Security Analysis of Federated Learning in Healthcare

1. Introduction

Federated learning (FL)[1] has emerged as an increasingly promising framework in privacy-preserving machine learning. In FL, local models (clients) are trained on private datasets and only share their gradient updates to a global model (server). The server aggregates these updates to provide a holistic model. Since raw data is not shared, this provides an added layer of privacy, which is important in data-sensitive sectors like healthcare. With greater AI integration into our lives, individual data urgently needs to be secured.

While FL remains largely at the research stage, its adoption is growing through pilot programs with NVIDIA launching a pilot programme with 20+ hospitals[39], Owkin[72], and the MELLODDY project[73].

This research aims to review the main security risks of FL, namely 1) malicious clients, 2) gradient inversion, and 3) membership inference, examining their impacts on healthcare. A policy framework is further proposed to mitigate those risks.

2. Background

AI in healthcare has significant potential for improving patient welfare, such as in early disease identification or specialised healthcare[6, 7, 8].

For example, melanoma has a 32% 5-year survival rate if detected at a metastatic stage, compared to 99% when detected earlier[70]. Differentiating between early-stage melanoma and benign skin lesions is visually difficult, often leading to misdiagnosis or a delay in diagnosis. AI technology can significantly help in identification, enabling clinicians to make more accurate diagnoses and offer therapeutic interventions earlier in patients' disease trajectories. Preliminary research has found that the overall classification accuracy of AI was on par with that of two dermatologists (72.1% vs 66.0% and 65.6%, respectively)[71].

However, research has found data protection risks associated with AI models[2]. Memorisation of data may occur through overfitting abundant parameters to small datasets or optimising the generalisation of long-tailed data distributions[3]. Even when training data is not memorised, malicious actors can infer personal information through gradient inversion attacks, which reverse-engineer the gradients sent to the server to reveal private information[4]. Against this, the existing privacy-preserving strategies (such as data sanitisation) provide limited privacy protection when applied to AI[5].

This raises the question of how personal data may be processed to train AI models. Medical privacy laws could hinder research on AI systems. To encourage further research in this area, AI systems should be designed to be more secure through regulations that ensure greater peace of mind when using patient data for research. As security practices have been found to affect the quality and latency of the model[9], researchers must also strike a balance between data privacy and AI performance.

This research differs from previous work by focusing specifically on healthcare and conducting targeted experiments using healthcare datasets to illustrate the practical consequences of security threats, unlike past healthcare research[16, 17]. By grounding the analysis in real-world healthcare scenarios, the study provides domain-specific insights that are often absent from broader AI policy discussions. These security threats are collated into a policy proposal to provide actionable insights for governments, researchers, and hospitals.

3. Methodology

3.1 Datasets

Three different datasets were used for testing:

The MNIST dataset is a collection of handwritten digits from 0 to 9 containing 60,000 training images and 10,000 testing images. The MNIST images are high contrast, greyscale, centred, and normalised, making them suitable for machine learning algorithms and as a good starting point for each security test.

The Chest X-Ray Dataset[19] includes 5,863 images split into 2 categories: normal scans and pneumonia scans. The Chest X-ray images (anterior-posterior) were selected from retrospective cohorts of pediatric patients one to five years old from Guangzhou Women and Children's Medical Centre, Guangzhou. All chest X-ray imaging was performed as part of the patients' routine clinical care. Two expert physicians graded the image diagnoses before being cleared for AI training. This dataset is more complex than MNIST, such that gradient inversion can be investigated with different difficulty levels. Furthermore, as imaging scans are increasingly seen as a viable method for AI to improve clinical processes by image classification, such image data will likely be adopted in clinical settings. Therefore, it is important to identify the risks associated with AI usage.

The Cardiovascular Dataset includes 70,000 records of patient data and 11 features (age, height, weight, gender, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoking, alcohol intake, physical activity) to arrive at the target (presence of cardiovascular disease). Prediction datasets are also being researched for their accuracy in identifying early signs of illnesses[20], hence this dataset investigates the unique risks of textual data compared to the previously used image data.

In this FL setup, non-IID data, where individual data points are not independent and identically distributed (IID), is used to train individual clients. This is more representative of real-life datasets, as different hospitals may see different patient distributions. Illness clusters are also likely to form around certain hospitals instead of being evenly distributed.

3.2 Attacking Methods

3.2.1 Model Poisoning

In FL, as the server only aggregates results from various clients, it cannot verify the accuracy of local models. This poses a threat where adversaries pose as malicious clients by providing misleading gradients to poison the server model[10, 11]. The final model may be highly inaccurate. If medical professionals rely on a poisoned model for decision-making, patients might receive harmful medical advice.

Two attacks are used to investigate the impact of malicious clients on model performance: 1) label flipping involves altering the true labels of the data points, and 2) random label assignment introduces randomness by assigning labels arbitrarily, skewing model training. Both can be performed with black-box access to the model, where the adversary may not exactly know the structure of the model, allowing them to launch attacks more easily across different types of data.

3.2.2 Gradient Inversion

Despite clients preserving their training data, research has shown that adversaries can still recover training data from gradient or weight updates, violating the promise of data privacy in FL[12, 13, 14]. These gradient inversion attacks operate by optimising over the input space to search for samples whose gradient matches the observed gradient. It aims to reverse-engineer the gradients intercepted from the model to recover sensitive training data[21, 22]. These attacks remain effective even when clients utilise secure aggregation[9] to avoid revealing individual gradients.

Gradient inversion can allow attackers access to patient data. This poses the highest breach of privacy, as it directly exposes personal health data, violating confidentiality laws in many countries. From a discussion with a medical expert, medical information that is particularly sensitive would be any item that can identify the patient. If identified, attackers may blackmail patients or commit identity fraud, and employers or insurers may use sensitive data to discriminate against patients. Since images contain multitudes of data points, they could more easily and uniquely identify the patient. Images may also be able to reveal much more about their patient than purely textual data (which may or may not have their labels attached). Hence, both image and textual datasets were used in testing.

In this project, the image reconstruction attack works as follows:

1. During backpropagation, model gradients will contain partial information about input data when they flow through model layers. These gradients are captured when sent to the global model.
2. Gradient descent is performed on these gradients to reconstruct the original input images.
3. Generative Adversarial Networks (GANs), which have been found to improve image reconstruction[23], were used to regenerate image data based on gradients intercepted from the model. GANs can reconstruct images by learning underlying data distributions through adversarial training between a generator and a discriminator.
4. Starting from noise, the adversary uses the intercepted gradients to update this image in the direction that minimises the difference between the gradient and the reconstructed gradient, reconstructed using a trained GAN model.
5. After several iterations, the adversary can generate an image that resembles the original training data.

Similarly to steps 1-2, the textual reconstruction attack receives the gradients and rescales the reconstructed tensors to resemble the original data.

3.2.3 Membership Inference

When AI systems memorise training data points, the model behaves differently on data depending on whether it was included in the training set[3]. Adversaries can infer sensitive information about training data even with only black-box access to the model, creating significant privacy risks.

In healthcare, membership inference attacks[18] are dangerous because attackers can either 1) identify the model function by feeding it different scans or 2) identify if a person has been involved in model training or visits a hospital for a certain illness if their scans are identified. This can be dangerous if data is used from small clinics catering to specific illnesses or personnel (e.g. athletes or politicians), possibly revealing the medical history of their clientele. Specifically, attackers may input physically observable data into the model (e.g. gender, height, weight) and the model could reveal an individual with those characteristics in the system.

The gradients are intercepted during model training to train a separate membership inference classifier to predict whether a given data point was part of the training set. This involves confidence scores, where if the model is overly confident in its prediction of a data point during training, it may be more easily inferred as part of the training set. The attack's effectiveness is evaluated by measuring the accuracy of the classifier in distinguishing between training and non-training data.

3.3 Protection Methods

3.3.1 Robust Aggregation[24]

When combining the learned parameters across all client models, the simplest aggregation method uses the mean of client weights as the server's model weights. This runs the risk of even one malicious client significantly skewing the results, particularly with few clients.

Hence, the median of client weights can be used. Furthermore, gradient clipping caps the size of gradients so no single client can exert excessive influence on the global model, especially if their gradients are unusually dissimilar due to malicious intent or outlier data. Furthermore, should clients repeatedly provide gradients that stray from other clients, they are removed from consideration to avoid poisoning the entire dataset.

3.3.2 Homomorphic Encryption

Homomorphic encryption allows computations to be performed directly on encrypted data without exposing raw inputs[25]. If an adversary intercepts intermediate values during inference, the still encrypted form prevents meaningful data extraction, protecting against gradient inversion and input reconstruction attacks. TenSEAL[18, 26], a library that supports homomorphic encryption based on Cheon-Kim-Kim-Song (CKKS), is used.

3.3.3 Differentially Private Stochastic Gradient Descent[15]

Differentially private stochastic gradient descent (DPSGD) improves privacy by injecting noise into the gradients during optimisation[27]. Hence, even if gradients are leaked during training, the noisy gradients can be made unusable in both inversion and inference attacks.

3.4 Computational Environment and Model Architecture

All computations were performed on a local machine with a dual-core Intel CPU (2.3 GHz, 8 GB RAM).

Sequential models were integrated with TensorFlow Federated. Using fewer parameters through layer stacking and simpler architectures allows for faster aggregation and easier convergence across the network. The model can also be easily adapted to a variety of classification tasks. Each client can train the same model architecture with minimal computational burden, making it highly scalable across many clients. Defensive layers can be easily incorporated into the Sequential model to test various protective strategies.

In the following experiments, Accuracy is used to evaluate model quality.

4. Results

4.0 Preliminary Results

4.0.1 Data Splits

	IID Data	Non-IID Data
Accuracy	0.849	0.829

Table 1: Accuracy Trained on IID Data and Non-IID Data

The data from Table 1 was collated using the Chest X-ray Dataset. This dataset was used due to its more complex model architecture and healthcare applicability. Non-IID data (individual data points are not independent and identically distributed) leads to reduced model performance and slower convergence than IID data, as wildly different training data creates low consensus for optimised gradients. There could be significant weight divergence across clients, making it difficult to aggregate a globally optimal model. However, non-IID data is more representative of real-life datasets and distributions, making it necessary for research.

4.0.2 Federated Learning Architecture

Number of Clients	1	3	5	8	10	15
Accuracy	0.849	0.829	0.811	0.806	0.796	0.779

Table 2: Accuracy against Number of Clients

From Table 2, accuracy worsens as the number of clients increases. This could pose future problems in collaboration with many hospitals, as the accuracy of the final model may not be as good as having fewer clients. Hospitals may have to pre-aggregate their datasets to reduce the number of clients present. Hospitals will have to contend with a trade-off between accuracy and the extent of data sharing.

4.0.3 Gradient Inversion

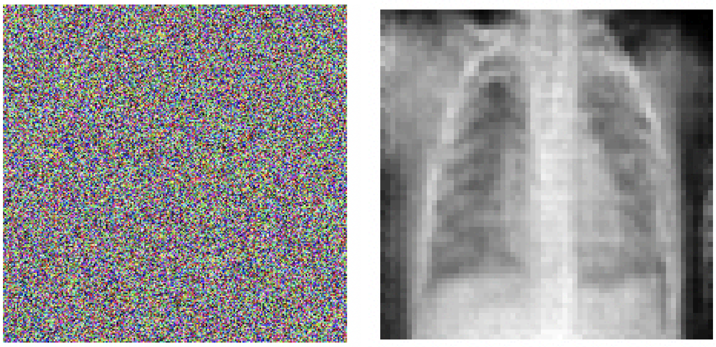


Fig 1. Reconstructed Images Using Models Without GAN (left) and With GAN (right)

A gradient inversion is conducted using the Chest X-Ray Dataset, seen in Fig 1. Without a GAN, the model cannot identify any features. Hence, a black-box attack is much less successful than an attack where the function of the type of training images is already known to the attacker, for example, brain tumour scans. The attacker then trains on publicly available data of brain tumour scans to regenerate training images more successfully. This also highlights the importance of limiting public access to medical imaging data, even with patient consent and data anonymisation.

4.1 Model Poisoning

4.1.1 MNIST

Percentage of Malicious Clients (%)	0	20	40	60	80
Base Model Accuracy	0.969	0.854	0.726	0.574	0.421
Protected Model Accuracy	0.973	0.963	0.886	0.838	0.090

Table 3: Accuracy against Percentage of Malicious Clients on Base Model and Protected Model

4.1.2 Cardiovascular Dataset

Percentage of Malicious Clients (%)	0	20	40	60	80
Base Model Accuracy	0.716	0.711	0.495	0.341	0.294
Protected Model Accuracy	0.715	0.710	0.687	0.362	0.295

Table 4: Accuracy against Percentage of Malicious Clients on Base Model and Protected Model

From Tables 3 and 4, when the percentage of malicious clients increases, the accuracy of the unprotected base model decreases, since malicious gradients overwhelm the final aggregated model. With the protected model, as the percentage of malicious clients rises, accuracy remains relatively high initially. Even with no malicious clients, robust aggregation can prevent the server from catering to extreme outliers, improving its generalisability.

However, if malicious clients overwhelm the entire client base (Table 3), the model risks ignoring the true clients and straying far from the optimised model.

Furthermore, the protected model could underperform (Table 4) if all clients are trustworthy but have non-IID splits, which is likely to occur when aggregating information from different medical institutions. Certain clients could have an abundance of one training category, while another client has none. Clients will learn differently, which may cause the weights to vary significantly, leaving a final server weight that may not represent any client. Hence, it is recommended that each client model should contain a minimum percentage of each category for better results.

Hence, with evidence of its success, secure aggregation is recommended to protect FL models. The global model should be regularly tested to identify if malicious clients are overwhelming the entire server.

4.2 Gradient Inversion

4.2.1 MNIST

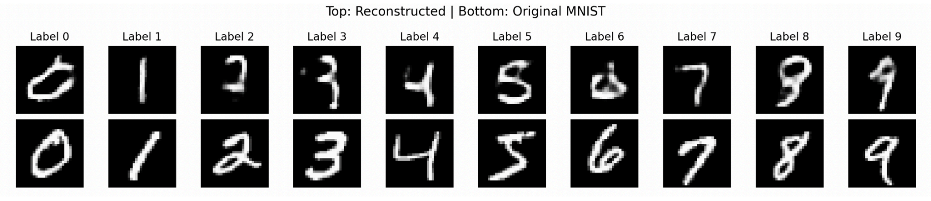
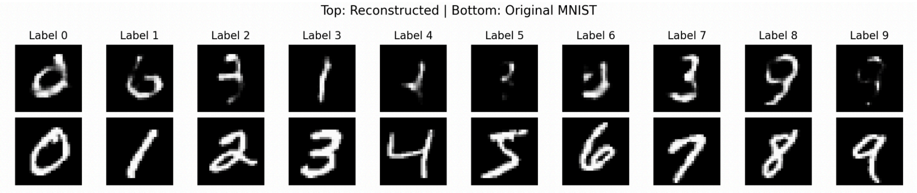
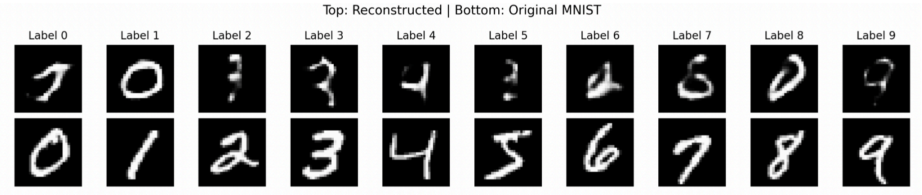
Model	Reconstructed Image	Accuracy	Time (s)
Base Model		0.978	154.14
TenSEAL		0.977	198.43
DPSGD		0.978	229.70

Table 5: Reconstructed and Original Images with Base Model, TenSEAL Protected Model, and DPSGD Protected Model

4.2.2 Chest X-Ray Dataset

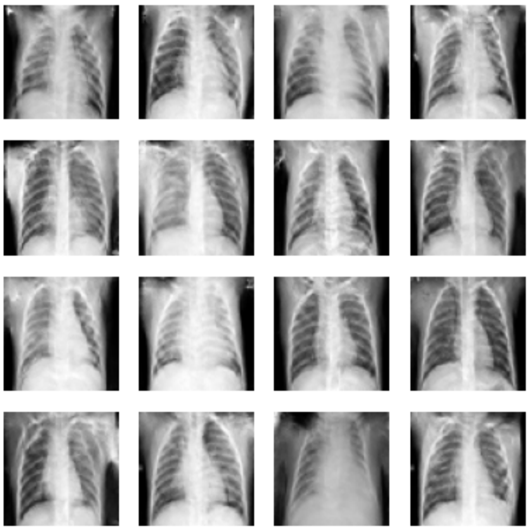


Fig 2. Reconstructed Images Using GAN Models

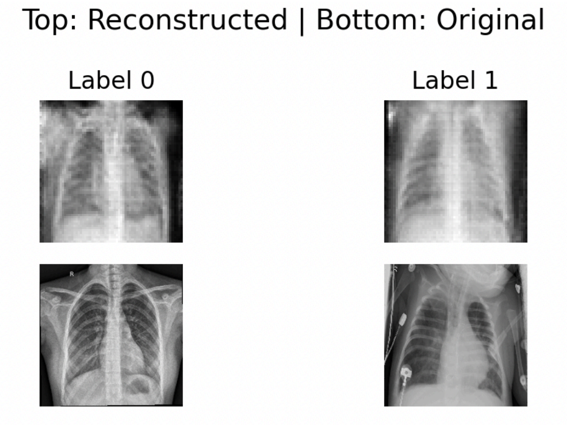
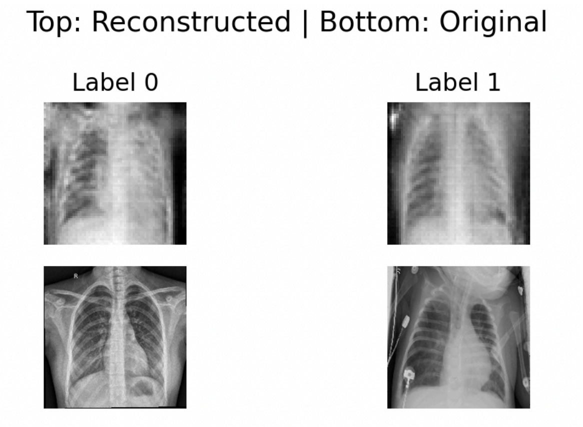
Base Model	 Top: Reconstructed | Bottom: Original Label 0 Label 1
DPSGD	 Top: Reconstructed | Bottom: Original Label 0 Label 1

Table 6: Reconstructed and Original Images with Base Model and DPSGD Protected Model (Label 0: Normal Lung; Label 1: Pneumonia Lung)

Loss per Image Type	Base Model	DPSGD
Normal	0.0193	0.1364
Pneumonia	0.0007	0.0049

Table 7: Loss for Reconstructed Images with Base Model and DPSGD Protected Model

The images from Fig 2 are reconstructed images using gradient inversion. A medical doctor was consulted and observed that the images are identifiably different from actual chest X-rays, as chest X-rays are much sharper with a defined rib structure. However, the top right image has been noted for its similarity to an actual X-ray by the doctor. The image on the top left has also been correctly identified as a child's X-ray with a serious case of pneumonia. Hence, data features and individual information can be extracted purely from gradient updates, highlighting the dangers of gradient inversion. A doctor has highlighted that medical information that is particularly sensitive would be any item that can identify the patient, such as a rare deformity or tattoos. For example, skin scans could uniquely identify patients if tattoos are present.

From Table 5, when the model is encrypted with TenSEAL and DPSGD, the gradient inversion attacks are less successful in retrieving key features of the training data. Similarly in Table 6, although the rib structure is visible, the images are slightly less distinct in colour. In Table 5, the accuracy remains similar while the time taken increases, though latency may be exaggerated due to computational limitations. This trade-off is recommended as privacy protection should be paramount. However, the model can still retrieve data features due to the GAN model, as mentioned in Section 4.0.3.

In Table 7, loss can quantify the difference between the training images and the regenerated images. After 500 epochs, the final loss values for the DPSGD model are higher than the unprotected base model. Evidently, DPSGD can reduce the closeness of the regenerated image to the original image.

4.2.3 Cardiovascular Dataset

Each value in the results array corresponds to a feature value (age, gender [1 or 2], height, weight, systolic blood pressure, diastolic blood pressure, cholesterol [1, 2, or 3], glucose [1, 2, or 3], smoking [0 or 1], alcohol intake [0 or 1], physical activity [0 or 1]).

Model	Data Source	Results	Percentage Difference
Base Model	Reconstructed Data	[52, 1, 162, 71, 121, 82, 1, 1, 0, 0, 1]	0.85%
	Actual Data	[50, 1, 163, 73, 122, 81, 1, 1, 0, 0, 1]	
DPSGD	Reconstructed Data	[58, 2, 167, 69, 124, 82, 1, 1, 0, 0, 1]	22.54%
	Actual Data	[64, 1, 170, 80, 140, 90, 3, 1, 0, 0, 1]	

Table 8: Reconstructed and Original Data with Base Model and DPSGD Protected Model

In Table 9, the prediction accuracy measures how accurately the model can predict if the patient has cardiovascular disease.

Batch Size	Prediction Accuracy	Data Source	Results
1	0.716	Reconstructed Data	[55, 1, 165, 78, 131, 90, 1, 1, 0, 0, 1]
		Actual Data	[55, 1, 165, 78, 130, 90, 1, 1, 0, 0, 1]
4	0.713	Reconstructed Data	[56, 1, 167, 79, 128, 84, 1, 1, 0, 0, 1]
		Actual Data	[54, 2, 168, 78, 125, 82, 1, 1, 0, 0, 1]
16	0.710	Reconstructed Data	[55, 2, 168, 72, 120, 81, 1, 1, 0, 0, 1]
		Actual Data	[52, 2, 168, 74, 124, 81, 1, 1, 0, 0, 1]
32	0.709	Reconstructed Data	[52, 1, 162, 71, 121, 82, 1, 1, 0, 0, 1]
		Actual Data	[50, 1, 163, 73, 122, 81, 1, 1, 0, 0, 1]

Table 9: Reconstructed and Original Data Against Batch Size

In Table 8, the percentage difference indicates how far the reconstructed data is from the actual training data. An attacker may be able to uncover extensive training data purely from gradients from the base model, without access to any original data. The base model is highly insecure while DPSGD can prevent the close reconstruction of training data.

The batch size is investigated in Table 9. The batch size represents the number of training samples used per iteration. Higher batch sizes tend to provide faster training times, but lower accuracy and higher computational power is required. When tested with a batch size of 1, the attack can reconstruct every data point precisely. With lower batch sizes, the model may memorise each training example more, causing gradients to pinpoint data used more closely. Even though the prediction accuracy decreases, a higher batch size should be a valid trade-off against privacy. It is recommended that batch sizes of at least 8 be used to prevent such close reconstruction.

4.3 Membership Inference

4.3.1 MNIST

	Base Model	DPSGD
ROC-AUC	0.537	0.479
Accuracy	0.978	0.975

Table 10: ROC-AUC and Accuracy with Base Model and DPSGD Protected Model

4.3.2 Chest X-Ray Dataset

	Base Model	DPSGD
ROC-AUC	0.589	0.538
Accuracy	0.709	0.707

Table 11: ROC-AUC and Accuracy with Base Model and DPSGD Protected Model

The ROC-AUC, area under the Receiver Operating Characteristic curve, indicates how accurately membership is inferred. DPSGD can reduce the success rate of distinguishing between training and test data. Meanwhile, accuracy is not significantly affected by DPSGD, making it a valuable trade-off, as seen in Tables 10 and 11.

However, it is noted that the success rate of the base models is not especially high. More sophisticated membership inference attacks may be possible. Therefore, improved privacy protection methods may be needed.

4.4 Analysis of Results

Given the extent of private information revealed in each of these attacks, it is important to establish standards of protection for AI models. As these defences are effective, the following framework will encourage their usage in FL models. However, some private data can still be revealed despite these defences. It is recommended that further research be done on more complex defence methods to protect against various FL threats. Still, as this research was done with full knowledge of the model architecture, it may be harder for true attackers to threaten a model that they have only black-box access to. It is also recommended that healthcare data be kept private (unlike now, where chest scans were available on Kaggle), so that GANs cannot be trained well for gradient inversion attacks. Additionally, as research was done without a CPU, there were computational limitations that reduced the scope of research (e.g., increasing the number of epochs or clients, testing more complicated architectures). More research could be done on faster computers that could attack or defend more successfully.

5. Policy Proposal

5.1 Why This Matters

There has been significant promise in integrating AI into healthcare to improve patient welfare[62], whether in early identification of illnesses or specialised healthcare. In medical imaging alone, there have been applications found in tumour detection in oncology[6], genomic characterization[7], tumour subtyping[28], grading prediction[29], outcome risk assessment[30, 31] or risk of relapse quantification[32, 33], chest X-ray analysis[34] and retinal fundus imaging[35]. However, based on the results in Section 4, it is evident that AI systems are not fully secure. These models may be poisoned to produce skewed results (Section 4.1), or the training data could be leaked (Section 4.3) or even fully reconstructed (Section 4.2). This could reduce trust in AI systems, limiting the extent of research possible. To encourage further research in this area, AI systems should be designed to be more secure.

As more healthcare data is kept in electronic records, individual medical data could be sold as commodities if hacked. Marketers could potentially tailor medical advertisements based on personal health information; companies might use employee data to justify discrimination. In more extreme scenarios, governments could misuse sensitive medical history, such as records related to HIV status or abortion care, for targeted surveillance.

Furthermore, in medicine, doctors have the right to a patient's data for research or big data analysis only after the patient has provided informed consent, that is, they are fully aware of the condition, treatment risk and benefits, and how the data will be used. However, to transform personal health information into high-quality medical data that can be used as a production factor, it is usually necessary to go through multiple levels of acquisition, analysis and use, with multiple institutions (including medical institutions, companies, public health departments, etc.) jointly participating in it. This gradually weakens each individual's control over the subsequent transformation and derivation of their own health information into big data. The difficulty of implementing and enforcing the principle of informed consent increases layer by layer.

As security practices have been found to affect the latency of the model (from 3-7× for homomorphic encryption, and up to 2× in Section 4.2)[36, 37, 38], researchers must also strike a balance between data privacy and AI performance. Recent political changes have also shown the fractured state of regulation in AI[40], requiring politicians to achieve a consensus to encourage future research. Therefore, the following proposal aims to establish regulations to ensure greater peace of mind and consensus for higher ethical standards when using patient data for research.

5.2 Current Landscape

5.2.1 Key Challenges

Within most medical research, the option to opt out must be included. However, while AI companies such as OpenAI claim to comply with the right to erasure, it is unclear how this is achieved because personal information may be contained in multiple forms in AI. This increases the complexity of identifying and isolating specific data points in the latent space of the model. Simply deleting data from a training dataset is a superficial solution, as training data may still be extracted from associated information encapsulated within the model's parameters. It is also unfeasible to retrain models for every data removal request. This undermines the integrity of the deletion process and perpetuates potential privacy violations[41].

In addition, medical research laws may follow the data minimisation principle, where data controllers should collect personal data only as relevant and necessary for a specific purpose. However, limiting the amount of data available for analysis could significantly restrict the accuracy of medical AI, which could produce stress on patients in the event of misdiagnoses[42].

Furthermore, current security practices rely on data anonymisation techniques, where identifiable patient data is replaced with ID numbers or otherwise encoded. However, there is no standardised set of recommendations on anonymising clinical trial datasets, and researchers in clinical trials still consider that anonymisation techniques alone are insufficient to protect patient privacy[43]. Some organisations continue to rely on outdated anonymisation methods, such as k-anonymity and l-diversity, despite their well-documented vulnerabilities[44, 45, 46].

5.2.2 Policy Case Studies

Privacy laws can differ significantly between countries. The following are a few examples of current regulations on data usage.

5.2.2.1 Singapore

According to a medical researcher, personal data is fully anonymised with a detailed data management plan, after patient consent. The patient data generally comes from hospital medical records, so companies that participate in research and handle patient data will have to sign NDAs. However, if the anonymisation is reversed, ethics review boards do not factor in, as the data has already been consented to be released.

Privacy protection relies on oversight by hospital ethics committees and the ethical review processes of academic journals. Those boards follow regulations outlined in the Personal Data Protection Act (PDPA)[47] by the Infocomm Media Development Authority (IMDA). The government agency IMDA oversees the responsible adoption of AI across both public and private sectors. The Personal Data Protection Commission (PDPC), a commission within the IMDA, regulates data privacy.

The government has developed various frameworks and tools to guide AI deployment and promote the responsible use of AI, including:

- The Model AI Governance Framework (2020) provides detailed guidance to private sector organisations to address key ethical and governance issues when deploying AI solutions
- AI Verify2, an AI governance testing framework and toolkit designed to help organisations validate the performance of their AI systems against AI ethics principles through standardised tests.
- The National Artificial Intelligence Strategy 2.03 (2023) outlines Singapore's ambition and commitment to building a trusted and responsible AI ecosystem

However, currently, there are no specific laws in Singapore that directly regulate AI.

As the IMDA has control over data protection and technological development, my policy can be implemented by the IMDA to regulate AI companies and medical institutions.

5.2.2.2 China

China has several foundational data protection laws[48], including the Cybersecurity Law, Data Security Law, and Personal Information Protection Law:

1. Cybersecurity Law emphasises that network operators must strictly protect the user information they collect and establish robust systems for user information protection.
2. Data Security Law requires that data processing activities comply with laws and regulations, respect public ethics and professional conduct, and fulfil obligations to protect data security.
3. Personal Information Protection Law stipulates that personal information processors must follow the principles of legality, legitimacy, necessity, and good faith.

In 2022, Administrative Measures for Cybersecurity in Medical and Health Institutions[49] further clarified the full-cycle responsibilities of medical institutions in data collection, storage, use, sharing, and destruction.

Additionally, local governments have issued more detailed implementation guidelines and regional policies tailored to their contexts. For example, according to a researcher in the area, some provinces have introduced regional health information platform data security management standards, specifying detailed requirements for data sharing and exchange.

5.2.2.3 United States

The U.S. adopts a multi-department regulatory model for managing medical data, involving:

1. Food and Drug Administration (FDA) evaluates the security and effectiveness of data related to medical devices.
2. Centers for Medicare & Medicaid Services (CMS) oversees the use of data related to coverage decisions under Medicare and Medicaid.
3. Department of Health and Human Services (HHS) is responsible for privacy protection, especially under the Health Insurance Portability and Accountability Act (HIPAA).

The FDA has proposed a regulatory framework for AI/ML-based software as a medical device[50]. However, this is simply a framework and could face challenges given the recent Big Beautiful Bill[40] that could prevent AI regulation.

Furthermore, the U.S. does not have a dedicated law specifically for medical data privacy. Instead, it relies on existing frameworks such as HIPAA, which applies to “covered entities” like hospitals and insurers, but offers limited regulation over non-clinical or consumer health data (e.g., fitness apps), leaving certain areas relatively underregulated. Given the fractured state of AI regulation, the following policy could instead be included in the HHS as an amendment to HIPAA to focus primarily on patient privacy.

5.2.2.4 European Union

The General Data Protection Regulation (GDPR), implemented in 2018, forms the core of the EU’s data privacy regime. Personal health data is classified as special category data, subject to stricter protections. The patient’s informed consent must be obtained, and the purpose of using health data must be clearly stated. The data minimisation principle must also be followed, where only necessary data should be collected and processed.

Additionally, the European Group on Ethics in Science and New Technologies (EGE) has published guidelines that 1) prohibit the misuse of sensitive data, 2) require algorithmic transparency from companies using medical data.

Specific to AI, the EU has implemented the AI Act (AIA)[51] in 2022 as one of the first pieces of legislation on AI use in the world. The AIA classifies AI into risk categories, where healthcare AI is classified under “high-risk AI” and subject to strict obligations. These include robust data governance, human oversight, traceability, and transparency of algorithms. The AIA complements the GDPR to ensure compliance with overall ethical and privacy standards.

However, differences in implementation across EU member states present challenges in consistent enforcement and policy execution at the national level. Some also argue that overly strict rules may hinder AI innovation, especially in healthcare startups and research institutions.

5.3 Key Stakeholders

This policy should apply to researchers and healthcare professionals developing AI healthcare models. It is recommended that the policy be implemented first in medical establishments such as hospitals, as they have on-the-ground access to patient data and hold greater responsibility in protecting patients. This can also simplify the excessive regulation needed to cover any individual involved in AI research on such data in the future. The policy should also be integrated alongside existing data protection laws in the country to minimise extra manpower required to uphold the policy.

5.3.1 Governments

Governments stand to benefit from this policy, which helps build public trust in emerging health technologies and allows national healthcare systems to reap the benefits of AI without centralising sensitive data. Countries can remain competitive in AI innovation and improve their citizens' quality of life.

However, governments are responsible for keeping pace with rapidly evolving AI and privacy technologies. Failure to do so could result in policy gaps or unaddressed system vulnerabilities. Enforcing compliance across hospitals, research institutions, and AI companies requires investment in oversight infrastructure, legal frameworks, and technical expertise that may be unevenly distributed across countries or jurisdictions. While the policy aligns with broader goals of data protection and digital innovation, its complexity and cost may pose challenges in implementation.

Governments can integrate this policy into existing healthcare privacy policies. As AI regulation is still developing, this policy is important to protect citizens. The policy can also be based on more foundational healthcare protection policies, which can improve its stability and adoption.

5.3.2 Hospitals

Each country has different healthcare systems, broadly split into public/non-profit hospitals and private/for-profit hospitals for this discussion. All hospitals will benefit from access to improved AI tools that support clinical decision-making while ensuring patient data is handled ethically and securely. With federated learning, data remains on-site, reducing the risk of breaches and misuse, and aligning with typically stronger public accountability requirements of these institutions.

Public hospitals may have greater access to public research institutions such as universities or government research institutions, which can encourage ease of integration for data into research. The framework also encourages closer partnerships with research institutions and AI developers, which could bring technical and financial resources into underfunded hospital systems. However, the technical and operational demands of implementing these measures, such as maintaining secure infrastructure, managing consent, and participating in secure aggregation protocols, could burden already resource-constrained public hospitals.

Private hospitals may see strategic advantages in adopting the proposed framework, particularly in maintaining patient trust and competitive positioning in a market increasingly concerned with data privacy. Private hospitals can access cutting-edge tools without sacrificing confidentiality or regulatory compliance by participating in collaborative AI development while retaining control over their patient data. However, implementing the technical components of federated learning requires financial investment and skilled personnel, which not all private institutions may have. Furthermore, compliance demands may reduce flexibility in pursuing external partnerships.

Still, research has shown that AI models matched the performance of human radiologists when acting as the second reader of mammography scans can streamline breast cancer diagnoses by cutting radiologists' workloads by at least 30%[68]. Given the evidence of early success of AI in healthcare decision-making, hospitals can improve patient care by making improved decisions using AI.

5.3.3 Research Institutions

Research institutions stand to gain immensely through their expanded access to real clinical data while minimising privacy concerns that could limit access under current regulations. Also, the emphasis on data security may ease the institutional review process.

However, the reduced access to raw data may impede more exploratory analysis. Researchers must also coordinate across multiple institutions, complicating the research process. For institutions without significant technical infrastructure, meeting the computational demands of privacy-preserving techniques may pose additional barriers to participation.

Still, data protection is necessary to ensure trust in research institutions such that individuals remain willing to provide their data in the future.

5.3.4 AI Companies

Similarly to research institutions, AI companies gain access to rich datasets and opportunities to build more robust, generalisable models through partnerships with hospitals and research institutions. The FL approach allows companies to develop healthcare solutions while adhering to strict privacy laws, enhancing their credibility and marketability in a highly regulated domain.

However, the framework also imposes significant limitations on direct data access, which may hinder model optimisation. Companies must also comply with technical and legal standards across multiple jurisdictions, which require extensive expertise and resources. For startups and smaller firms, these requirements may be a barrier to entry.

The framework actively encourages collaboration between companies and hospitals to pool resources together and share their expertise, mitigating some of the costs.

5.3.5 Patients

Patients are arguably the most protected under this policy, which strongly emphasises personal data privacy. Patients enjoy greater data privacy as FL ensures that sensitive health data remains within the institution where it was collected, substantially reducing the risk of data breaches or unauthorised access. Patients may benefit indirectly from AI models trained across multiple institutions, which can improve clinical decision-making. Research has shown outpatient diagnostic errors of 5.08%, or approximately 12 million US adults every year. About half of these errors could potentially be harmful. If AI can reduce even 5% of misdiagnoses, 0.6 million US adults can be protected every year[69].

However, there are some important costs to consider. Due to how AI models are trained, patients may have limited agency or visibility in how their data contributes to AI development, which may raise concerns about transparency. Some patients may be reluctant to allow their data to be used due to mistrust, reducing the volume and diversity of data available for model training. If the training data is not representative, resulting models may underperform for underrepresented populations, possibly worsening healthcare disparities.

With this framework, patients may trust AI models more, potentially becoming more willing to contribute to these models.

5.4 Timeline

The implementation plan will be as follows:

1. First, medical institutions should incorporate the proposed framework into their internal data protection policies. This may involve migrating patient data to secure cloud infrastructures to meet privacy requirements. To manage potential technical or financial challenges, hospitals can collaborate with national or regional health IT governance bodies for infrastructure support and funding.

2. Next, academic institutions, particularly those conducting AI medical research, should work with their ethics committees to introduce specific review protocols for AI-related data use. This may include assessments of model explainability, dataset bias, and patient consent practices. In tandem, universities may need to upgrade their computational infrastructure (e.g., expand GPU capacity or secure access to research clouds) to meet the technical demands of privacy-preserving large-scale AI model training.

3. In the longer term, governmental bodies should establish clear regulations governing collaborations between AI companies, medical institutions, and academic institutions. These regulations should include formalised data sharing contracts and mandatory audits. Rather than creating new legislation from scratch, these AI-specific rules could be incorporated as amendments to existing data protection laws, such as HIPAA (US) or the GDPR (EU), ensuring consistency with current legal frameworks.

5.5 Proposed Framework: Singapore

To improve existing policy, the following framework is designed for medical research involving AI. The framework focuses on ensuring that AI models are trained using medical data in a way that protects patient privacy and complies with existing legal regulations. The framework has a legal basis in Singapore, given better on-the-ground knowledge of research boundaries and legal practices.

Data Collection

- Individuals must provide informed consent before their data is used for research. This is currently already regulated in the PDPA. Their consent should include:
    - Clear descriptions of how their data will be used.
    - The purpose of AI training, including any risks and benefits.

    - The option to opt out at any time. However, as mentioned in Section 5.2.1, the opting-out process may not be easily achieved as the model has already converted training data into vector space, making it difficult to isolate individual data points.
- As IMDA often provides programmes and grants[63] for Singapore's technological development, some of its funding can be set aside for healthcare research. IMDA can run pilot programmes where the health data of deceased individuals is used to train AI models. Deceased individuals may provide a greater collection of data that includes more data points, which could improve model performance. However, this could create biased models, which is later addressed in the Model Evaluation section.

Data Security

- Basic security standards
    - Use advanced encryption (AES-256) for data storage and transmission. The PDPA currently regulates that an organisation must protect personal data in its possession or under its control by making reasonable security arrangements to prevent unauthorised access, collection, use, disclosure, copying, modification or disposal, or similar risks; and the loss of any storage medium or device on which personal data is stored. While no specific actions are outlined in the Act, the IMDA can encourage secure encryption in its best practices section for organisations. All systems processing medical data for AI research should be hosted on secure platforms (e.g., HIPAA-compliant cloud infrastructure[53]). Systems processing medical data must comply with national standards outlined in the Cybersecurity Act 2018[51].
    - All data used in model training should be de-identified, where personal identifiers are replaced with unique pseudonyms, as is currently practised.
- Federated Learning security standards
    - The IMDA can encourage the adoption of FL to minimise data sharing. This can be included in the Emerging Technologies[64] IT Standards and Frameworks. Data can remain within local medical institutions. If local medical institutions are unable to train the model accurately, its data may be aggregated on a city or state level, up to where sufficient data is available for training. Central servers should only receive aggregated, noise-added updates, preventing reconstruction of any contributing data source.
    - On untrusted systems, homomorphic encryption techniques are recommended so that model updates are encrypted. This ensures no party can access the raw data or reverse-engineer the updates.
        - However, HE is known for its slow performance and computational requirements[54, 55], which some institutions may not have the capacity to handle. Hence, where homomorphic encryption is not feasible due to performance constraints, secure multiparty computation is recommended, based on early research of its relative success[56].
    - Differential privacy
        - Implement differential privacy during the model aggregation process to prevent the leakage of sensitive patient data through model updates. Differential privacy is found to be successful in preventing both gradient inversion (Section 4.2) and membership inference (Section 4.3) attacks.
    - Aggregation
        - Secure aggregation techniques, such as Bonawitz et al.'s protocol[57], should be used to ensure that model updates are combined so that no party can see or

extract individual updates. Secure aggregation can reduce the effects of model poisoning (Section 4.1) and membership inference (Section 4.3)

Data Sharing

- Data Sharing Agreements (DSAs)
    - Formalise partnerships between hospitals, researchers, and AI companies with clear Data Sharing Agreements. Currently, IMDA provides a Data Sharing Framework[65] and samples. As it oversees AI companies, IMDA is at a prime spot for facilitating DSAs between AI companies, researchers, and hospitals. DSAs can include the following:
        - Explicitly outline data ownership, usage rights, and data protection responsibilities.
        - Specify which parties have access to the data and the conditions under which data may be shared or reused.
        - NDAs must be included to protect research results before publishing, and raw data
- Third-party data use
    - Any third-party AI companies or research institutions accessing the data must adhere to the same data protection standards with legally binding contracts that include penalties for violations. These contracts can be investigated by the PDPC for violations against personal data protection. The PDPA Offences and Penalties Section[66] can keep organisations in line. Specific penalties can be outlined in an amendment in the future.

Model Evaluation

- Model performance
    - Research on AI must ensure that the data used to train AI models represents diverse populations. In particular, this highlights how relying on data from deceased individuals to limit privacy risks could skew results due to age or sampling bias. Reasonable effort must be made to address these questions:
        - Effectiveness: Does the model improve patient outcomes?
        - Fairness: Is the model fair across different demographic groups (e.g., gender, race, socio-economic status)?
        - Privacy: Does the model ensure data protection and patient confidentiality?
    - The model must also be repeatedly tested on unseen data to verify that no malicious clients are involved in training, as discussed in mitigation techniques for model poisoning in Section 4.1. A malicious client is defined as an untrusted component that can collude in attacks during FL by affecting the global FL model's performance through participation in the training process.
- Model refinement
    - Continuous refinement based on feedback from healthcare workers, researchers, and patients is necessary to prevent model drift[58], where model performance degrades over time. This ensures that the models evolve to incorporate the latest clinical knowledge. There exists some work on correcting deployed models in a way that does not require re-training end-to-end (e.g. fine-tuning[59], and in-context learning[60, 61]). Still, more work remains, especially for AI systems with many interacting parts.

This framework extends current data protection legislation and ethical research standards. Hence, the IMDA can initially include this framework in its best practices. The PDPC could amend the PDPA to cover AI data protection further. As the IMDA regulates the overall technology sector, it can work closely with IT companies to encourage secure AI adoption. The government can regulate universities and hospitals (through the Ministry of Health) to integrate this framework into existing ethics boards to oversee research methodologies and data handling practices in healthcare research. Existing ethics boards should ensure that at least one individual with prior experience in AI is involved to understand technical details in data ingestion and model training. These boards will ensure that all partners (e.g., hospitals, researchers, AI companies) follow strict governance protocols for sharing and processing data. Much like how current boards operate, if the researchers do not outline sufficiently secure practices, their research will not be allowed to proceed. When submitting their research to journals, journals may not accept research that does not outline secure practices, although this may require further international coordination.

5.6 Future Updates

As AI models become more sophisticated, attacks will naturally increase in difficulty. To contend with more insidious attacks, it is necessary to consistently update the framework by working closely with researchers at the forefront of FL research, such as those in Singapore's national universities. Singapore's small size works to its advantage. The government can closely regulate institutions for its security and facilitate collaboration between various sectors.

6. Conclusion

This paper provides experimental results on various privacy attacks against federated learning models specific to healthcare uses. It is found that the privacy attacks can reveal significant training information, although mitigation methods demonstrate some success in preventing these attacks. To encourage further healthcare research into AI, the proposed policy lists several concrete directions for AI regulation that could improve the adoption of secure healthcare AI systems. Further research into machine unlearning is recommended to completely purge patient data in compliance with data privacy laws. Furthermore, this work can be extended to other data-sensitive sectors, especially as AI is integrated into areas such as defence[67].

7. References

1. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics* (pp. 1273-1282). PMLR.
2. Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., ... & Wallace, E. (2023). Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)* (pp. 5253-5270).
3. Feldman, V. (2020, June). Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd annual ACM SIGACT symposium on theory of computing* (pp. 954-959).
4. Model inversion attacks that exploit confidence ... (n.d.). https://rist.tech.cornell.edu/papers/mi-ccs.pdf
5. Brown, H., Lee, K., Mireshghallah, F., Shokri, R., & Tramèr, F. (2022, June). What does it mean for a language model to preserve privacy?. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency* (pp. 2280-2292).
6. McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., Ledsam, J. R., … Shetty, S.

(2020). International evaluation of an AI system for breast cancer screening. *Nature*, *577*(7788), 89–94. https://doi.org/10.1038/s41586-019-1799-6

7. Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., Naidich, D. P., & Shetty, S. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine*, *25*(6), 954–961. https://doi.org/10.1038/s41591-019-0447-x
8. Kaissis, G., Ziegelmayer, S., Lohöfer, F., Steiger, K., Algül, H., Muckenhuber, A., Yen, H. Y., Rummeny, E., Friess, H., Schmid, R., Weichert, W., Siveke, J. T., & Braren, R. (2019). A machine learning algorithm predicts molecular subtypes in pancreatic ductal adenocarcinoma with differential response to gemcitabine-based versus FOLFIRINOX chemotherapy. *PloS one*, *14*(10), e0218642. https://doi.org/10.1371/journal.pone.0218642
9. Podschwadt, R., Takabi, D., & Hu, P. (2021). Sok: Privacy-preserving deep learning with homomorphic encryption. *arXiv preprint arXiv:2112.12855*.
10. Xie, Y., Fang, M., & Gong, N. Z. (2025). Model poisoning attacks to federated learning via multi-round consistency. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 15454-15463).
11. Feng, C., Li, Y., Gao, Y., Celdrán, A. H., von der Assen, J., Bovet, G., & Stiller, B. (2025). DMPA: Model Poisoning Attacks on Decentralized Federated Learning for Model Differences. *arXiv preprint arXiv:2502.04771*.
12. Geiping, J., Bauermeister, H., Dröge, H., & Moeller, M. (2020). Inverting gradients-how easy is it to break privacy in federated learning?. *Advances in neural information processing systems*, *33*, 16937-16947.
13. Zhu, L., Liu, Z., & Han, S. (2019). Deep leakage from gradients. *Advances in neural information processing systems*, *32*.
14. Yang, W., Wang, S., Wu, D. *et al.* Deep learning model inversion attacks and defenses: a comprehensive survey. *Artif Intell Rev* **58**, 242 (2025). https://doi.org/10.1007/s10462-025-11248-0
15. Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep Learning with Differential Privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS '16). Association for Computing Machinery, New York, NY, USA, 308–318. https://doi.org/10.1145/2976749.2978318
16. Dhade, P., & Shirke, P. (2023). Federated Learning for Healthcare: A Comprehensive Review. Engineering Proceedings, 59(1), 230. https://doi.org/10.3390/engproc2023059230
17. Nguyen, D. C., Pham, Q. V., Pathirana, P. N., Ding, M., Seneviratne, A., Lin, Z., ... & Hwang, W. J. (2022). Federated learning for smart healthcare: A survey. ACM Computing Surveys (Csur), 55(3), 1-37.
18. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017, May). Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)* (pp. 3-18). IEEE.
19. Kermany, Daniel; Zhang, Kang; Goldbaum, Michael (2018), “Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification”, Mendeley Data, V2, doi: 10.17632/rscbjbr9sj.2
20. Al-Dekah, A. M., & Sweileh, W. (2025). Role of artificial intelligence in early identification and risk evaluation of non-communicable diseases: a bibliometric analysis of global research trends. BMJ open, 15(5), e101169. https://doi.org/10.1136/bmjopen-2025-101169
21. Qian, J., Wei, K., Wu, Y., Zhang, J., Chen, J., & Bao, H. (2024, August). Gi-smn: Gradient inversion attack against federated learning without prior knowledge. In *International Conference on Intelligent Computing* (pp. 439-448). Singapore: Springer Nature Singapore.
22. Wu, R., Chen, X., Guo, C., & Weinberger, K. Q. (2023, July). Learning to invert: Simple adaptive attacks for gradient inversion in federated learning. In *Uncertainty in Artificial Intelligence* (pp. 2293-2303). PMLR.
23. Hitaj, B., Ateniese, G., & Perez-Cruz, F. (2017, October). Deep models under the GAN: Information leakage from collaborative deep learning. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security* (pp. 603-618).

24. Campos, E. M., Gonzalez-Vidal, A., Hernández-Ramos, J. L., & Skarmeta, A. (2024). FedRDF: A robust and dynamic aggregation function against poisoning attacks in federated learning. *IEEE Transactions on Emerging Topics in Computing*.
25. Jin, W., Yao, Y., Han, S., Gu, J., Joe-Wong, C., Ravi, S., ... & He, C. (2023). FedML-HE: An efficient homomorphic-encryption-based privacy-preserving federated learning system. *arXiv preprint arXiv:2303.10837*.
26. Benaissa, A., Retiat, B., Cebere, B., & Belfedhal, A. E. (2021). Tenseal: A library for encrypted tensor operations using homomorphic encryption. *arXiv preprint arXiv:2104.03152*.
27. S. Song, K. Chaudhuri and A. D. Sarwate, "Stochastic gradient descent with differentially private updates," 2013 IEEE Global Conference on Signal and Information Processing, Austin, TX, USA, 2013, pp. 245-248, doi: 10.1109/GlobalSIP.2013.6736861.
28. Pinker, K., Chin, J., Melsaether, A. N., Morris, E. A. & Moy, L. Precision medicine and radiogenomics in breast cancer: new approaches toward diagnosis and treatment. Radiology 287, 732–747 (2018).
29. Lu, H. et al. A mathematical-descriptor of tumor-mesoscopic-structure from computed-tomography images annotates prognostic- and molecularphenotypes of epithelial ovarian cancer. Nat. Commun. 10, 764 (2019).
30. Kaissis, G. et al. A machine learning model for the prediction of survival and tumor subtype in pancreatic ductal adenocarcinoma from preoperative difusion-weighted imaging. Eur. Radiol. Exp. 3, 41–41 (2019).
31. Cui, E. et al. Predicting the ISUP grade of clear cell renal cell carcinoma with multiparametric MR and multiphase CT radiomics. Eur. Radiol. 30, 2912–2921 (2020).
32. Varghese, B. et al. Objective risk stratifcation of prostate cancer using machine learning and radiomics applied to multiparametric magnetic resonance images. Sci. Rep. 9, 1570 (2019).
33. Elshafeey, N. et al. Multicenter study demonstrates radiomic features derived from magnetic resonance perfusion images identify pseudoprogression in glioblastoma. Nat. Commun. 10, 3170 (2019).
34. Rajpurkar, P. et al. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. Preprint at https://arxiv.org/ abs/1711.05225 (2017).
35. Poplin, R. et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat. Biomed. Eng. 2, 158–164 (2018).
36. Florian Bourse et al. Fast Homomorphic Evaluation of Deep Discretized Neural Networks. In Advances in Cryptology, 2018.
37. Pejic, I., Wang, R., & Liang, K. (2022). Effect of homomorphic encryption on the performance of training federated learning generative adversarial networks. arXiv preprint arXiv:2207.00263.
38. Ehsan Hesamifard et al. Deep neural networks classification over encrypted data. In ACM Conference on Data and Application Security and Privacy, 2019.
39. NVIDIA & Mass General Brigham Hospital Federated Learning Project Predicts COVID-19 Patient Oxygen Need Using 20 Days of Data From 20 Hospitals. (2020, 10 07). Synced.
40. H.R.1 - One Big Beautiful Bill Act 119th Congress (2025-2026), Sec 0012
41. De Cristofaro, E. (2020). An overview of privacy in machine learning. arXiv preprint arXiv:2005.08679.
42. Melanie Sloan, Michael Bosley, Caroline Gordon, Thomas A Pollak, Farhana Mann, Efthalia Massou, Stephen Morris, Lynn Holloway, Rupert Harwood, Kate Middleton, Wendy Diment, James Brimicombe, Elliott Lever, Lucy Calderwood, Ellie Dalby, Elaine Dunbar, David D'Cruz, Felix Naughton, 'I still can't forget those words': mixed methods study of the persisting impact on patients reporting psychosomatic and psychiatric misdiagnoses , Rheumatology, Volume 64, Issue 6, June 2025, Pages 3842–3853, https://doi.org/10.1093/rheumatology/keaf115
43. Rodriguez, A., Tuck, C., Dozier, M. F., Lewis, S. C., Eldridge, S., Jackson, T., Murray, A., & Weir, C. J. (2022). Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review. Clinical trials (London, England), 19(4), 452–463. https://doi.org/10.1177/17407745221087469
44. Gadotti, A., Rocher, L., Houssiau, F., Creţu, A.M., De Montjoye, Y.A.: Anonymization: The imperfect science of using data while preserving privacy. Science Advances 10(29), eadn7053 (2024)

45. Olatunji, I. E., Rauch, J., Katzensteiner, M., & Khosla, M. (2024). A review of anonymization for healthcare data. Big data, 12(6), 538-555.
46. Tabiri Aning, Edmond and Agnihotri, Nishant, Data Security: A Study of Data Anonymization And Several Techniques (March 13, 2024). Available at SSRN: https://ssrn.com/abstract=4757543 or http://dx.doi.org/10.2139/ssrn.4757543
47. Personal Data Protection Act 2012 - Singapore Statutes Online. (n.d.). Singapore Statutes Online. Retrieved July 24, 2025, from https://sso.agc.gov.sg/Act/PDPA2012
48. Article 1226 of the Civil Code of the People's Republic of China
49. National Health Commission. Notice on Issuing the Measures for the Administration of Cybersecurity in Medical and Health Institutions [EB/OL]. http://www.nhc.gov.cn/guihuaxxs/s10743/202208/50e2ef41b7554ae894053bcac32b79f0.shtml, 2022-8-8.
50. Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD), FDA, 2025
51. Cybersecurity Act 2018 - Singapore Statutes Online. (n.d.). Singapore Statutes Online. Retrieved July 24, 2025, from https://sso.agc.gov.sg/Acts-Supp/9-2018/
52. Hine, Emmie and Novelli, Claudio and Taddeo, Mariarosaria and Floridi, Luciano, Supporting Trustworthy AI Through Machine Unlearning (November 24, 2023). Science and Engineering Ethics, volume 30, issue 5, 2024[10.1007/s11948-024-00500-5], Available at SSRN: https://ssrn.com/abstract=4643518 or http://dx.doi.org/10.1007/s11948-024-00500-5
53. (OCR), O. for C.R. (2023) Cloud computing, HHS.gov. Available at: https://www.hhs.gov/hipaa/for-professionals/special-topics/health-information-technology/cloud-computing/index.html (Accessed: 21 July 2025).
54. C. Gentry and S. Halevi, "Implementing gentry's fully-homomorphic encryption scheme," in Annual international conference on the theory and applications of cryptographic techniques. Springer, 2011, pp. 129– 148.
55. J. H. Cheon, A. Kim, M. Kim, and Y. Song, "Homomorphic encryption for arithmetic of approximate numbers," in Advances in Cryptology– ASIACRYPT 2017: 23rd International Conference on the Theory and Applications of Cryptology and Information Security, Hong Kong, China, December 3-7, 2017, Proceedings, Part I 23. Springer, 2017, pp. 409– 437.
56. Secure Multiparty generative AI. (n.d.). https://arxiv.org/html/2409.19120v1
57. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., ... & Seth, K. (2016). Practical secure aggregation for federated learning on user-held data. arXiv preprint arXiv:1611.04482.
58. Bayram, F., Ahmed, B. S., & Kassler, A. (2022). From concept drift to model degradation: An overview on performance-aware drift detectors. Knowledge-Based Systems, 245, 108632.
59. E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. Lora: Low-rank adaptation of large language models. In International Conference on Learning Representations, 2022.
60. J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems, 2022.
61. S. Chan, A. Santoro, A. Lampinen, J. Wang, A. Singh, P. Richemond, J. McClelland, and F. Hill. Data distributional properties drive emergent in-context learning in transformers. Advances in Neural Information Processing Systems, 35:18878–18891, 2022.
62. Kaissis, G. A., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence, 2(6), 305–311. doi:10.1038/s42256-020-0186-1
63. How We Can Help. (2024, July 11). IMDA. Retrieved July 24, 2025, from https://www.imda.gov.sg/how-we-can-help?
64. IMDA. (n.d.). Emerging Technologies. IT Standards and Frameworks. https://www.imda.gov.sg/regulations-and-licensing-listing/ict-standards-and-quality-of-service/it-standards-and-frameworks/emerging-technologies

65. IMDA. (n.d.). About the Trusted Data Sharing Framework. IMDA. Retrieved July 24, 2025, from https://www.imda.gov.sg/how-we-can-help/data-innovation/trusted-data-sharing-framework
66. Personal Data Protection Act 2012 - Singapore Statutes Online. (n.d.). Singapore Statutes Online. Retrieved July 24, 2025, from https://sso.agc.gov.sg/Act/PDPA2012?ProvIds=P110-#pr51-
67. CNN. (2025, July 15). US Department of Defense awards contracts to Google, Musk's xAI. CNN. Retrieved July 24, 2025, from https://edition.cnn.com/2025/07/15/business/us-department-defense-google-musk-xai
68. Sharma N et al. Retrospective large-scale evaluation of an AI system as an independent reader for double reading in breast cancer screening. doi: 10.1101/2021.02.26.21252537.
69. Singh H, Meyer AN, Thomas EJ. The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations. BMJ Qual Saf. 2014 Sep;23(9):727-31. doi: 10.1136/bmjqs-2013-002627. Epub 2014 Apr 17. PMID: 24742777; PMCID: PMC4145460.
70. Melanoma: Statistics. http://www.cancer.net/about-us/cancernet-editorial-board.
71. Esteva A et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118 (2017).
72. "Understanding Complex Biology through AI | Owkin." www.owkin.com, www.owkin.com.
73. Heyndrickx, W., Mervin, L., Morawietz, T., Sturm, N., Friedrich, L., Zalewski, A., Pentina, A., Humbeck, L., Oldenhof, M., Niwayama, R., Schmidtke, P., Fechner, N., Simm, J., Arany, A., Drizard, N., Jabal, R., Afanasyeva, A., Loeb, R., Verma, S., . . . Ceulemans, H. (2023). MELLODDY: Cross-pharma Federated Learning at Unprecedented Scale Unlocks Benefits in QSAR without Compromising Proprietary Information. Journal of Chemical Information and Modeling, 64(7), 2331–2344. https://doi.org/10.1021/acs.jcim.3c00799