Abschlussbericht PrivacyUmbrella

Privatheit von Daten sicherstellen durch Umfassende Bereitstellung von Anonymisierungsverfahren

Förderkennzeichen	16KISA001K
Laufzeit/ Fördersumme	15.12.2022 – 14.12.2025 / 579.989,02 Euro
Leitung	Prof. Dr. Lena Wiese Fraunhofer Institut für Toxikologie und Experimentelle Medizin ITEM Leiterin Arbeitsgruppe Bioinformatik Nikolai-Fuchs-Str. 1, 30625 Hannover Tel. +49 511 5350-303 lena.wiese@item.fraunhofer.de https://www.item.fraunhofer.de/bioinformatics
Projektmitarbeiter	MSc. Mohan Xu / Pronaya Prosun Das
Kooperationspartner	• Universitätsmedizin der Johannes Gutenberg-Universität Mainz, Institut für Med. Biometrie, Epidemiologie und Informatik (IMBEI) • Goethe-Universität Frankfurt (GUF), Institut für Medizininformatik • MCS Data Labs GmbH (MCS)
Datum	21. April 2026

Fraunhofer ITEM

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Inhaltsverzeichnis

1 Zusammenfassung **2**

2 Einleitung **3**

3 Erhebungs- und Auswertungsmethodik **3**
3.1 Realer HAR-Datensatz 3
3.2 Datengenerierung mittels Adversarial Learning 4

4 Durchführung, Arbeits- und Zeitplan **5**
4.1 Zielstellung 5
4.2 Arbeits-/Zeitplan 6
4.2.1 AP 4.1 Anonymisierungsregeln 6
4.2.2 AP 4.2 Optimierungsmodell 8
4.2.3 AP 4.3 ML-Resilienz 11

5 Ergebnisse **16**
5.1 Webanwendung 16
5.2 Neuronale Netze 16

6 Wichtigste Positionen des zahlenmäßigen Nachweises **17**

7 Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten **17**

8 Verwertbarkeit der Ergebnisse / Zukunftsprognose **17**

9 Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen **18**

10 Erfolgte Veröffentlichungen **19**
10.1 Publikationen 19
10.2 Verbreitung und Öffentlichkeitsarbeit der Projektergebnisse 19

Literatur **22**

1 Zusammenfassung

Bei der Erhebung personenbezogener Daten werden üblicherweise die eindeutigen Identifikatoren der einzelnen Nutzer entfernt, um ihre Identität zu verschleiern. Mit dieser Methode können jedoch Bedenken hinsichtlich des Schutzes der Privatsphäre nicht angemessen berücksichtigt werden, da die einfache Entfernung eindeutiger Kennungen nicht die vollständige Anonymität der Daten garantiert. Die Anwendung von Algorithmen des maschinellen Lernens auf anonymisierte Daten kann trotz ihres geänderten Zustands unbeabsichtigt sensible Informationen über Einzelpersonen innerhalb des Datensatzes offenbaren. In unserem Forschungsbeitrag wurde ein k-Anonymisierungsansatz vorgestellt, der den Simulated Annealing-Algorithmus nutzt, um die Herausforderung des Ausgleichs zwischen Datenschutz und Nutzen zu bewältigen, insbesondere im Zusammenhang mit der Klassifizierungsgenauigkeit.

Im Rahmen des Projektes wurde außerdem auf die Einbindung der neuentwickelten Technologien in existierende mobile Endgeräte des Industriepartners (MCS) abgezielt, wobei die beschränkten Rechenkapazitäten dieser Geräte berücksichtigt wird. Vor diesem Hintergrund war ein weiteres Ergebnis die Entwicklung von ressourceneffizienten Algorithmen für ein multitaskfähiges Federated-Learning-Framework mit integrierter Wissensdestillation, das speziell auf heterogene Geräteumgebungen und variierende Labelverteilungen ausgerichtet ist.

2 Einleitung

Das Projekt "PrivacyUmbrella" adressierte Herausforderungen im Bereich Datenschutz und Datennutzbarkeit, speziell im Kontext medizinischer und personenbezogener Daten. Im Mittelpunkt stand die Entwicklung eines umfassenden Anonymisierungssystems, das verschiedene etablierte Verfahren, wie k-Anonymität, Differential Privacy und Privacy-Preserving Record Linkage kombiniert und standardisierte Privacy-Metriken etabliert. Dabei wurden insbesondere technische Grundlagen geschaffen, um sensible medizinische Daten auch für weitergehende Analysen nutzbar zu machen, ohne die Privatsphäre einzelner Patientinnen und Patienten zu gefährden. Dazu wurde zum Einen ein Optimierungsmodell entwickelt, das unter Einsatz von Heuristiken wie "Simulated Annealing" die Auswahl geeigneter Generalisierungsgrade und weiterer Parameter automatisiert. Zusätzlich untersuchte das Projekt Verfahren, um anonymisierte Datensätze mittels Machine-Learning-Techniken hinsichtlich ihrer Anfälligkeit für mögliche Re-Identifikationsattacken zu evaluieren und die Resilienz gegen solche Angriffe zu verbessern. Im Rahmen des Projekts wurden zudem spezielle Herausforderungen im Umgang mit verschachtelten und Graph-basierten medizinischen Datenformaten adressiert. Insbesondere wurde auf Basis von HL7 FHIR-Ressourcen und der Open-Source-Registerplattform OSSE erarbeitet, wie komplex strukturierte Datensätze effektiv anonymisiert und gleichzeitig nutzbar bleiben können. Ein weiteres zentrales Ziel des Projekts war die Untersuchung der Anwendbarkeit von verteilten Anonymisierungsverfahren direkt auf mobilen Endgeräten und Wearables. Federated Learning wurde hier genutzt, um Machine-Learning-Modelle dezentral zu trainieren und so die Verarbeitung sensibler Daten direkt auf mobilen Geräten sicherzustellen. Die Entwicklungen wurden prototypisch umgesetzt und in der praktischen Erprobung weiter optimiert. Die Ergebnisse weisen nach, dass das Konzept erfolgreich den angestrebten Datenschutz gewährleistet, ohne die Nutzbarkeit der Daten für analytische Zwecke maßgeblich einzuschränken. Die Forschungsergebnisse legen den Grundstein für weitere Verbesserungen und die Nutzung des entwickelten Systems in vielfältigen medizinischen sowie anderen datenintensiven Anwendungsszenarien.

3 Erhebungs- und Auswertungsmethodik

Das Gesamtziel des geplanten Projekts ist es, das Potenzial der Anonymisierung für eine effizientere Gesundheitsversorgung zu nutzen; es soll erreicht werden durch Entwicklung und Validierung eines Optimierungsmodells, um eine optimale Kombination von Einzelverfahren zu erhalten. Um die besonderen Schutzbelange von medizinischen personenbezogenen Daten zu berücksichtigen, besteht der im Projekt verfolgte Ansatz in der Entwicklung hoch-spezialisierter Anonymisierungslösungen vor dem Hintergrund moderner Daten-Analyse für medizinische Anwendungsfälle und mit der Möglichkeit der Anbindung an die Medizininformatik-Initiative. Vor diesem Hintergrund wurden Daten im Bereich Human-Activity-Recognition (HAR) neu erhoben und Datensätze und Systeme im medizinischen Bereich den Auswertungen zugrunde gelegt.

3.1 Realer HAR-Datensatz

Ein unter realen Bedingungen generierter und von den Probanden kategorisierter Smart-Health-Datensatz für körperliche Aktivitäten (Radfahren, Laufen, Wandern, Pilates, Sitzen) wurde erhoben mittels der Smartwatches, die von MCS bereitgestellt wurden. Der insgesamt 115h umfassende Datensatz wurde über einen Zeitraum von vier Monaten erfasst und beinhaltet multimodale Aktivitätsaufnahmen von bis zu 9 Sensoren (u.a. Herzrate, Blutsauerstoffsättigung, EKG, PPG, Hauttemperatur, usw.).

Dieser Datensatz soll mit anonymisierten Rohdaten als auch in Form einer heruntegerechneter Version (1 Hz-Rate) mit aus den Rohdaten abgeleiteten weiteren Merkmalen öffentlich zur freien Verfügung bereitgestellt werden. Die Anomyisierung erfolgte durch eine Pseudonymisierung der im Datensatz inhaltenen Individuen und der zufälligen Verschiebung der zugeordnenen Zeitmarken. In Folgeprojekten soll der Datensatz erweitert, veröffentlicht werden und Anonymisierungstechniken darauf angewendet werden.

Copyright: MCS Datalabs

3.2 Datengenerierung mittels Adversarial Learning

Die in unserer Arbeit entwickelte Methodik [1] befasst sich mit der Integration von Rényi-Differential Privacy (RDP) mit einem Generative Adversarial Network (GAN), das durch einen Variational Autoencoder (VAE) erweitert wurde, was zu unserer RDP-VAEGAN-Architektur führt (siehe Abb. 1).

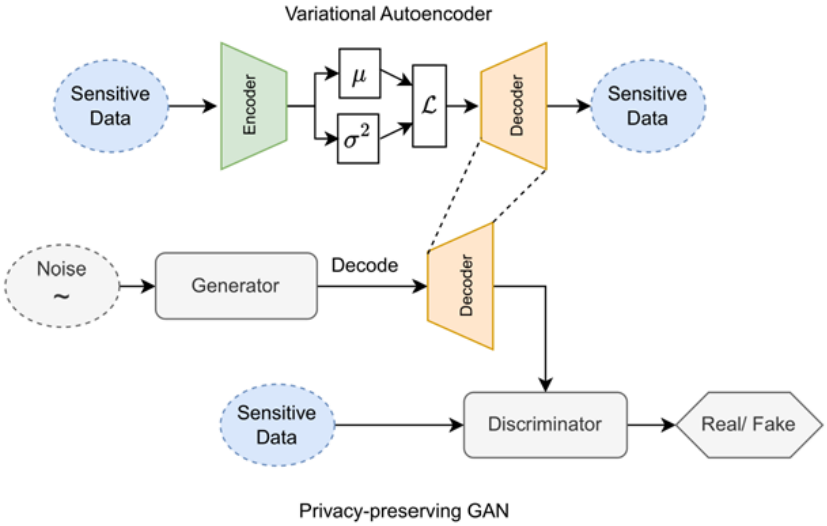


Abbildung 1: Block-Diagramm GAN framework

Dieses anspruchsvolle Setup ist darauf ausgelegt, qualitativ hochwertige synthetische Daten zu generieren, die die statistischen Eigenschaften der Originaldaten beibehalten und gleichzeitig robuste Datenschutzgarantien sicherstellen können. Die Architektur kombiniert die Fähigkeiten von GANs und VAEs zur Generierung synthetischer Daten. Das GAN ist dafür verantwortlich, die Verteilung der Eingangsdaten zu erlernen und neue Dateninstanzen zu erzeugen, die diese Verteilung nachahmen. Die VAE-Komponente hilft dabei, die Eingangsdaten in einen kontinuierlichen latenten Raum zu transformieren, was die Handhabung gemischter Datentypen (kontinuierlich und diskret), wie sie häufig in medizinischen Datensätzen vorkommen, erleichtert. Die GAN-Architektur besteht aus zwei Komponenten:

- **Generator (G)**: Nimmt zufälliges Rauschen z auf und gibt synthetische Daten $\hat{x}$ aus. Er versucht, Daten zu erstellen, die von realen Daten nicht zu unterscheiden sind.
- **Diskriminator (D)**: Erhält entweder echte Daten x oder synthetische Daten vom Generator und versucht, zwischen den beiden zu unterscheiden. Der Diskriminator gibt eine Wahrscheinlichkeit aus, die angibt, ob die Eingabedaten echt oder synthetisch sind.

Das VAE wird danach integriert, um die Daten in einen kontinuierlichen latenten Raum zu überführen, was den Generierungsprozess durch das GAN erleichtert, insbesondere bei der Verarbeitung von diskreten und gemischten Datentypen. Das VAE besteht aus:

- **Encoder**: Wandelt echte Daten in eine latente Repräsentation um.
- **Decoder**: Rekonstruiert Daten aus dem latenten Raum, die im Trainingsprozess des Generators verwendet werden.

4 Durchführung, Arbeits- und Zeitplan

4.1 Zielstellung

Die folgende Auflistung fasst die in diesem Teilvorhaben angestrebten innovativen Elemente zusammen:

- Zielstellung des Projekts war es, den Schutz sensibler Informationen in relationalen Tabellen durch die Weiterentwicklung bestehender Anonymisierungsverfahren zu verbessern. Zu diesem Zweck wurde ein vorhandenes k-Anonymitätsmodell vorgestellt. Dieses Modell nutzte Generalisierungs- und Suppressionstechniken zur Erzeugung eines Lösungsraums und setzte Simulated Annealing als probabilistisches Optimierungsverfahren ein, um innerhalb dieses Lösungsraums eine annähernd optimale Anonymisierungslösung zu identifizieren. Im Rahmen des Projekts wurde dieses Modell um das Konzept der t-Closeness als zusätzlichen Verfeinerungsmechanismus ergänzt. Ziel dieser Erweiterung war es, das Risiko der Attributoffenlegung weiter zu reduzieren und zu untersuchen, wie sich die Integration von t-Closeness-Beschränkungen in das bestehende k-Anonymitätsframework auf die Gesamtleistung sowie die Effektivität des Anonymisierungsprozesses auswirkte.
- Unser wesentliches Ziel bei der Herleitung von synthetischen Datensätzen war die Wahrung des Datenschutzes während des Trainings des GAN und VAE. Dies wurde durch die Anwendung von Rényi-Differential Privacy erreicht. Unsere Arbeit zeigt, dass durch das Training des Diskriminators mit differential privacy und die Verwendung des VAE zur Vorverarbeitung der Eingabe für den Generator das gesamte System sicherstellt, dass keine sensiblen Informationen aus den synthetischen Daten abgeleitet werden können.
- Ziel des Projekts im Bereich föderiertes Anonymisieren war die Entwicklung und Evaluation eines Lernverfahrens, das den praktischen Herausforderungen verteilter Trainingsumgebungen gerecht wird. Föderiertes Lernen unterstützte mehrere Nutzer dabei, gemeinsam ein globales Modell auf Basis ihrer jeweils lokalen, personenbezogenen Datensätze zu trainieren. In der praktischen Umsetzung zeigte sich jedoch, dass die Leistungsfähigkeit durch die Heterogenität der Clients – insbesondere hinsichtlich Rechenkapazität, Kommunikationsbedingungen und individueller Anforderungen – sowie durch Datenschutzrisiken beim Austausch von Modellparametern eingeschränkt wurde. Vor diesem Hintergrund wurde im Rahmen des Projekts ein Multi-Task-Federated-Learning-Framework auf Basis von Wissensdistillation (MFedKD) konzipiert und implementiert. Ziel war es, Geräte- und Label-Heterogenität effektiv zu adressieren und zugleich die Übertragung sensibler Modellparameter zu reduzieren.

  Die Leistungsfähigkeit des entwickelten Ansatzes wurde anhand des ISIC2019-Dermoskopie-Datensatzes evaluiert. Die experimentellen Ergebnisse zeigten, dass MFedKD eine konkurrenzfähige Modellgüte erreichte und gleichzeitig den Rechen- und Kommunikationsaufwand für ressourcenbeschränkte Clients signifikant verringerte. Darüber hinaus bestätigten Rekonstruktionsexperimente auf Basis der Gradienteninversion die Wirksamkeit des Ansatzes im Hinblick auf den Schutz der Privatsphäre der Nutzer.

4.2 Arbeits-/Zeitplan

Die Beschreibung des aktuellen ITEM-Arbeitsstands im Vergleich zum Plan lautet wie folgt:

4.2.1 AP 4.1 Anonymisierungsregeln

AP 4.1.1 Erweiterung der einzelnen Anonymisierungsverfahren. Das übergeordnete Ziel der Datenanonymisierung besteht darin, die Zuordnung von Datensätzen zu den ursprünglichen Personen zu verhindern und somit deren Privatsphäre zu schützen sowie Datenmissbrauch und Datenschutzverletzungen vorzubeugen. Die k-Anonymität erfreut sich aufgrund ihrer Einfachheit und Effektivität beim Schutz vor Re-Identifizierung großer Anerkennung. Die vorliegende Arbeit entwickelt einen Optimierungsalgorithmus, der mithilfe der Simulated-Annealing-Heuristik automatisch ein geeignetes Generalisierungsniveau zur Erreichung der k-Anonymität auswählt. Der Algorithmus umfasst folgende Hauptschritte:

(1) **Datenvorbereitung und Initialisierung** Nach dem Import des Datensatzes erfolgen die Anpassung der Klassenverteilung mittels Imputation sowie die Entfernung von Ausreißern. Die ausgewählten Quasi-Identifikatoren (QIDs) dienen zur Konstruktion der Domänengeneralisierungshierarchie (DGH). Anschließend werden Einschränkungen definiert, darunter der maximale Unterdrückungsanteil Suppmax sowie der Parameter k. Der Algorithmus startet mit einer zufällig generierten Initiallösung S_0, welche eine Kombination der Generalisierungsstufen aller Quasi-Identifikatoren repräsentiert.

(2) **Temperatursteuerung** Die Temperatur wird initial auf T_0 gesetzt und in jeder Iteration i gemäß der Abkühlrate α exponentiell reduziert:

$$T_i = T_0 \cdot \alpha^i \tag{1}$$

Die Iteration wird fortgesetzt, bis eines der folgenden Abbruchkriterien erfüllt ist: Entweder wurde die maximale Anzahl aufeinanderfolgender Iterationen ohne Lösungsverbesserung erreicht, oder die Temperatur unterschreitet den definierten Schwellenwert ($T_i \leq T_{\text{min}}$).

(3) **Exploration** Im iterativen Durchlauf durchsucht der Algorithmus den Lösungsraum systematisch, indem ein Quasi-Identifikator-Attribut (QID) zufällig ausgewählt und dessen Generalisierungsgrad angepasst wird. Der Auswahlprozess lässt sich mathematisch wie folgt formalisieren: Gegeben sei eine Menge von Quasi-Identifikator-Attributen $A = \{a_1, a_2, \ldots, a_n\}$. Aus A wird ein Attribut a_i gewählt, wobei $1 \leq i \leq n$. Für a_i existiert eine Menge zulässiger Generalisierungsgrade $G(a_i) = \{g_1, g_2, \ldots, g_m\}$, aus der ein Grad g_j ausgewählt wird, wobei $1 \leq j \leq m$.

(4) **Iteration und Konvergenz** Innerhalb dieser iterativen Schleife reduziert der Algorithmus schrittweise die Temperatur und steuert dadurch die Suche im Lösungsraum. Mit sinkender Temperatur verschärfen sich die Akzeptanzkriterien, wodurch Verbesserungen der Zielfunktion zunehmend priorisiert werden. Dieser graduelle Übergang von Exploration zu Exploitation gewährleistet, dass der Algorithmus seine Suche sukzessive auf vielversprechende Regionen des Lösungsraums konzentriert und schließlich zu einer optimalen oder nahezu optimalen Lösung konvergiert.
Die Terminierung des Simulated-Annealing-Prozesses erfolgt, sobald entweder die maximale Iterationsanzahl erreicht wurde oder ein stationärer Zustand eintritt, in dem weitere Exploration keine signifikante Verbesserung der Bewertung mehr erzielt.

AP 4.1.2 Verfeinerung der Datenanonymisierung und Verbesserung der Analysierbarkeit der Daten. Angesichts der Herausforderung, bei der k-Anonymisierung ein ausgewogenes Verhältnis zwischen Privatsphäre und Datennutzbarkeit zu erzielen, präsentiert das Arbeitspaket einen Simulated-Annealing-Algorithmus, der die Klassifikationsgenauigkeit optimiert und sich zugleich dem globalen Optimum der Anonymisierung annähert.
Als Optimierungsziel dient der AUC-ROC-Wert, ein etabliertes Maß für die Klassifikationsgenauigkeit [2]. Bei der Evaluation einer Nachbarlösung erfolgt die Entscheidung über deren Akzeptanz anhand des Metropolis-Kriteriums, welches Änderungen basierend auf dem Zielfunktionswert und der aktuellen Temperatur probabilistisch akzeptiert oder verwirft. Die Akzeptanzwahrscheinlichkeit ist definiert als $p = e^{-\Delta c/T_i}$, wobei Δc die Änderung des Zielfunktionswertes und T_i die aktuelle Temperatur bezeichnet. Der resultierende Algorithmus erzielt ein optimales Gleichgewicht zwischen zwei zentralen Zielen: der Maximierung der Klassifikationsgenauigkeit und der strikten Einhaltung vordefinierter Nebenbedingungen.
Im Rahmen der Evaluation wird der vorgestellte Algorithmus mit ARX [3] anhand zweier realer Datensätze verglichen: dem UCI Adult Dataset [4] sowie dem MIMIC-III-Datensatz [5, 6]. Der Fokus liegt dabei auf Klassifikationsaufgaben. Durch den Vergleich der Klassifikationsergebnisse vor und nach der Anonymisierung lässt sich der Einfluss verschiedener Anonymisierungsverfahren auf die Modellgenauigkeit quantifizieren und somit die Qualität der anonymisierten Datensätze bewerten.

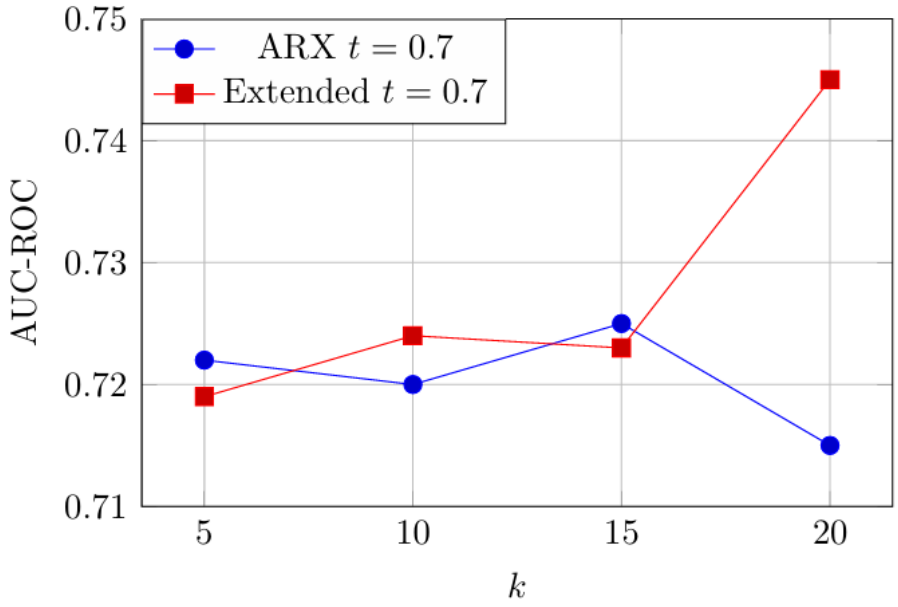


Abbildung 2: Vergleich der AUC-ROC-Verläufe zwischen ARX und erweitertem Modell für verschiedene k-Werte auf dem Adult-Datensatz mit einer maximal zulässigen Unterdrückung von 30%

Im Vergleich zu ARX ermöglicht die vorgestellte Methode die Identifikation geeigneter Kombinationen von Generalisierungsstufen, die sowohl die k-Anonymität des Datensatzes gewährleisten als auch akzeptable AUC-ROC-Werte bei nachgelagerten Klassifikationsaufgaben sicherstellen.

AP 4.1.3 Komplexe medizinische geschachtelte Daten und Graph-Daten. Das Arbeitspaket adressiert spezifische Herausforderungen im Umgang mit verschachtelten und graphenbasierten medizinischen Datenformaten. Zu diesem Zweck wurde eine Lösung auf Basis der HL7-FHIR-Ressource und der Open-Source-Registerplattform OSSE entwickelt, die demonstriert, wie komplexe strukturierte Datensätze effektiv anonymisiert werden können, ohne

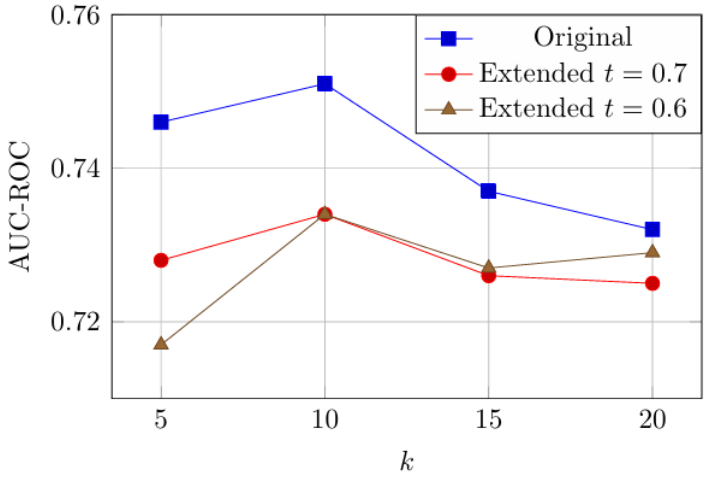
deren Nutzbarkeit zu beeinträchtigen.
Zur Validierung der Datenverarbeitung sowie der Verschlüsselungs- und Anonymisierungstechniken wurden synthetische Patientendaten gemäß dem HL7-FHIR-Standard [7] generiert. Diese synthetischen Daten ermöglichen eine uneingeschränkte Nutzung unter realen Bedingungen, ohne Patientenrechte zu tangieren. Eine Genehmigung aus ethischen oder datenschutzrechtlichen Gründen ist daher nicht erforderlich. Der HL7-FHIR-Standard gewährleistet ein einheitliches Datenformat und bildet somit eine solide Grundlage für die Datenintegration. Ergänzend wurde eine Methode zur Konvertierung von FHIR-Daten in das tabellarische CSV-Format entwickelt und validiert.
Die Ergebnisse belegen, dass das CSV-Format für die Implementierung der ausgewählten Anonymisierungstechniken besser geeignet ist als das hierarchische FHIR-Format. Innerhalb der FHIR-Datenstruktur wurden systematisch personenbezogene Datenfelder identifiziert, die potenzielle Angriffsvektoren für Re-Identifizierungsrisiken darstellen. Dabei erfolgte eine Unterscheidung zwischen direkten und indirekten Identifikatoren, um Anonymisierungsmaßnahmen gezielt applizieren zu können. Der verwendete Datensatz wurde zudem mit dem Kerndatensatz der Medical Informatics Initiative als Referenzmodell abgeglichen. Dies gewährleistete Datenkonsistenz und Datenqualität als zentrale Voraussetzung für effektive Datenanalyse- und Verarbeitungsprozesse. Gleichzeitig wurde eine hohe Interoperabilität zwischen verschiedenen Systemen und Anwendungen unterstützt.

4.2.2 AP 4.2 Optimierungsmodell

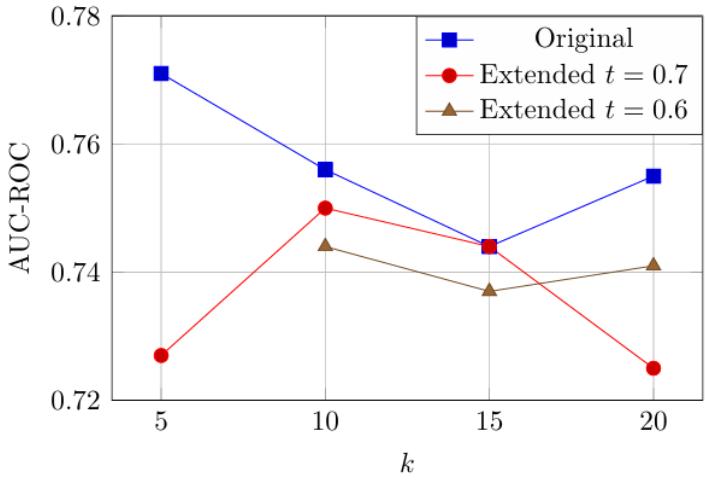
AP 4.2.1 Anonymisierungsverfahren. Das Arbeitspaket untersuchte den Einfluss von Informationsverlust und AUC-ROC-Werten auf den UCI Adult-Datensatz [4] sowie den MIMIC-III-Datensatz [5, 6] durch Variation der maximalen Unterdrückungsgrenze (Suppmax) und des Anonymitätsparameters k. Nach der Vorverarbeitung umfasst der Adult-Datensatz acht Quasi-Identifikatoren: Alter, Berufskategorie, Bildungsniveau, Familienstand, Beruf, ethnische Zugehörigkeit, Geschlecht und Nationalität. Der MIMIC-III-Datensatz enthält neun Quasi-Identifikatoren, darunter demografische Informationen, Vitalparameter, Laborwerte und Medikamentenstatus. Mithilfe von CatBoost [8] wurden für die anonymisierten Datensätze zwei Klassifikationsaufgaben durchgeführt: die Vorhersage, ob das Jahreseinkommen 50.000 US-Dollar übersteigt, sowie die Prognose der Einjahresmortalität.

Für die binäre Klassifikationsaufgabe wurden maximale Unterdrückungsgrenzen von 10%, 20% und 30% festgelegt, wobei k im Bereich von 2 bis 20 variiert wurde. Die fehlenden Datenpunkte im MIMIC-III-Datensatz resultieren daraus, dass bei bestimmten Parameterkombinationen (Suppmax = 10% mit $k \in \{10, 15\}$ sowie Suppmax = 20% mit $k = 15$) keine Lösung ermittelt werden konnte, die bei niedrigen Unterdrückungsgrenzen ein ausreichendes Anonymitätsniveau gewährleistet. Hervorzuheben ist, dass eine Lockerung der maximalen Unterdrückungsgrenze datensatzunabhängig zu verbesserten AUC-ROC-Werten führt. Dies ist darauf zurückzuführen, dass bei höheren Unterdrückungsraten in der Regel geringere Generalisierungsstufen zur Erfüllung der Anonymitätsanforderungen genügen, wodurch klassenspezifische Unterschiede erhalten bleiben.

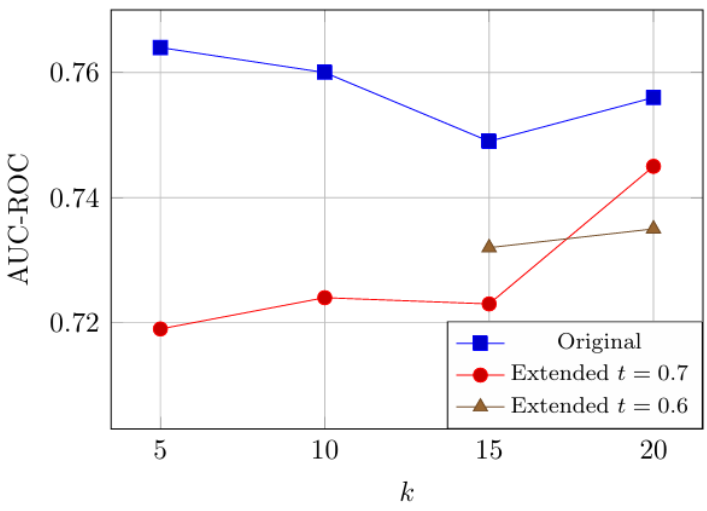
AP 4.2.2 Implementierung. Ergänzend wurde ein universelles Anonymisierungssystem entwickelt, das mit der Registersoftware OSSE (Open Source Registry System for Rare Diseases) [9] kompatibel ist. Über eine REST-API ermöglicht das System externen Anwendungen das Hochladen von Datensätzen sowie zugehöriger Schemadefinitionen und das Anfordern entsprechender Anonymisierungsrichtlinien. Das System verfügt über eine integrierte Zugriffskontrolle und unterstützt eine flexible Anonymisierungskonfiguration auf Attributebene.
Das Anonymisierungssystem operiert als eigenständiger Webserver und empfängt Datensät-



(a) 10 %

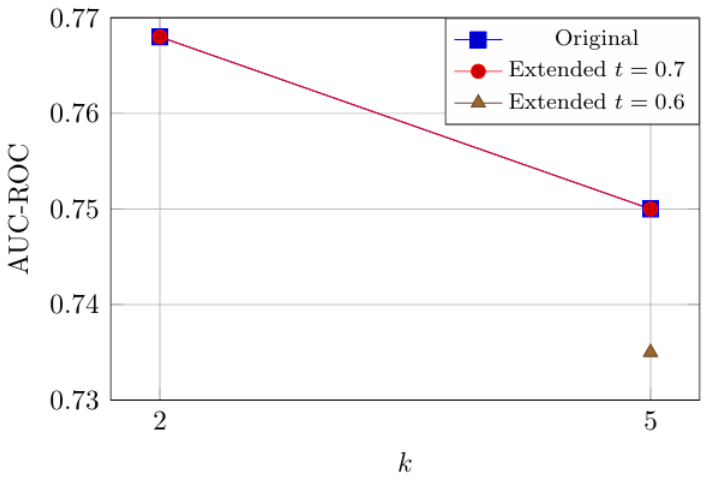


(b) 20 %

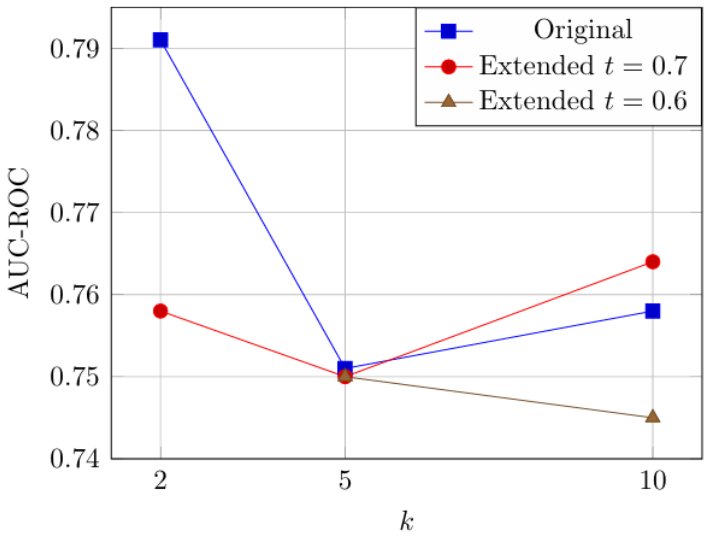


(c) 30 %

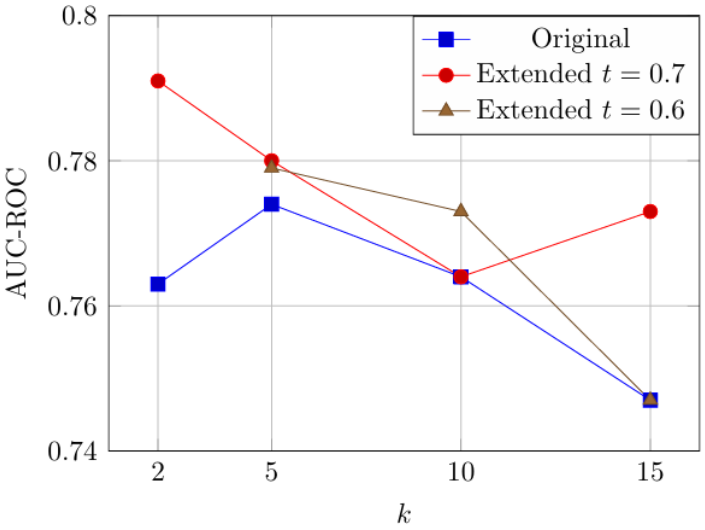
Abbildung 3: Vergleich der AUC-ROC-Verläufe zwischen Original- und erweitertem Modell für verschiedene k-Werte auf dem Adult-Datensatz mit einer maximal zulässigen Unterdrückung von 10%, 20%, 30%.



(a) 10 %



(b) 20 %



(c) 30 %

Abbildung 4: Vergleich der AUC-ROC-Verläufe zwischen Original- und erweitertem Modell für verschiedene k-Werte auf dem MIMIC-III-Datensatz mit einer maximal zulässigen Unterdrückung von 10%, 20%, 30%.

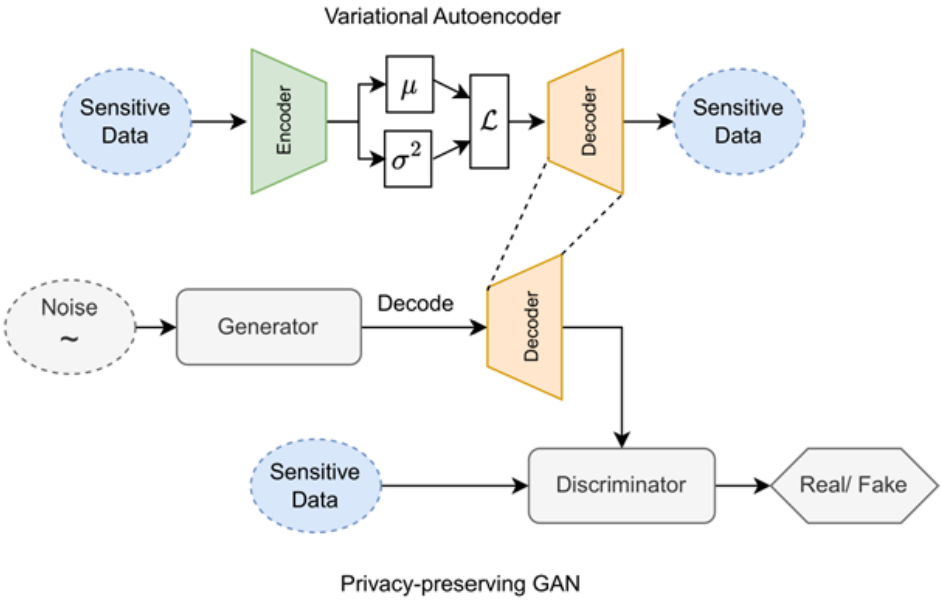


Abbildung 5: Zeitplan der Arbeitspakete

ze im CSV-Format sowie Metadaten im JSON-Format. Letztere dienen der Definition von Datenstrukturen, Quasi-Identifikatoren, Generalisierungshierarchien und Klassifikationszielen. Dieses Design ermöglicht es OSSE und vergleichbaren Plattformen, Anonymisierungsanfragen via HTTP zu übermitteln, wodurch eine effiziente Plattforminteroperabilität sowie eine nahtlose Integration gewährleistet werden.

4.2.3 AP 4.3 ML-Resilienz

AP 4.3.1 De-Anonymisierung durch Data Mining. Die in unserer Arbeit entwickelte Methodik befasst sich mit der Integration von Rényi-Differential Privacy (RDP) mit einem Generative Adversarial Network (GAN), das durch einen Variational Autoencoder (VAE) erweitert wurde, was zu unserer RDP-VAEGAN-Architektur führt (siehe Abb. 1).
Die GAN-Architektur besteht wie oben schon dargestellt aus zwei Komponenten:

- **Generator (G)**: Nimmt zufälliges Rauschen z auf und gibt synthetische Daten $\hat{x}$ aus. Er versucht, Daten zu erstellen, die von realen Daten nicht zu unterscheiden sind.
- **Diskriminator (D)**: Erhält entweder echte Daten x oder synthetische Daten vom Generator und versucht, zwischen den beiden zu unterscheiden. Der Diskriminator gibt eine Wahrscheinlichkeit aus, die angibt, ob die Eingabedaten echt oder synthetisch sind.

Das VAE wird danach integriert, um die Daten in einen kontinuierlichen latenten Raum zu überführen, was den Generierungsprozess durch das GAN erleichtert, insbesondere bei der Verarbeitung von diskreten und gemischten Datentypen. Das VAE besteht aus:

- **Encoder**: Wandelt echte Daten in eine latente Repräsentation um.
- **Decoder**: Rekonstruiert Daten aus dem latenten Raum, die im Trainingsprozess des Generators verwendet werden.

Unser wesentliches Ziel ist die Wahrung des Datenschutzes während des Trainings des GAN und VAE. Dies wird durch die Anwendung von Rényi-Differential Privacy erreicht. Unsere Arbeit zeigt, dass durch das Training des Diskriminators mit differential privacy und die Verwendung

des VAE zur Vorverarbeitung der Eingabe für den Generator das gesamte System sicherstellt, dass keine sensiblen Informationen aus den synthetischen Daten abgeleitet werden können.

Das Experiment wurde auf zwei Datensätzen durchgeführt: Early Prediction of Sepsis from Clinical Data (Sepsis) [10] und dem Cardiovascular Diseases Dataset [11]. Den Ergebnissen in Tabelle 1 zufolge liegen die Resultate von RDP-VAEGAN näher an den Real-Daten und sind daher besser als die beiden Benchmark-Methoden DPGAN und MedGAN darin, relevante Merkmale zu erfassen.

Tabelle 1: Vergleich verschiedener Verfahren anhand des $F1$-Scores im dimensionsweisen Vorhersagesetting; Klassifikatoren (außer "Real Data") sind mit synthetischen Daten trainiert.

Dataset	Real Data	DPGAN	MedGAN	RDP-VAEGAN
Sepsis	0.53	0.39	0.43	0.45
Cardiovascular Disease	0.49	0.27	0.31	0.38

Zur Bewertung des Modells geben wir den durchschnittlichen AUROC- (Area Under the Receiver Operating Characteristic) und AUPRC- (Area Under the Precision-Recall Curve) Score an. Tabelle 2 enthält die AUROC-Ergebnisse, und Tabelle 3 zeigt die AUPRC-Ergebnisse im Detail.

Tabelle 2: Vergleich verschiedener Modelle hinsichtlich der AUROC im Kontext der $(1, 10^{-5})$-DP-Einstellung. Für $\epsilon = \infty$ wurde unser RDP-VAEGAN-Modell ohne Privacy-Enforcement-Komponente implementiert. MedGAN verfügt ebenfalls über keine Privacy-Enforcement-Mechanismen. Alle Modelle wurden mit synthetischen Daten trainiert.

Dataset	$\epsilon = \infty$	DPGAN	MedGAN	RDP-VAEGAN
Sepsis	0.86	0.69	0.81	0.75
Cardiovascular Disease	0.77	0.62	0.73	0.70

Tabelle 3: Vergleich verschiedener Modelle hinsichtlich der AUPRC im Kontext der $(1, 10^{-5})$-DP-Einstellung. Für $\epsilon = \infty$ wurde unser RDP-VAEGAN-Modell ohne Privacy-Enforcement-Komponente implementiert. MedGAN verfügt ebenfalls über keine Privacy-Enforcement-Mechanismen. Alle Modelle wurden mit synthetischen Daten trainiert.

Dataset	$\epsilon = \infty$	DPGAN	MedGAN	RDP–VAEGAN
Sepsis	0.83	0.67	0.80	0.76
Cardiovascular Disease	0.76	0.60	0.72	0.68

Wenn $\epsilon = \infty$ ist, werden keine Datenschutzbeschränkungen angewendet. Basierend auf den Informationen in diesen Tabellen übertrifft unser Modell DPGAN in überwachten Lernumgebungen. Obwohl MedGAN sowohl bei AUROC als auch bei AUPRC besser abschneidet als DPGAN und RDP-VAEGAN, gewährleistet es keine Privatsphäre für die generierten Daten.

Die obigen experimentellen Ergebnisse zeigen die Leistung verschiedener Modelle im Rahmen des DP-Frameworks in überwachten und unüberwachten Lernsettings. Die Ergebnisse belegen, dass Variational Autoencoder Variablen, ob diskret, kontinuierlich oder eine Mischung aus beiden, effektiv verarbeiten können. Im Vergleich zu anderen Modellen weisen die künstlich generierten Daten unseres Systems eine überlegene Leistung bei Klassifikationsaufgaben auf, und alle Modelle arbeiten unter demselben Privacy-Budget.

AP 4.3.2 Verfahrensoptimierung und Kontrolle der Anonymisierungsregeln. Dieses Arbeitspaket reduziert Risiken zentraler Datenspeicherung und -übertragung durch dezentrale Verarbeitung und stärkt damit das Nutzervertrauen. Ohne wesentliche Beeinträchtigung der Datenverwendbarkeit berücksichtigt die Lösung heterogene Umgebungen (Rechenleistung, Netzbedingungen) sowie unterschiedliche Nutzerbedürfnisse. Sie ermöglicht die effiziente Ausführung von Anonymisierungsregeln auf lokalen Geräten (Smartphones, Wearables) und gewährleistet gleichzeitig starken Datenschutz.

Um die Herausforderungen heterogener Clients zu bewältigen und eine effektive Modellaggregation in der Cloud zu gewährleisten, schlagen wir den partiell trainierten, heterogenen Modellaggregationsalgorithmus cFedPT [12] vor. Dieser Algorithmus kann dynamisch Neuronen aus heterogenen Clientmodellen extrahieren und in das Training einbeziehen, wodurch Rechen- und Kommunikationskosten reduziert werden. Experimentelle Ergebnisse auf den Datensätzen ISIC2019 [13, 14, 15] und ICBHI [16] zeigen, dass cFedPT nicht nur Trainingskosten und Modellleistung entsprechend den Clientressourcen ausgleicht, sondern in Experimenten auch eine überlegene Modellleistung im Vergleich zu anderen Methoden erzielt (Tabelle 4).

Tabelle 4: Die F1-gewichteten Punktzahlen von cFedPT und anderen State-of-the-Art-Methoden.

Method	Homogeneous		Heterogeneous	
	ISIC2019	ICBHI	ISIC2019	ICBHI
Federated Dropout	80.84 ± 0.30	89.58±0.37	70.81 ± 2.71	92.74±1.03
FedPrune	76.60 ± 0.85	90.22±1.07	69.05 ± 1.38	92.52±1.11
HeteroFL	81.28 ± 0.27	91.96±0.29	74.37 ± 1.46	**94.37±0.30**
FedDSE	82.64 ± 0.33	92.03±0.39	74.82 ± 0.91	93.29±0.33
cFedPT	**85.02 ± 0.91**	**92.56±0.47**	**76.13 ± 0.67**	93.98±0.48

Obwohl föderiertes Lernen Vorteile wie Datenschutz und effiziente Datenverarbeitung bietet, ist es aufgrund der unterschiedlichen Rechenleistung der Clients und der Instabilität der Betriebsumgebung eingeschränkt. Wir schlagen daher FedRAS [17] vor, ein Framework für föderiertes Lernen mit zwei Strategien, das ressourcenbewusstes partielles Training und aufmerksamkeitsgesteuerte Selbstdestillation kombiniert. Dieses Framework ordnet trainierbare Modellteile dynamisch entsprechend dem Ressourcenstatus des Clients zu und gleicht lokale und globale Modellrepräsentationen durch einen Selbstdestillationsmechanismus an. Dadurch verbessert sich die Wirkung des föderierten Lernens in zwei Dimensionen: clientübergreifende Zusammenarbeit und lokale-globale Modellausrichtung.

Um Leistungsunterschiede zwischen FedRAS und anderen auf partiellem Training basierenden Methoden des föderierten Lernens quantitativ zu vergleichen, zeigt Tabelle 5 die durchschnittliche Genauigkeit von FedRAS und drei repräsentativen Methoden (Federated Dropout, HeteroFL und FedRolex) über alle Clients, basierend auf den durchschnittlichen Rechenkosten des Trainings (gemessen an der Parameteranzahl pro Aktualisierungsrunde). Tabelle 3 vergleicht die Unterschiede zwischen der in FedRAS angewandten Wissensdestillationsstrategie und anderen föderierten Lernmethoden, die Wissensdestillation einführen (FedKEM, DepthFL und FedBR), hinsichtlich der Modellleistung und der Anzahl der in das Training einbezogenen Parameter.

Experimentelle Ergebnisse zeigen, dass FedRAS nicht nur die durch partielles Training bedingte Leistungsverschlechterung behebt, sondern auch den Rechen- und Kommunikationsaufwand effektiv reduziert, abhängig von der Rechenleistung und Bandbreite der Clients. Im Vergleich zu Basismethoden, die entweder nur partielles Training oder nur Wissensdestillation nutzen, erzielt FedRAS Leistungsverbesserungen von 5.94% bzw. 3.65%. Dank seines

Tabelle 5: Vergleich der durchschnittlichen Genauigkeit von FedRAS mit anderen partiellen Trainingsmethoden.

Settings	Federated Dropout	HeteroFL	FedRolex	FedRAS
r=4	58.02 ± 0.85	68.22 ± 0.77	64.47 ± 0.18	**76.87±0.26**
r=2	60.81 ± 0.63	72.62 ± 0.46	69.41 ± 0.43	**78.56±0.40**

Tabelle 6: Vergleich der durchschnittlichen Genauigkeit von FedRAS und anderen Methoden der Wissensdestillation.

Metrics	DepthFL	FedKEM	FedBR	FedRAS
Acc (%)	65.67 ± 0.36	74.59 ± 0.31	75.07 ± 0.47	**78.72±0.41**
Params (MB)	46.14	48.66	**42.65**	43.33

ressourcenbewussten und heterogenen Parameter-Upload-Mechanismus verringert FedRAS zudem potenzielle Datenschutzrisiken bei der Übertragung von Modellparametern während des föderierten Lernens und gewährleistet gleichzeitig Konvergenz und Modellleistung.

Um den spezifischen Anforderungen von Clients in verschiedenen medizinischen Anwendungen gerecht zu werden, schlagen wir das föderierte Lernframework MFedKD vor. Es beinhaltet einen kollaborativen Trainingsmechanismus mit Wissensdestillation und eine flexible Strategie zum Hochladen von Parametern. In zwei Experimenten evaluieren wir die Auswirkungen der Wissensdestillation in MFedKD auf den Datenschutz der einzelnen Clients. Tabelle 7 zeigt Bildrekonstruktionsangriffe, die Gradientenlecks des Modells anhand von 30 Trainingsbeispielen simulieren. Ein höherer LPIPS-Wert deutet auf eine geringere Rekonstruktionsqualität der Trainingsbeispiele und damit auf ein reduziertes Risiko von Datenschutzverletzungen hin.

Tabelle 7: LPIPS-Vergleich auf dem ISIC-Datensatz mit CNN5 und ResNet18.

Upload	Client 1	Client 2	Client 3	Client 4	Average LPIPS
CNN5	0.61±0.0033	0.63±0.0055	0.61±0.0055	0.60±0.0010	0.61
ResNet18	0.59±0.0047	0.54±0.0062	0.51±0.0082	0.54±0.0090	0.55

Bei allen Clients lagen die LPIPS-Werte von CNN5 über denen der entsprechenden ResNet-Ergebnisse. Dies deutet darauf hin, dass die Verwendung eines Studentennetzwerks (CNN5) für die Parameterinteraktion zwischen Client und Server den Datenschutzverlust effektiv reduziert.

Abbildung 6 bewertet in Kombination mit Local Differential Privacy das Gleichgewicht, das das MFedKD-Framework zwischen Modellleistung und Datenschutz erreicht. Ein kleineres ϵ erhöht das Datenschutzbudget und führt zu mehr Rauschen im Modell, wodurch der Datenschutz verbessert wird, aber potenziell die Modellleistung beeinträchtigt wird.

Mit steigendem Datenschutzniveau sinkt die Gesamtgenauigkeit der clientseitigen Modelle. Modelle mit Wissensdestillation schneiden jedoch durchweg besser ab als Modelle ohne.

Durch die Kombination dieser beiden Evaluierungsexperimente zu Datenschutz und Modellleistung im föderierten Lernen konnten wir bestätigen, dass der Einsatz von Wissensdestillation den Zielkonflikt zwischen Datenschutz und Modellleistung mildert. Er optimiert nicht nur die Modellleistung innerhalb desselben Datenschutzbudgets, sondern reduziert auch effektiv das Risiko potenzieller Informationslecks während der Modellinteraktionen. Dadurch entsteht ein föderiertes Lernframework, das sowohl effizient als auch datenschutzfreundlich ist.

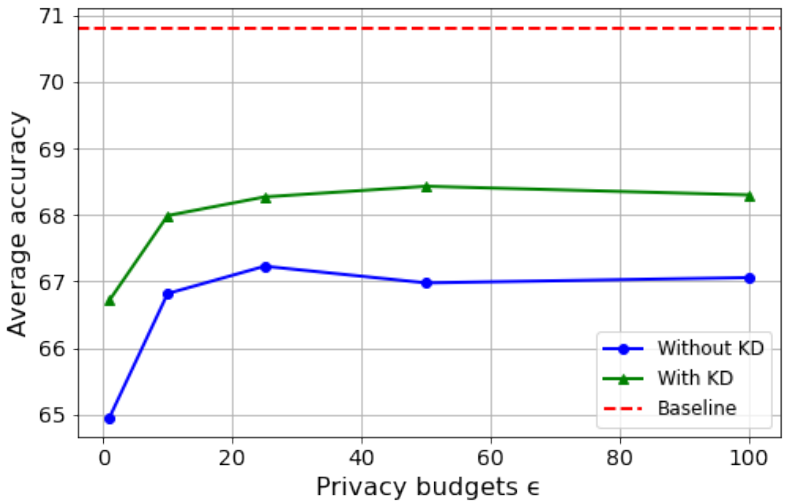


Abbildung 6: Durchschnittliche Genauigkeit über verschiedene Datenschutzbudgets unter lokaler differentieller Privatsphäre

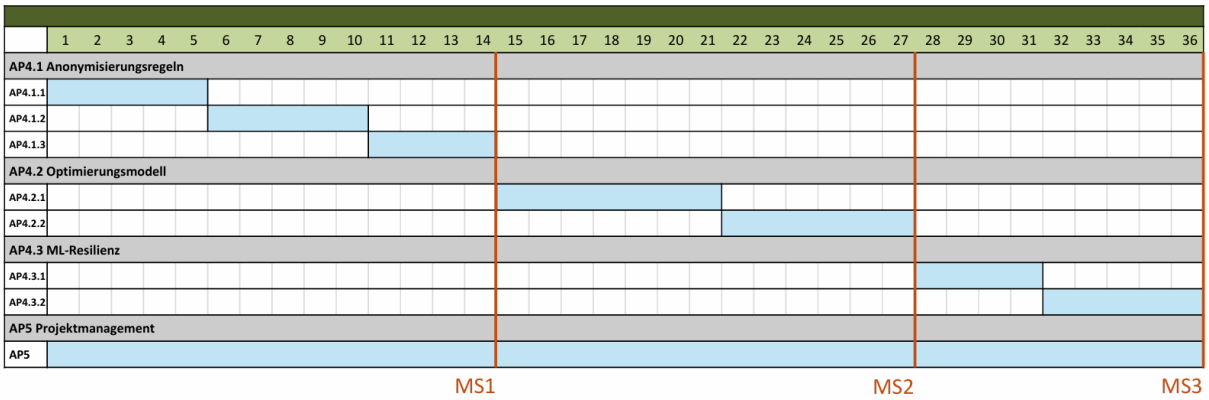


Abbildung 7: Zeitplan der Arbeitspakete

5 Ergebnisse

Im Folgenden werden die Ergebnisse anhand der eingesetzten Methoden dargestellt.

5.1 Webanwendung

Eine besondere Herausforderung in der registerbasierten Forschung besteht darin, sicherzustellen, dass Datensätze geteilt oder analysiert werden können, ohne die Privatsphäre Einzelner zu gefährden. Techniken wie k-Anonymität oder andere Anonymisierungsmodelle werden häufig eingesetzt, um das Risiko einer Wiedererkennung zu reduzieren. Die Umsetzung dieser Techniken erfordert jedoch die Lösung komplexer Optimierungsprobleme. Wir setzten ein allgemein verwendbares Anonymisierungssystem um, welches basierend auf "simulated annealing", k-Anonymität garantiert. Es ist für die Verwendung mit der Register-Software OSSE (Open Source System for Electronic Data Capture) konzipiert. "Simulated Annealing" sucht nach einem optimalen Gleichgewicht aus maximaler Privatsphäre und minimalem Informationsverlust. Das System bietet eine API-Schnittstelle an um externen Anwendungen zu erlauben, Daten und ihre Schemas hochzuladen, Anonymisierungsstrategien anzufragen und diese zu empfangen. Es enthält einen einfachen Zugriffsschutz und erlaubt die Verarbeitung flexibler Anonymisierungskonfigurationen auf Attributsebene. Die Leistungsfähigkeit wird anhand einiger Metriken gegen das bereits etablierte ARX Werkzeug mit zwei Beispieldatensätzen verglichen. Verwendet werden "UCI Adult" und "UCI Obesity Estimation" mit Metriken zur Klassifizierungsqualität (accuracy, balanced accuracy und F1-Score) und Datennutzbarkeit ("Discernibility Metric"). Hier bietet sich das Simulated Annealing (simuliertes Abkühlen) als attraktive Methode an, um die genannten Probleme zu bewältigen: Durch die schrittweise Verringerung des "Temperatur"-Parameters konvergiert Simulated Annealing zu Lösungen, die einen akzeptablen Kompromiss zwischen analytischer Validität und Schutz vor Wiedererkennung bieten. Dies stimmt mit dem Kernprinzip von OSSE überein, sichere, aber dennoch nützliche Datenaustausche zu ermöglichen. Um die Kompatibilität mit OSSE, einer Java-basierten Webanwendung, sicherzustellen, folgt die Architektur dieses Projekts einem ähnlichen dienstorientierten Design. Konkret wird das Anonymisierungssystem als eigenständiger Webserver implementiert, der seine Funktionalität über eine RESTful API bereitstellt. Dieser Ansatz ermöglicht es OSSE und ähnlichen Systemen, Anonymisierungsanfragen über HTTP einzureichen und gewährleistet dadurch Plattforminteroperabilität und eine einfache Integration. Das System ist so konzipiert, dass es Datensätze im CSV-Format empfängt, begleitet von Metadaten im JSON-Format, die die Datenstruktur beschreiben. Diese Metadaten enthalten Definitionen von Quasi-Identifikatoren, Generalisierungsstufen und Klassifikationszielen.

5.2 Neuronale Netze

Im Projekt wurde ein Multi-Task-Föderiertes-Lernframework erstellt, dass Knowledge Distillation einsetzt, um große, präzise Modelle in kleinere, effizientere Modelle zu komprimieren. Dadurch sollen sowohl heterogene Geräte als auch heterogene Labels integriert werden, um so ein globales Modell zur Klassifikation zu trainieren. Experimentelle Ergebnisse mit dem öffentlichen Dermoskopiedatensatz ISIC2019 zeigen, dass das Framework die Rechen- und Kommunikationskosten für ressourcenbeschränkte Clients (Endgeräte) erheblich reduziert und gleichzeitig hochleistungsfähigen Clients die flexible Auswahl von neuronalen Netzwerken ermöglicht. Verschiedene Clients bearbeiten verschiedene Klassifizierungsaufgaben für die acht Hautläsionstypen im ISIC2019-Datensatz: melanozytärer Nävus (NV), aktinische Keratose (AK), vaskuläre Läsion (VASC), gutartige Keratose (BKL), Dermatofibrom (DF), Melanom (MEL), Plattenepithelkarzinom (SSC), Basalzellkarzinom (BCC): (1) Client 1 verwendet ein einfaches binäres

Klassifikationssystem, das Bilder als "gesund" oder "ungesund" kategorisiert, um schnell Hautzustände zu identifizieren, die weitere Aufmerksamkeit erfordern. (2) Client 2 nutzt ein dreistufiges Klassifikationssystem: "gesund", "gutartig" und "bösartig"; die Unterscheidung zwischen gutartigen und bösartigen Zuständen ist relevant für eine detaillierte Analyse im klinischen Umfeld. (3) Client 3 verfeinert die Klassifikation weiter und kategorisiert Bilder in vier Gruppen: MEL, BCC, SCC und "Sonstige". Dies ist besonders relevant für Spezialanalysen, die verschiedene Formen von Hautkrebs unterscheiden müssen. (4) Client 4 verwendet das detaillierteste Klassifikationssystem, das die Bilder in acht verschiedene Hautläsionstypen gemäß dem ISIC2019-Datensatz unterteilt. Dieses System unterstützt umfassende diagnostische Prozesse und präzise Behandlungsplanung. Die experimentellen Ergebnisse zeigen, dass Clients, die Knowledge Distillation nutzen, eine bessere Leistung erzielen. Zukünftige Forschungen sollen die Robustheit und Vielseitigkeit des Ansatzes durch Anwendung des Frameworks auf weitere Datensätze und Szenarien validieren.

6 Wichtigste Positionen des zahlenmäßigen Nachweises

Der Hauptteil der Finanzmittel wurde wie geplant über die gesamte Laufzeit des Projekts für wissenschaftliche Mitarbeiter:innen (insbesondere die Entwicklung der Software durch MSc. Mohan Xu / MSc. Pronaya Prosun Das sowie die Entwicklung der Webanwendung durch BSc. Despina Tawadros) eingesetzt. Das Projekt wurde auf Seiten des ITEM wie geplant ausgeführt. Es waren keine Abweichungen vom Finanzplan nötig.

7 Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten

Die Notwendigkeit und Angemessenheit ergibt sich durch folgende Ergebnisse im Bereich Anonymisierung und Datensicherheit, die mehrere wichtige Erkenntnisse zeigen:

1. Simulated Annealing kann lokale Optima überwinden und mehrere Anonymisierungsstrategien erkunden, was es besonders geeignet macht für die hochgradig nichtlineare Optimierungslandschaft im Spannungsfeld zwischen Datenschutz und Datennutzbarkeit.

2. Wahrung des Datenschutzes während des Trainings eines GAN und VAE zur Generierung synthetischer Daten wird erreicht durch die Anwendung von Rényi-Differential Privacy.

3. Federated Learning heterogener Clients kann effizient implementiert werden und eine effektive Modellaggregation in der Cloud umgesetzt werden.

8 Verwertbarkeit der Ergebnisse / Zukunftsprognose

Der Gesundheitsmarkt befindet sich im Zuge der Digitalisierung in einer dynamischen Expansionsphase; insbesondere der Bereich Smart Health mit Wearables bietet hier ein hohes kommerzielles Potenzial mit steigender Nachfrage nach innovativen Lösungen für patientenzentrierte Diagnostik, Therapiebegleitung und präventive Gesundheitsversorgung.

Die Marktaussichten des Projekts basieren auf drei zentralen Komponenten unseres Konsortiums: Technologisches Alleinstellungsmerkmal ist die datenschutzkonforme Verarbeitung sensibler Gesundheitsdaten; ein direkter Marktzugang durch den beteiligten Industriepartner

ist sichergestellt; ein klinisch relevanter Nutzen ergibt sich durch präzise diagnostische und therapiebegleitende Anwendungen, die auf Wearables-Daten basieren.

Ein entscheidender Wettbewerbsvorteil liegt in der vollständig dezentralen Anonymisierung von Sensordaten direkt auf mobilen Endgeräten – noch vor einer zentralen Datenspeicherung oder -aggregation. Durch diesen frühestmöglichen Schutz personenbezogener Daten erfüllen wir nicht nur höchste Datenschutzstandards, sondern ermöglichen auch skalierbare, ressourcenschonende Verarbeitung trotz begrenzter Rechenleistung auf Wearables. Dieses Verfahren eröffnet attraktive Verwertungschancen in mehreren Zielmärkten, darunter:

- Datenschutzkonforme Plattformen für Wearables (B2B-Vertrieb an Hersteller und Gesundheitsdienstleister),
- Erweiterte Gesundheitsmonitoring-Systeme für chronisch erkrankte Patient:innen (z. B. in der Diabetes-, Herz-Kreislauf- oder Krebsnachsorge)
- sowie KI-gestützte Diagnostiktools für telemedizinische Anwendungen, die auf hochwertigen, anonymisierten Datensätzen aufbauen.

Zum Projektabschluss liegt ein voll funktionsfähiger Prototyp des Anonymisierungssystems vor, der nahtlos mit bestehenden Wearables integrierbar ist. Eine Validierung und weitere Entwicklungsstufen sind in konkreten Anschlussprojekten geplant. Durch die frühe Positionierung im Wachstumsmarkt Smart Health entstehen vielfältige kommerzielle Verwertungsmöglichkeiten – von der direkten Lizenzierung über SaaS-Modelle bis hin zu datengetriebenen Dienstleistungen für das Gesundheitswesen der Zukunft.

9 Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen

Herkömmliche datenschutzwahrende Ansätze stützen sich typischerweise auf zentralisierte Architekturen, bei denen sensible Daten in einem zentralen Speicher gesammelt, gespeichert und verarbeitet werden [18, 19, 20]. Obwohl in vielen Szenarien wirksam [21], führen solche Architekturen zu einem Single Point of Failure und erhöhen das Risiko von Datenpannen, unbefugtem Zugriff und Re-Identifizierungsangriffen [22]. Darüber hinaus birgt die Übertragung großer Datenmengen sensibler Informationen auf zentralisierte Server zusätzliche Risiken während der Übertragung und wirft Bedenken hinsichtlich des Datenbesitzes und der Benutzerkontrolle auf. Dies gilt auch für aktuelle angewandte Systeme wie ARX [3]. Der Fokus unseres Projektes lag daher auf der Dezentralisierung und der Einbindung von ressourcenbeschränkten Geräten. Datenschutz ist ein weiterer kritischer Faktor gerade in medizinischen Anwendungen, da Patientendaten vertraulich bleiben müssen, um den Datenschutzgesetzen zu entsprechen und das Vertrauen der Patienten zu erhalten. Um dieses Problem zu lösen, werden datenschutzfreundliche Techniken zur Generierung synthetischer Daten [23] verwendet, um realistische synthetische Daten zu erzeugen und gleichzeitig den Datenschutz der Patienten zu wahren. Diese Techniken umfassen die Erstellung synthetischer Daten, die statistisch echten Patientendaten ähneln, jedoch keine identifizierbaren Patientendaten enthalten; hier konnte wir vorhandene Methoden im Rahmen des Projekts um einen neuen Ansatz erweitern.

10 Erfolgte Veröffentlichungen

10.1 Publikationen

Folgende Veröffentlichungen wurden verfasst, die in Tabelle 8 kurz inhaltlich beschrieben werden.

- Privacy-Preserving Medical Data Generation Using Adversarial Learning [1]
- Constrained Adaptive Partial Training for Federated Learning on Heterogeneous Clients [12]
- FedRAS: A Dual-Strategy Framework for Federated Learning on Heterogeneous Devices [17]
- Optimizing Personalized Federated Learning on Heterogeneous Problems via Knowledge Distillation [24]

10.2 Verbreitung und Öffentlichkeitsarbeit der Projektergebnisse

Neben der oben dargestellten Publikationslage sind Prototypen entstanden, die in Repositorien[1][2][3] zur Verfügung gestellt werden. Die Projektergebnisse wurden zudem auf Konferenzen präsentiert.

Literatur

[1] Pronaya Prosun Das, Despina Tawadros, and Lena Wiese. Privacy-preserving medical data generation using adversarial learning. In *International Conference on Information Security*, pages 24–41. Springer, 2023.

[2] William Winkler. Using simulated annealing for k-anonymity. Technical report, Research Report 2002-07, US Census Bureau Statistical Research Division, 2002.

[3] Florian Kohlmayer, Fabian Prasser, Claudia Eckert, Alfons Kemper, and Klaus A Kuhn. Flash: efficient, stable and optimal k-anonymity. In *2012 International conference on privacy, security, risk and trust and 2012 international confernece on social computing*, pages 708–717. IEEE, 2012.

[4] Arthur Asuncion, David Newman, et al. Uci machine learning repository, 2007.

[5] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[6] Alistair Johnson, Tom Pollard, and Roger Mark. MIMIC-III Clinical Database. *PhysioNet*, September 2016. doi: 10.13026/C2XW26. URL https://doi.org/10.13026/C2XW26. Version 1.4.

[1] https://github.com/deFlip99/PPRL-using-Bloomfilters
[2] https://github.com/tina5980g/MA_OSSE/tree/main
[3] https://github.com/desstaw/DataPrivacy_SimulatedAnnealing

Publikation	Technische Ergebnisse
[1]	• Mithilfe von Adversarial Learning haben wir neuartige, datenschutzfreundliche Verfahren zur Generierung synthetischer Daten entwickelt, die realistische synthetische Daten erzeugen und gleichzeitig die Privatsphäre der Patienten schützen. • Das Konzept der Differential Privacy (DP) wird als formales System angewendet und erweitert, das einen Rahmen zur Quantifizierung und Kontrolle von Datenschutzverlusten bietet. • Die Arbeit untersucht die Vorteile von DP bei der Gewährleistung eines hohen Datenschutzniveaus, ohne die Nutzbarkeit der Daten wesentlich zu beeinträchtigen.
[12]	• Diese Arbeit untersucht eine heterogene Modellaggregationsmethode auf Basis partiellen Trainings, bei der Clients nur Teilmodelle eines globalen Modells trainieren, um Rechen- und Kommunikationskosten zu reduzieren. • Wir schlagen den adaptiven partiellen Trainingsalgorithmus (cFedPT) vor, der bei der Neuronenauswahl sowohl aktuelle als auch historische Aktivierungen berücksichtigt und dadurch Über- bzw. Untertraining vermeidet. • Experimente zu Bild- und Audioklassifikationsaufgaben bestätigen die Wirksamkeit von cFedPT in föderierten Lernumgebungen mit heterogenen Clients.
[17]	• Wir schlagen ein duales föderiertes Lernframework (FedRAS) für heterogene Endgeräte vor, das zwei komplementäre Modellaggregationsstrategien integriert: partielles Modelltraining und Wissensdistillation. • Jeder Client trainiert adaptiv den Teil des Modells, der seiner verfügbaren Ressourcenkapazität entspricht, und lädt die Modellparameter in rollierender Weise hoch. Dadurch wird die Ressourcennutzung maximiert und Leerlaufzeiten leistungsstarker Geräte reduziert. • Zur weiteren Leistungssteigerung bei partiellem Training integrieren wir einen Aufmerksamkeitsmechanismus in das Selbst-Distillationsframework, der es lokalen Modellen ermöglicht, sich effektiver am globalen Modell auszurichten. • Durch das Hochladen lediglich heterogener und unvollständiger Teilmengen von Modellparametern verringert unser Ansatz potenzielle Datenschutzrisiken, die mit der Parameterübertragung im föderierten Lernen verbunden sind.
[24]	• Um Herausforderungen des Federated Learning (Heterogenität der Clients in Bezug auf Rechenkapazität, Kommunikationsbedingungen) zu bewältigen, schlägt dieser Artikel ein Multi-Task-Federated-Learning-Framework basierend auf Knowledge Distillation (MFedKD) vor, das effektiv die Heterogenität von Geräten und Labels adressiert und gleichzeitig die Übertragung von Modellparametern reduziert. • Experimentelle Ergebnisse auf dem ISIC2019-Dermoskopie-Datensatz zeigen, dass MFedKD überdurchschnittliche Leistungen erzielt, während der Rechen- und Kommunikationsaufwand für ressourcenbeschränkte Clients deutlich gesenkt wird. • Darüber hinaus bestätigen Rekonstruktionsexperimente basierend auf der Umkehrung von Gradienten die Wirksamkeit von MFedKD beim Schutz der Privatsphäre der Nutzer

[7] Stephany N Duda, Nan Kennedy, Douglas Conway, Alex C Cheng, Viet Nguyen, Teresa Zayas-Cabán, and Paul A Harris. Hl7 fhir-based tools and initiatives to support clinical research: a scoping review. *Journal of the American Medical Informatics Association*, 29 (9):1642–1653, 2022.

[8] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, 2018.

[9] Tim Vincent Knecht, Jens Göbel, Axel Zieschank, Holger Storf, and Jessica Vasseur. Open source registry system for rare diseases (osse): Development and usability evaluation of an initial audit trail prototype. In *German Medical Data Sciences 2025: GMDS Illuminates Health*, pages 112–121. IOS Press, 2025.

[10] Matthew A Reyna, Christopher S Josef, Russell Jeter, Supreeth P Shashikumar, M Brandon Westover, Shamim Nemati, Gari D Clifford, and Ashish Sharma. Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. *Critical care medicine*, 48(2):210–217, 2020.

[11] S. Ulianova. Cardiovascular disease dataset. https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset, Jan 2019.

[12] Mohan Xu and Lena Wiese. Constrained adaptive partial training for federated learning on heterogeneous clients. In *International Symposium on Algorithmic Aspects of Cloud Computing*, pages 43–57. Springer, 2025.

[13] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):180161, 2018.

[14] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

[15] Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*, 2019.

[16] Bruno M Rocha, Dimitris Filos, Lea Mendes, Ioannis Vogiatzis, Eleni Perantoni, Evangelos Kaimakamis, Panagiotis Natsiavas, Ana Oliveira, Cristina Jácome, Alda Marques, et al. A respiratory sound database for the development of automated classification. In *International conference on biomedical and health informatics*, pages 33–37. Springer, 2017.

[17] Mohan Xu and Lena Wiese. Fedras: a dual-strategy framework for federated learning on heterogeneous devices. *Journal of Cloud Computing*, 2026.

[18] Pierangela Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001. doi: 10.1109/69.971193.

[19] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002. doi: 10.1142/S0218488502001648.

[20] Djordje Slijepčević, Maximilian Henzl, Lukas Daniel Klausner, Tobias Dam, Peter Kieseberg, and Matthias Zeppelzauer. k-anonymity in practice: How generalisation and suppression affect machine learning classifiers. *Computers & Security*, 111:102488, 2021. doi: 10.1016/j.cose.2021.102488.

[21] Kristen LeFevre, David J. DeWitt, and Raghu Ramakrishnan. Mondrian multidimensional k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)*, pages 25–25. IEEE, 2006. doi: 10.1109/ICDE.2006.101.

[22] Arvind Narayanan and Vitaly Shmatikov. How to break anonymity of the netflix prize dataset, 2007. URL http://arxiv.org/abs/cs/0610105. arXiv preprint arXiv:cs/0610105.

[23] Fan Liu, Zhiyong Cheng, Huilin Chen, Yinwei Wei, Liqiang Nie, and Mohan Kankanhalli. Privacy-preserving synthetic data generation for recommendation systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1379–1389, 2022.

[24] Mohan Xu and Lena Wiese. Optimizing personalized federated learning on heterogeneous problems via knowledge distillation. Europ. Interdisc. Cybersec. Conf. (under review).