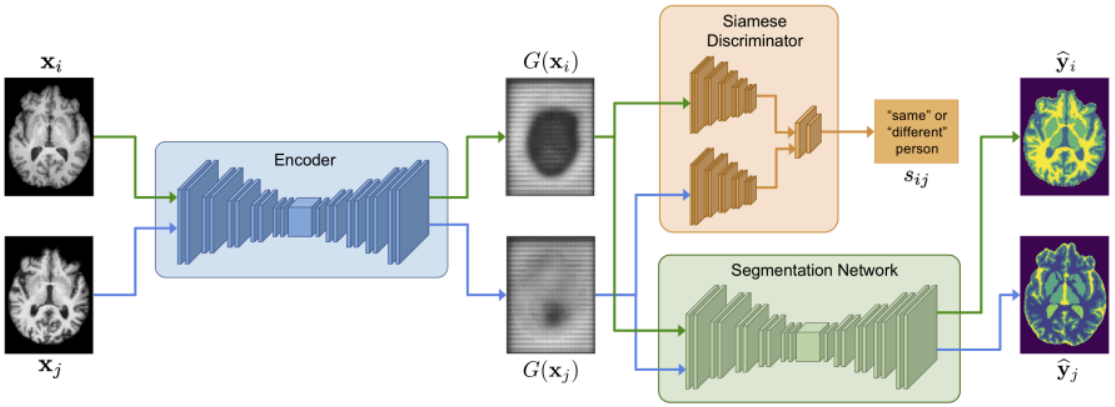
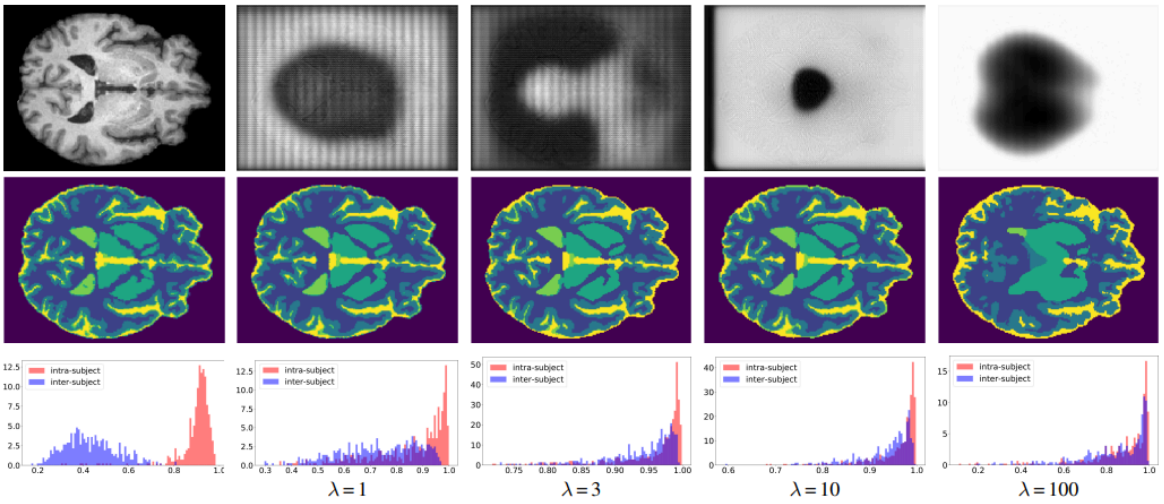
1.https://arxiv.org/pdf/1909.04087
Якщо брати саму логіку вирішення проблеми,то у нас дуже схоже,певно навіть індентично,є дві моделі:
1.**Encoder** - ховає ідентичність пацієнта,накладає шум;
2. **Discriminator** - вона намагається вгадати ідентичність з наявних даних;
3.**Medical Analysis Network** - намагається поставити діагноз (або виконати сегментацію) на основі даних від Encoder.



a) Training configuration

Мережа (Segmentation Network) аналізує патерни у вхідних даних. Вона навчена, що навіть якщо форма черепа прихована, текстура тканин мозку залишилася незмінною. Вона класифікує кожен піксель зображення як здорова тканина, патологія або фон.



Ряд 1 (Input): Оригінальний знімок МРТ, де чітко видно структуру голови.
Ряд 2 (Encoded): Те, що бачить мережа медичного аналізу. Це зображення є деформованим,для конфіденційності інформації.
Ряд 3 (Prediction/Ground Truth): Результат роботи мережі медичного аналізу. Це чорно-біла маска, де білим кольором виділено потрібну зону мозку.

Клієнтська сторона: На машині користувача запускається кодувальник. Він обробляє необроблені дані (МРТ) локально, щоб видалити конфіденційні ідентифікаційні дані,

перш ніж вони залишать пристрій Серверна сторона: Хмарний сервер отримує лише безпечні дані. Він запускає мережу сегментації та повертає результат Це моделює безпечну архітектуру,де сервер виконує складну логіку, при цьому не маючи доступу до читання приватних необроблених даних Обробка 3D-об'ємів: Система обробляє 3D-об'єми MPT (H x W x D). Оскільки ці 3D-масиви занадто великі для пам'яті (VRAM), автори реалізують стратегію на основі латок.Вони розрізають тривимірний об'єм на менші підкуби розміром.Під час тестування об'єм розділяється на рівномірно розташовані ділянки, обробляється окремо, а потім математично зшивається знову разом для реконструкції повного 3D-об'єкта Ефективність: крок кодування займає 0,08 секунди, сегментація 0,1 секунди

TABLE II

INTRA-SUBJECT AND INTER-SUBJECT PREDICTION ACCURACY ON TEST EXAMPLES OBTAINED BY THRESHOLDING MS-SSIM SCORES, USING THE ADVERSARIAL DISCRIMINATOR ($D_{\mathbf{adv}}$), OR TRAINING A SEPARATE DISCRIMINATOR ON THE ENCODED IMAGE ($D_{\mathbf{new}}$). THE MAP COLUMN IS THE MEAN AVERAGE PRECISION OF A TOP-k RETRIEVAL ANALYSIS USING THE SIAMESE DISCRIMINATOR'S EMBEDDING AS REPRESENTATION. RESULTS ARE REPORTED FOR NON-ENCODED IMAGES OR ENCODED IMAGES FOR DIFFERENT λ VALUES.

		Accuracy			mAP
		MS-SSIM	D_{adv}	D_{new}	
Non-encoded		0.970	—	0.953	0.850
Encoded	$\lambda = 1$	0.564	0.520	0.598	0.189
	$\lambda = 3$	0.533	0.537	0.615	0.152
	$\lambda = 10$	0.510	0.523	0.577	0.141
	$\lambda = 100$	0.503	0.516	0.513	0.087

Non-encoded (не закодовано) - це звичайні, оригінальні знімки.
Значення(D new == Accuracy) **0.953**(95.3%**)** - якщо взяти звичайний знімок, система вгадує пацієнта у 95% випадків. Це погано для приватності.
Encoded (Закодовано) - це робота їхньої моделі Privacy-Net.
Значення(D new == Accuracy) падає до **0.513** (51.3%).
Система вгадує пацієнта лише у **51%** випадків. Оскільки варіантів завжди мінімум два (це та людина або ні), то 50% це сліпе вгадування.

TABLE III

SEGMENTATION DICE SCORE ON THE PPMI TEST SET FOR DIFFERENT VALUES OF λ. NON-ENC REFERS TO THE MODEL TRAINED WITH NON-ENCODED IMAGES.

	GM	WM	Nuclei	CSF int.	CSF ext.	Overall
Non-enc	0.941	0.853	0.657	0.665	0.825	0.848
$\lambda = 1$	0.925	0.824	0.580	0.598	0.752	0.812
$\lambda = 3$	0.899	0.793	0.549	0.550	0.693	0.778
$\lambda = 10$	0.881	0.796	0.555	0.531	0.685	0.771
$\lambda = 100$	0.847	0.692	0.454	0.405	0.513	0.684

Overall(F1)**:**
Non-enc (Звичайна, не захищена мережа): Точність **0.848** (84.8%).

lambda = 1 (Захищена мережа, м'який захист): Точність **0.812** (81.2%).
Висновок*:* Втрата якості зовсім невелика (лише 3.6%), але дані вже захищено.

lambda = 100(Максимальний захист): Точність **0.684** (68.4%).

Висновок: При дуже сильному викривленні точність діагнозу падає суттєвіше.

Для оцінки якості роботи мережі медичного аналізу (Segmentor) у статті використовується метрика **Dice Score** (Table III). У задачах сегментації ця метрика є еквівалентом **F1 Score**, оскільки вона об’єднує в собі точність (Precision) та повноту (Recall).
Значення **F1 Score (Dice)** для захищеної моделі становить **0.812** (lambda = 1).

Окремо оцінюється **Accuracy (точність ідентифікації)**, але в контексті приватності (Table II, метрика D new), де її зниження з **0.953** до **0.598** свідчить про успішність захисту даних.

2. https://dl.acm.org/doi/pdf/10.1145/3702638
Є три види таких атак:
1.Біла скринька (повна інформація): Зловмисник має вихідний код. Він може бачити ваги моделі, архітектуру та градієнти
FGSM (Fast Gradient Sign Method): Швидка атака, що додає шум у напрямку градієнта функції втрат.
PGD (Projected Gradient Descent): Ітеративна версія FGSM, потужніша і вважається стандартом для тестування.
C&W (Carlini & Wagner): Оптимізаційна атака, яка створює дуже непомітні зміни, але потребує більше часу.
DeepFool: Атака, що шукає найкоротший шлях до зміни класифікації.

2.Чорна скринька (Oracle Access): Зловмисник діє як користувач веб-API. Він надсилає вхідні дані та отримує вихідні (прогноз), але не може бачити внутрішній код або градієнти. Він повинен використовувати оптимізацію нульового порядку, щоб вгадати градієнт на основі вихідних балів
Query-based: Роблять багато запитів до моделі, щоб підібрати шум.
Transfer-based: Генерують атаку на одній моделі й переносять її на іншу.

3.Без рамки (переносність): Зловмисник створює приклад атаки на власній локальній моделі та сподівається, що він перенесеться на цільову систему. Це спирається на логіку, що різні мережі часто вивчають схожі слабкі місця у функціях

Логіка захисту ідентична логіці першого дослідження:захисник намагається мінімізувати помилку, тоді як атакуючий одночасно максимізує помилку. Це змушує модель вивчати плавну межу прийняття рішень, яку важко порушити
Таке навчання є обчислювально ресурсоємним, оскільки воно навчає мережу в жорсткому режимі
Методи захисту:
Adversarial Training (Змагальне тренування): Найефективніший метод. У тренувальний набір даних додають зашумлені зображення, щоб модель вчилася їх розпізнавати.
Preprocessing (Попередня обробка / Denoising): Очищення зображення від шуму перед подачею в нейромережу.
Detection (Виявлення): Створення окремого модуля, який визначає, чи є зображення справжнім, чи атакованим, і відхиляє підозрілі.

Стандартне навчання (NAT) займає приблизно 16 секунд,а робустне навчання (PGD-AT) триває приблизно секунди
Для сегментації різниця більша (13,8 секунди проти 148,5 секунди)

Вони виклали свій код на **GitHub.** Силка на репозиторій - https://github.com/tomvii/Adv_MIA
Там є все,тренування,метрики,їхня модель.

3.https://www.researchgate.net/profile/Miaomiao-Zhang-9/publication/356206421_Defending_Medical_Image_Diagnostics_Against_Privacy_Attacks_Using_Generative_Methods_Application_to_Retinal_Diagnostics/links/67bd606e207c0c20fa95c6fc/Defending-Medical-Image-Diagnostics-Against-Privacy-Attacks-Using-Generative-Methods-Application-to-Retinal-Diagnostics.pdf

Атака: У статті розглядаються **Privacy Attacks** (атаки на конфіденційність), а конкретно — **Membership Inference Attacks (MIA)**.

> Суть атаки: Зловмисник намагається дізнатися, чи використовувалися дані конкретного пацієнта (наприклад, зображення сітківки) для тренування моделі. Якщо атака успішна, це порушує медичну таємницю.

Автори тестують різні сценарії, зокрема **Graybox loss-threshold attack** (атака, що базується на аналізі функції втрат моделі).

Захист: Основний метод захисту — використання **Generative Methods** (GANs — генеративних змагальних мереж).

Метод*:* Замість того щоб давати дослідникам реальні дані пацієнтів, лікарня (data sourcer) тренує GAN, щоб створити **синтетичний набір даних (**proxy dataset).

Цей синтетичний набір передається розробникам моделі. Він зберігає статистичні властивості хвороби, але не містить реальних зображень людей.

Для покращення результатів вони використовують техніки **MMD** (Maximum Mean Discrepancy**)** та **Mixup**.

Table 1: Task and Attack Accuracies (%) of various defenses with different percentages of synthetic data. 95% CIs are in parenthesises, and numbers in bold are the best.

	Raw Data Access %	Defense	$Acc_{Task,D}$	$Acc_{Attack,D}$			
				Blackbox		Graybox	
				Loss-Thre	Label-Only	Loss-Thre	Label-Only
Synthetic Data Only	0%	No Defense	68.77 (0.91)	49.84 (0.69)	49.90 (0.69)	**49.85 (0.69)**	49.95 (0.69)
		MMD+Mixup	73.30 (0.87)	**49.67 (0.69)**	**49.79 (0.69)**	49.99 (0.69)	**49.79 (0.69)**
Synthetic/ Real Mixture	25%	No Defense	73.54 (0.86)	53.35 (0.69)	53.10 (0.69)	54.98 (0.69)	53.10 (0.69)
		MMD+Mixup	74.10 (0.86)	52.65 (0.69)	52.17 (0.69)	50.05 (0.69)	52.17 (0.69)
	50%	No Defense	72.95 (0.87)	57.49 (0.69)	56.70 (0.69)	60.31 (0.68)	56.70 (0.69)
		MMD+Mixup	73.80 (0.86)	51.91 (0.69)	52.86 (0.69)	50.08 (0.69)	52.86 (0.69)
	75%	No Defense	75.14 (0.85)	60.43 (0.68)	59.21 (0.68)	66.03 (0.66)	59.21 (0.68)
		MMD+Mixup	74.75 (0.85)	54.52 (0.69)	55.41 (0.69)	50.01 (0.69)	55.41 (0.69)
Real Data Only	100%	No Defense (Baseline)	73.24 (0.87)	64.25 (0.66)	62.70 (0.67)	66.64 (0.65)	62.70 (0.67)
		MMD+Mixup	**75.52 (0.84)**	61.80 (0.67)	59.23 (0.68)	50.08 (0.69)	59.23 (0.68)
		Memguard	73.24 (0.87)	63.87 (0.67)	62.70 (0.67)	63.87 (0.67)	62.70 (0.67)

Діагностика не страждає (Task Accuracy): Зазвичай методи захисту приватності знижують точність моделі. Однак тут метод **MMD+Mixup** навіть *покращив* точність діагностики: на реальних даних (100% Real Data) він показав точність **75.52%**, тоді як модель без захисту (No Defense) мала лише **73.24%**. Це означає, що модель стала краще виявляти хворобу.

Атака стає марною (Attack Accuracy): Ефективність атаки зловмисника (Graybox Loss-Thre) на незахищену модель становила **66.64%** (зловмисник міг вгадувати дані пацієнтів значно краще за випадковість). Після застосування захисту **MMD+Mixup** успішність атаки впала до **50.08%**.

Показник 50% означає сліпе вгадування. Тобто захист повністю нівелював здатність зловмисника розрізняти дані пацієнтів.

Table 2: $P1(D)_{Attack} \in [0, 100]$ where larger is better for various settings and defenses.

	Raw Data Access %	Defense	$P1(D)_{Attack}$			
			Blackbox		Graybox	
			Loss-Thre	Label-Only	Loss-Thre	Label-Only
Synthetic Data Only	**0%**	**No Defense**	58.01 (0.63)	57.97 (0.63)	58.00 (0.63)	57.94 (0.63)
		MMD+Mixup	**59.68 (0.62)**	**59.60 (0.62)**	59.46 (0.62)	**59.60 (0.62)**
Synthetic/ Real Mixture	**25%**	**No Defense**	57.09 (0.61)	57.27 (0.61)	55.85 (0.61)	57.27 (0.61)
		MMD+Mixup	57.78 (0.61)	58.14 (0.61)	59.67 (0.61)	58.14 (0.61)
	50%	**No Defense**	53.72 (0.61)	54.34 (0.61)	51.41 (0.60)	54.34 (0.61)
		MMD+Mixup	58.23 (0.61)	57.53 (0.61)	59.56 (0.61)	57.53 (0.61)
	75%	**No Defense**	51.84 (0.60)	52.88 (0.60)	46.79 (0.58)	52.88 (0.60)
		MMD+Mixup	56.55 (0.61)	55.86 (0.61)	59.91 (0.61)	55.86 (0.61)
Real Data Only	**100%**	**No Defense (Baseline)**	48.05 (0.59)	49.43 (0.60)	45.84 (0.59)	49.43 (0.60)
		MMD+Mixup	50.74 (0.59)	52.95 (0.60)	**60.11 (0.61)**	52.95 (0.60)
		Memguard	48.39 (0.59)	49.43 (0.60)	48.39 (0.59)	49.43 (0.60)

Ця таблиця використовує метрику **P1(D)** яка є сумарною оцінкою.

Беззаперечна перевага MMD+Mixup: У найскладнішому сценарії (Real Data Only, 100%), загальна оцінка ефективності для методу **MMD+Mixup** становить **60.11** балів. Для порівняння, звичайна модель без захисту (No Defense) отримала лише **45.84** бали.

Отже,використання генеративних методів (MMD+Mixup) є найкращою стратегією,воно дає найвищий рівень безпеки, зберігаючи при цьому найвищу точність медичного діагнозу.

4. https://arxiv.org/pdf/2310.06227

Дослідження зосереджено на **Evasion Attacks (Атаки ухилення)** та їхній **Transferability (Переносимості)** у середовищі федеративного навчання (Federated Learning - FL).

Тип атаки: Атакуючий (шкідливий клієнт) використовує оновлення глобальної моделі, щоб генерувати змагальні приклади (adversarial examples), які потім можуть обманути моделі інших клієнтів.

У статті використовуються класичні градієнтні методи для створення атак:

FGSM (Fast Gradient Sign Method): Швидка атака для генерації шуму.

PGD (Projected Gradient Descent): Потужніша ітеративна атака.

Автори показують, що зловмисник може використовувати **градієнти попередніх глобальних моделей** (gradient information from prior global updates), щоб покращити атаку та її переносимість на інших клієнтів, навіть не маючи доступу до їхніх даних.

Автори оцінюють вразливість FL-мереж і зазначають, що поточні методи (наприклад, **Adversarial Training**, **Homomorphic Encryption** або **Secure Aggregation**) часто є недостатньо ефективними проти запропонованих сценаріїв або занадто ресурсомісткий для медичних систем.

Головний висновок щодо захисту: існує термінова потреба у перегляді протоколів безпеки, оскільки специфічні конфігурації доменів (domain-specific configurations) у медицині значно підвищують успішність атак.

TABLE I: Comparison of iterative models and their computational efficiency, on performing computations on one batch of data. ACC shows average client performance for unperturbed test data. AASR is average ASR on all clients.

Dataset	Attack type	ACC	AASR	time (sec)
Meningioma	PGD-20	84.12%	27.52%	3.423
	PGD-40		32.99%	6.794
Pathology	PGD-20	77.01%	60.98%	5.464
	PGD-40		82.27%	10.843
Glioma	PGD-20	61.84%	51.83%	3.420
	PGD-40		63.77%	6.793

Dataset (Набір даних): Тип медичних зображень, на яких тренувалася модель (Meningioma — пухлина мозку, Pathology — гістопатологія, Glioma — гліома).

Attack type (Тип атаки): Використовується атака **PGD** (Projected Gradient Descent).

PGD-20: Атака з 20 ітераціями (кроками).

PGD-40: Атака з 40 ітераціями (потужніша, але довша).

Збільшення потужності атаки (кількості ітерацій) підвищує її успішність (**AASR**), але пропорційно збільшує час обчислень**.** Для захисту це означає, що моделі на різних типах медичних даних мають різний рівень природної стійкості (Pathology захищати найважче).

Dataset	FGSM	PGD
Meningioma	84.80%	83.73%
Glioma	82.32%	74.89%
Pathology	90.67%	79.86%

FGSM (Fast Gradient Sign Method): Швидка атака, яка робить лише один крок у напрямку помилки. Вона проста і генерує менш точний, грубий шум.

PGD (Projected Gradient Descent): Ітеративна атака, яка робить багато маленьких кроків, щоб знайти найвразливіше місце моделі. Зазвичай у машинному навчанні PGD вважається значно сильнішою атакою.

Pathology: FGSM досягає **90.67%**, тоді як PGD — лише **79.86%**.

Glioma: FGSM **82.32%** значно ефективніша за PGD **74.89%**.

Meningioma: FGSM **84.80%** трохи краща за PGD **83.73%**.

Таблиця демонструє, що для атак у федеративній мережі (де зловмисник не знає точних моделей жертв) **простіша атака (FGSM) може бути більш небезпечною за складну (PGD)**, оскільки вона має вищу здатність до перенесення (transferability), особливо на вразливих даних типу Pathology (успішність понад 90%).