The predictions on the unlearning dataset remain comparable to those on the test dataset. However, as the volume of unlearning data increases, the reduction in the overall training dataset size leads to a moderate decline in model performance. This indicates that while both methods maintain consistency between the unlearning and test datasets, the scaling of unlearning tasks impacts the general performance of the models.

In contrast, as shown in Fig. 4, FIM exhibits limited forgetting on the unlearning subsets, and its performance deteriorates as the deletion size increases. Even with 128 deleted samples, residual influence remains. This is consistent with the method's local parameter regularization: Fisher-based approximations estimated on the pre-deletion distribution become inaccurate once subgroup removal induces a distribution shift. In high-dimensional, multimodal settings, approximation and estimation errors are amplified and scale with the deletion size (e.g., 512 samples), yielding weaker target suppression. Consistent with Fig. 5, FIM also struggles to preserve global utility, with marked test-set F1 drops for larger deletions. Together, these results indicate that, in our setting, FIM's local adjustments fail to track the new optimum and inadvertently perturb parameters supporting the remaining data, thereby undermining the model's overall predictive ability.

The DeltaGrad method exhibits the most aggressive unlearning performance, with its F1-score on the unlearning dataset rapidly declining in each deletion task. This reflects the method's strong capability for removing target data through gradient adjustment. However, this aggressive unlearning comes at the cost of substantial instability, particularly when 256 and 512 samples are deleted, as its F1-score continues to decline, nearing a collapsed state. This phenomenon indicates that DeltaGrad lacks convergence in large-scale deletion tasks, with its excessive interference with global model parameters not only erasing target data but also disrupting the model's adaptability to the remaining training data. Fig. 5 further corroborates this: DeltaGrad consistently achieves the lowest F1-score on the test dataset, highlighting its failure to preserve global model performance.

The CDR method's performance on the unlearning dataset falls between that of SISA and FIM. Its F1-score demonstrates a relatively stable downward trend, with acceptable unlearning effectiveness in the deletion of 128 samples. However, as the deletion scale increases, model performance degrades more severely. A plausible cause in deep, non-convex settings is that practical CDR implementations rely on curvature approximations, e.g., approximate Hessians, and often apply removals in batches; as the number of deletions increases, approximation errors compound and destabilize the updates, leading to noticeable F1 degradation. From the test dataset performance in Fig. 5, the CDR method shows slightly better global performance maintenance than FIM, but its F1-score also declines significantly with the increase in deleted samples, suggesting that its parameter optimization strategy remains limited in its applicability to large-scale tasks.

D. MG3 Performance

We conducted experiments using two unlearning methods, Retrain and FU, to evaluate the impact of removing specific features on the performance of the trained model. It is worth noting that other unlearning methods, including SISA, FIM, CDR, and DeltaGrad, are excluded from this analysis as they are primarily designed for sample-level unlearning. These methods operate on entire data points and lack mechanisms to remove individual input features. Adapting them for feature-level unlearning would require significant algorithmic changes, which are beyond the

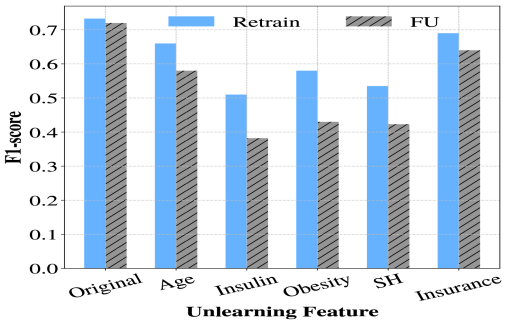


Fig. 6. F1-score on test dataset after unlearning different features. The systemic hypertension is denoted by SH.

scope of this study. Therefore, we focus on Retrain and FU, which naturally support manipulation at the feature level and are suitable for this evaluation. Consequently, their comparative performance in the context of feature unlearning was not evaluated in this subsection.

The dataset originally contained 21 features, from which we selected five features for this unlearning evaluation. These include three features, namely "Insulin", "Obesity", and "Systemic Hypertension" (SH), which were predicted to have a significant impact on the model's decision-making, and two features, "Age" and "Insurance", which were expected to have a relatively smaller influence on the model's predictions. Specifically, "Insulin" indicates whether the sample involves insulin use; "Obesity" reflects whether the sample self-reports obesity; "Systemic Hypertension" denotes whether the sample self-reports a hypertension condition; "Age" represents the sample's age, and "Insurance" captures the sample's self-reported insurance status. The goal is to compare the effect of unlearning these features on the model's overall performance. The experimental results are shown in Fig. 6.

The results of the Retrain method meet our expectations. The unlearning of the features, "Age" and "Insurance", had minimal effects on the model's F1-score. The relatively higher robustness of Retrain when unlearning less critical features highlights its capability to maintain the model's global performance while targeting specific attributes. In contrast, the removal of key predictive features of the model, such as "Insulin", "Obesity", and "Systemic Hypertension", led to a noticeable decrease in the model's performance. This demonstrates that the Retrain method effectively isolates the influence of specific features.

In comparison, the FU method exhibited a much larger performance degradation across all feature removal tasks, irrespective of the feature's predicted importance. This outcome can be attributed to FU's reliance on the influence function to update model parameters. Given the interdependency among features in complex datasets, removing even a single feature may inadvertently affect parameters associated with other related features. As a result, the FU method struggles to maintain the model's overall performance after feature removal.

E. MG4 Performance

Table IV shows the impact of five different machine unlearning methods on the SPD and DI of the model after removing a subgroup of female patients. The dataset removed contained 512 samples with a positive-to-negative ratio of 3:7. SPD is used to measure the bias between different groups within the model, here is the female group and male group, with values closer to 0 indicating greater fairness across groups. DI offers a ratio-based view, where values closer to 1 suggest balanced treatment across groups.

TABLE IV
SPD AND DI OF SIX MACHINE UNLEARNING METHODS AFTER UNLEARNING THE FEMALE PATIENT SUBGROUP. "ORIGINAL" INDICATES THE FAIRNESS METRICS BEFORE UNLEARNING.

	Original	Retrain	FIM	CDR	DeltaGrad	SISA
SPD	0.0413	0.1136	0.0754	0.0782	0.0819	0.0914
DI	1.08	1.23	1.15	1.16	1.16	1.18

After the removal of the female patient subgroup, the SPD for the fully retrained model increased to 0.1136, and the DI rose to 1.23, both showing increased gender disparity compared to the original model. The primary reason for this increase is that the removal of a large-scale female subgroup with a relatively high positive sample ratio caused a shift in the distribution of the training data, thereby increasing the bias between gender groups.

In contrast, the FIM and CDR methods exhibited better performance than full retraining in terms of both SPD and DI. These methods partially adjust the model parameters to weaken the influence of the target data, but they do not eliminate the impact of the removed data. Instead, they apply gradual modifications to the model. However, these gradual adjustments not only weaken the influence of the unlearned data but also affect the performance of the remaining data in the model. It subsequently reduces the model's overall performance. This indicates that FIM and CDR did not achieve complete unlearning but rather weakened the target data's influence, allowing the model to retain part of the memory of the original features. As a result, these methods achieved relatively lower SPD values and DI values closer to 1. The SPD and DI values for SISA fell between full retraining and FIM/CDR. SISA employs a slicing and aggregation strategy, where data is partitioned and trained on separate sub-models, followed by aggregation to achieve partial data unlearning. When retraining the sub-models containing the unlearned data, biases similar to those introduced by full retraining are incorporated. However, because only a subset of the models is retrained while the remaining sub-models remain unaffected, SISA is able to retain some of the original features. Consequently, the SPD and DI for SISA are higher than those of FIM and CDR but remain lower than those of full retraining.

In summary, after removing a specific female patient subgroup, the SPD and DI values of all methods increased compared to the original model, indicating an increase in the model's bias between gender groups. Full retraining is the most direct approach, but also results in the largest increase in bias, as it completely loses the ability to learn from the removed data. FIM and CDR mitigate the impact of unlearning on model performance by preserving some features, leading to smaller increases in SPD and more balanced DI values; however, this

TABLE V
PERFORMANCE OF SIX MACHINE UNLEARNING METHODS BASED ON MEDICAL AI UNLEARNING EVALUATION FRAMEWORK. ✓= CRITERION SATISFIED; ↓ = DEGRADED; - = NOT APPLICABLE.

	MG1	MG2	MG3	MG4
Retrain	✓	✓	✓	↓
SISA	✓	✓	-	↓
DeltaGrad	↓	↓	-	↓
FIM	↓	↓	-	↓
CDR	↓	↓	-	↓
FU	-	-	↓	-

comes at the cost of incomplete unlearning. SISA, on the other hand, strikes a balance between retaining model performance and achieving complete unlearning. Although its fairness metrics still deteriorated, it outperformed full retraining in terms of bias control.In sum, practical unlearning must preserve utility and fairness while remaining stealthy: post-unlearning fairness drift can itself leak which subgroup was deleted, undermining the plausible deniability of the deletion.

F. Discussion

The evaluation summarized in Table V reveals notable differences in the capabilities of various unlearning methods within the context of medical AI. Retrain, regarded as the gold standard, demonstrates strong performance across critical metrics such as maintaining fairness, accuracy, and model generalizability, although unlearning certain groups may introduce bias. However, it suffers from a significant drawback: its computational and time demands make it impractical for large-scale or frequent unlearning tasks, particularly in scenarios involving deep learning models with high-dimensional medical data. Methods like FIM and CDR exhibit weaker unlearning and global performance maintenance, reflecting their insufficient ability to adapt to the complex distributions inherent in medical datasets. Meanwhile, DeltaGrad's overly aggressive unlearning approach leads to severe instability, making it unsuitable for sensitive medical applications. The failure of parameter-based unlearning methods such as DeltaGrad, FIM, and CDR arises from their assumption that data samples have isolated, localized effects. In multi-modal medical models, however, deeply fused visual and structured features lead to gradient entanglement and feature attribution coupling, where forgetting one sample causes non-local disruptions. This mismatch between method assumptions and the entangled representation structure explains their poor performance in this setting. SISA, on the other hand, emerges as a promising alternative, showcasing competitive performance compared to Retrain across multiple metrics while requiring less computational overhead. Its shard-based aggregation and re-training mechanism effectively balances unlearning effectiveness and global performance maintenance. Nevertheless, the shard mechanism introduces limitations when dealing with large-scale deletions, as it can exacerbate distributional imbalances among shards. While this study highlights SISA's strengths, these limitations were not fully explored or addressed, underscoring the need for future research to refine unlearning methods that are not only efficient but also capable of accommodating the unique and stringent requirements of medical AI.

VII. LIMITATIONS AND FUTURE WORK

While the proposed evaluation framework offers a structured methodology for assessing machine unlearning in medical AI, several limitations remain that require further exploration.

First, the current framework was developed and validated based on the specific task of diabetic retinopathy classification using the mBRSET dataset. Although its design principles aim for generalizability, the framework has not yet been tested on other types of medical AI applications, such as electrocardiogram interpretation, pathology image analysis, or models that integrate multiple data modalities. Future work should investigate its applicability in these broader clinical scenarios to validate its general use. Second, our experiments focused on six representative unlearning methods. However, not all of these methods are compatible with every evaluation dimension included in our framework. For instance, certain methods cannot be readily applied to feature-level unlearning. Exploring adaptations or extensions of these methods to improve compatibility with a wider range of evaluation objectives would enhance the comprehensiveness of the framework. Finally, this study primarily serves as an evaluation benchmark and does not introduce novel unlearning algorithms. Future research could focus on developing unlearning methods specifically optimized for medical contexts, with an emphasis on privacy preservation, fairness assurance, and clinical safety. Such efforts should also consider practical constraints found in real-world healthcare systems, including continuously evolving datasets and heterogeneous institutional environments.

VIII. CONCLUSION

In this paper, we evaluated multiple machine unlearning methods in the context of medical AI using the mBRSET dataset, a comprehensive medical dataset encompassing diverse features and demographics. By designing task-specific datasets aligned with the goals of our Medical AI Unlearning Evaluation Framework, we systematically assessed the ability of each method to achieve effective unlearning while maintaining overall model performance. Our results demonstrate that traditional unlearning methods remain significant limitations when these methods are applied to medical AI scenarios. Retrain, though providing reliable unlearning, is computationally expensive and time-intensive, making it impractical for real-world applications involving large-scale models or dynamic data updates. SISA, while effective in many cases, struggles with unlearning discrete samples and can exhibit performance degradation in scenarios involving highly imbalanced data or complex relationships between features. Additionally, methods such as DeltaGrad, FIM, CDR, and FU, which focus on parameter updates, demonstrated inconsistent performance and failed to generalize effectively to the unique challenges posed by medical data, such as fairness, sensitivity, and dependencies among specific features. This paper highlights the critical need for more robust and efficient

unlearning methods tailored to the stringent requirements of medical AI. The findings in this study underscore the importance of developing specialized unlearning solutions that can simultaneously ensure privacy, fairness, and clinical reliability in medical AI systems.

REFERENCES

[1] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in health and medicine," *Nature Med.*, vol. 28, no. 1, pp. 31–38, 2022.

[2] C. Bhatt, I. Kumar, V. Vijayakumar, K. U. Singh, and A. Kumar, "The state of the art of deep learning models in medical science and their challenges," *Multimedia Syst.*, vol. 27, no. 4, pp. 599–613, 2021.

[3] T. Dhar, N. Dey, S. Borra, and R. S. Sherratt, "Challenges of deep learning in medical image analysis—improving explainability and trust," *IEEE Trans. Technol. Soc.*, vol. 4, no. 1, pp. 68–75, Mar. 2023.

[4] A. L. Beam, J. M. Drazen, I. S. Kohane, T.-Y. Leong, A. K. Manrai, and E. J. Rubin, "Artificial intelligence in medicine," *New England J. Med.*, pp. 1220–1221, 2023.

[5] A. Malhotra, E. J. Molloy, C. F. Bearer, and S. B. Mulkey, "Emerging role of artificial intelligence, Big Data analysis and precision medicine in pediatrics," *Pediatr. Res.*, vol. 93, no. 2, pp. 281–283, 2023.

[6] N. S. Gupta and P. Kumar, "Perspective of artificial intelligence in healthcare data management: A journey towards precision medicine," *Comput. Biol. Med.*, vol. 162, 2023, Art. no. 107051.

[7] L. Tang, J. Li, and S. Fantus, "Medical artificial intelligence ethics: A systematic review of empirical studies," *Digit. Health*, vol. 9, 2023, Art. no. 20552076231186064.

[8] J. Zhang and Z.-m. Zhang, "Ethics and governance of trustworthy medical artificial intelligence," *BMC Med. Inform. Decis. Mak.*, vol. 23, no. 1, 2023, Art. no. 7.

[9] A. Ashraf, S. Khan, N. Bhagwat, M. Chakravarty, and B. Taati, "Learning to unlearn: Building immunity to dataset bias in medical imaging studies," in *Proc. NeurIPS Workshop Mach. Learn. Health (ML4H)*, 2018.

[10] General Data Protection Regulation, "General data protection regulation (gdpr)," Intersoft Consulting, 2018.

[11] H. M. Edemekong and P. Annamaraju, "Health insurance portability and accountability act," Accessed: Jan. 15, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK500019/#:~:text=HHS%20initiated%

[12] J. Xu, Z. Wu, C. Wang, and X. Jia, "Machine unlearning: Solutions and challenges," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 8, no. 3, pp. 2150–2168, Jun. 2024.

[13] L. Bourtoule et al., "Machine unlearning," in *Proc. 2021 IEEE Symp. Secur. Privacy*, 2021, pp. 141–159.

[14] T. Shaik, X. Tao, H. Xie, L. Li, X. Zhu, and Q. Li, "Exploring the landscape of machine unlearning: A comprehensive survey and taxonomy," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 7, pp. 11676–11696, Jul. 2025.

[15] S. K. Sakib and M. Xie, "Machine unlearning in digital healthcare: Addressing technical and ethical challenges," in *Proc. AAAI Symp. Ser.*, 2024, vol. 4, no. 1, pp. 319–322.

[16] C. Wu et al., "A portable retina fundus photos dataset for clinical, demographic, and diabetic retinopathy prediction," *Sci. Data*, vol. 12, 2025, Art. no. 323, doi: 10.1038/s41597-025-04627-3.

[17] J. Vatter, R. Mayer, and H.-A. Jacobsen, "The evolution of distributed systems for graph neural networks and their origin in graph processing and deep learning: A survey," *ACM Comput. Surv.*, vol. 56, no. 1, pp. 1–37, 2023.

[18] J. Brophy and D. Lowd, "Machine unlearning for random forests," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 1092–1104.

[19] A. Ginart, M. Guan, G. Valiant, and J. Y. Zou, "Making AI forget you: Data deletion in machine learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 3518–3531.

[20] L. Graves, V. Nagisetty, and V. Ganesh, "Amnesiac machine learning," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 13, pp. 11516–11524.

[21] C. Guo, T. Goldstein, A. Hannun, and L. Van Der Maaten, "Certified data removal from machine learning models," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2020, pp. 3832–3842.

[22] Y. Wu, E. Dobriban, and S. Davidson, "DeltaGrad: Rapid retraining of machine learning models," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 10355–10366.

[23] A. Golatkar, A. Achille, and S. Soatto, "Eternal sunshine of the spotless net: Selective forgetting in deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9304–9312.

[24] A. Warnecke, L. Pirch, C. Wressnegger, and K. Rieck, "Machine unlearning of features and labels," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2023, doi: 10.14722/ndss.2023.23087.

[25] R. Wang, Y. F. Li, X. Wang, H. Tang, and X. Zhou, "Learning your identity and disease from research papers: Information leaks in genome wide association study," in *Proc. 16th ACM Conf. Comput. Commun. Secur.*, 2009, pp. 534–544.

[26] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, "Privacy in pharmacogenetics: An {End-to-End} case study of personalized warfarin dosing," in *Proc. 23 rd USENIX Secur. Symp.*, 2014, pp. 17–32.

[27] Y. Yang, H. Zhang, J. W. Gichoya, D. Katabi, and M. Ghassemi, "The limits of fair medical imaging AI in real-world generalization," *Nature Med.*, vol. 30, pp. 2838–2848, 2024.

[28] S. Tayebi Arasteh et al., "Preserving fairness and diagnostic accuracy in private large-scale AI models for medical imaging," *Commun. Med.*, vol. 4, no. 1, 2024, Art. no. 46.

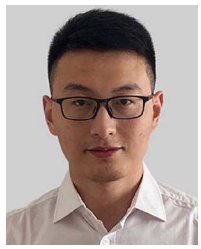
[29] C. McGenity et al., "Artificial intelligence in digital pathology: A systematic review and meta-analysis of diagnostic test accuracy," *npj Digit. Med.*, vol. 7, no. 1, 2024, Art. no. 114.

[30] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, "Membership inference attacks from first principles," in *Proc. 2022 IEEE Symp. Secur. Privacy*, 2022, pp. 1897–1914.

[31] M. Kurmanji, P. Triantafillou, J. Hayes, and E. Triantafillou, "Towards unbounded machine unlearning," in *Proc. 37th Int. Conf. Neural Inf. Process. Syst.*, 2024, pp. 1957–1987.

[32] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. Cambridge, MA, USA: MIT Press, 2023.

Xinghao Li received the B.E. degree from Southwest University, China, and the bachelor's and bachelor's (Hons.) degrees in information technology from Deakin University, Australia, respectively. He is currently working toward the Ph.D. degree with the Faculty of Information Technology, Monash University, Clayton, VIC, Australia. His research interests include machine unlearning, medical AI, and edge computing.



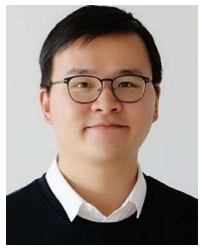
Chi Liu (Member, IEEE) received the B.Sc. degree from Wuhan University, Wuhan, China, the M.Sc. degree from the China Ship Research and Development Academy, China, and the Ph.D. degree from the School of Computer Science, University of Technology Sydney, Australia. He was a Research Fellow with the Faculty of Information Technology, Monash University, Clayton, VIC, Australia. He is currently an Assistant Professor with the Faculty of Data Science, City University of Macau, Macao, SAR, China. His research interests include multimedia privacy and security, trustworthy and responsible AI, and AI security and safety.

Youyang Qu (Senior Member, IEEE) received the B.S. degree in mechanical automation and the M.S. degree in software engineering from the Beijing Institute of Technology, Beijing, China, in 2012 and 2015, respectively, and the Ph.D, degree from the School of Information Technology, Deakin University, Australia, in 2019. He was a Research Fellow with Deakin University, Australia. He is currently a Research Scientist with Data61, Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia. His research interests include machine learning, Big Data, IoT, blockchain and corresponding security, and customizable privacy issues. He has more than 50 publications, including high-quality journals and conferences papers such as IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING, *ACM Computing Surveys*, and IEEE INTERNET OF THINGS JOURNAL. He is also active in the research society and has served as an organizing committee member in SPDE 2020, BigSecuirty 2021, Tridentcom 2021/2022.

Shujie Cui received the B.Sc. degree from Shandong University, Jinan, China, in 2011, the M.Sc. degrees in computer science from Shandong University, Jinan, China, and also from the University of Luxembourg, Luxembourg, in 2013, and the Ph.D. degree in computer science from the University of Auckland, Auckland, New Zealand. She was a Research Associate With the Department of Computer Science, City University of Hong Kong, Hong Kong. She is currently a Lecturer with the Faculty of Information Technology, Monash University, Clayton, VIC, Australia. Her research interests include cloud computing, applied cryptography, and security and privacy.

Cunjian Chen (Senior Member, IEEE) received the Ph.D. degree in computer science from West Virginia University. He was a Senior Research Associate with Michigan State University, East Lansing, MI, USA. He is currently an Adjunct Lecturer with Monash University, Clayton, VIC, Australia, and a Research Fellow with the Monash Suzhou Research Institute. He is currently an Associate Editor for *Neural Processing Letters* and was also an Associate Editor for *IET Image Processing*. Dr. Chen has contributed to the academic community in various roles, including Tutorial Chair for IJCB, Area Chair for ICME, ICIP, ICASSP, and FG, and Session Chair for ICASSP, ICME, and FG. He was the recipient of the Outstanding Area Chairs Award at ICME 2021.

Xingliang Yuan (Senior Member, IEEE) is currently an Associate Professor with the School of Computing and Information Systems, The University of Melbourne, Parkville, VIC, Australia. From 2017 to 2024, he was a Faculty Member with Monash University, Clayton, VIC, Australia. His research interests include designing systems and protocols to address real-world privacy and security challenges, and focuses on data security and privacy, secure networked systems, and trustworthy machine learning. His research has been supported by the Australian Research Council, CSIRO, the Australian Department of Home Affairs, Australian Department of Health and Aged Care, and the Oceania Cyber Security Centre. His work has been published in major venues of computer security and systems, such as CCS, S&P, USENIX Security, NDSS, IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, and IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY. He was the sole recipient of the Dean' s Award for Excellence in Research by an Early Career Researcher (2020) at Monash, and also the co-recipient of the Best Paper Award at the European Symposium on Research in Computer Security (ESORICS) 2021. He is also on the Editorial Board of IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING (TDSC) and IEEE TRANSACTIONS ON SERVICES COMPUTING (TSC). He was a Track Co-Chair of ICDCS' 24, WISE' 24, MSN' 24, and Program Co-Chair of Lamps@CCS' 24, SecTL@AsiaCCS' 23, and NSS' 22.



Zongyuan Ge (Senior Member, IEEE) received the Ph.D. degree in engineering from the Queensland University of Technology, Brisbane, Australia, in 2017. From 2016 to 2018, he was an Engineer with IBM Research Australia. From 2018 to 2021, he was a Senior Research Fellow with Monash University, Clayton, VIC, Australia, where he has been an Associate Professor since 2021. He has authored or coauthored more than 150 peer-reviewed publications and patents in journals and conferences such as *The Lancet Digital Health*, *The British Medical Journal*, *Bioinformatics*, *Hypertension*, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON MEDICAL IMAGING, NeurIPS, CVPR, ICCV, ICLR, KDD, AAAI, and MICCAI. His research interests include artificial intelligence, statistical analysis, machine learning, and the application of computer vision in medical AI for dermatopathology, neurodegenerative diseases, genomics, and ophthalmology. Prof. Ge was selected as one of the 200 Most Qualified Young Researchers in Computer Science and Mathematics by the Scientific Committee of the Heidelberg Laureate Forum Foundation in 2017.

Longxiang Gao (Senior Member, IEEE) received the Ph.D. degree in computer science from Deakin University, Australia. He was a Senior Lecturer with the School of Information Technology, Deakin University, and a Postdoctoral Research Fellow with IBM Research & Development, Australia. He is currently a Professor with the Qilu University of Technology (Shandong Academy of Sciences) and Shandong Computer Science Center (National Supercomputer Center in Jinan). His research interests include Fog/Edge computing, Blockchain, data analysis and privacy protection. Dr. Gao has more than 100 publications, including patent, monograph, book chapter, journal and conference papers. Some of his publications have been published in the top venue, such as IEEE TRANSACTIONS ON MOBILE COMPUTING, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, IEEE INTERNET OF THINGS JOURNAL, IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, and IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING. He has being Chief Investigator (CI) for more than 20 research projects (the total awarded amount is over $5million), from pure research project to contracted industry research.