A-SIT

Zentrum für sichere Informationstechnologie - Austria

Authentizität & Korrektheitsgarantien in Federated Learning: Überblick & Methoden

Authentizität & Korrektheitsgarantien in Federated Learning: Überblick & Methoden

Autor: Karl W. Koch

Tel.: +43 316 873 - 5517
E-Mail: karl.koch@tugraz.at

Datum: 20.12.2024

Kurzfassung:

Mithilfe von maschinellem Lernen (ML) - *bzw. oft als KI/AI bezeichnet* - und einer Vielzahl an Daten von verschiedensten Nutzern, können nahezu alle Anwendungen/Services verbessert werden; z.B. Analyse von Vitalaktivitäten via SmartWatches bzw. „Wearables" generell, Aufzeichnung vom Fahrverhalten oder verbesserte Krebsanalysen via MRI-Bilder. Einerseits werden solche Szenarien immer attraktiver, weil die entsprechenden ML-Technologien mittlerweile praktisch relevant sind, und andererseits ist dabei die Wahrung der Privatsphäre eine der größten Herausforderungen. Weiters können Anwender eines ML-Modells die Frage stellen, ob und inwiefern einem ML-Modell vertraut werden kann?

Um ein globales ML-Modell basierend auf Daten von vielen End-Nutzer-Geräten zu trainieren, und auch die Privatsphäre der Nutzer zu bewahren, hat Google 2016/17 Federated Learning (FL) ins Leben gerufen. Bei einem FL-Framework, welches praktische Relevanz aufweist, muss man - zusätzlich zur Wahrung der Privatsphäre - auf Vertrauen bzw. Trust im ganzen Zyklus einer FL-Epoche achten. Um ein zuverlässigeres bzw. vertrauenswürdiges aktualisiertes ML-Modell zu erzielen, müssen daher **Authentizitäts- und Korrektheitsgarantien für Trainings-Daten und ML-Parameter/-Gewichte im FL-Prozess** hinzugefügt werden. Im Zuge des FL-Prozessflusses gibt es sechs wichtige Vertrauenspunkte bzw. Trust-Punkte (TPs) in vertrauenswürdigem FL („Trusted FL"). Um nun beides zu erreichen **– Schutz der Privatsphäre und Vertrauenswürdigkeit –** muss FL abermals upgegradet werden.

Deshalb werden in diesem Bericht Methoden für Trusted Federated Learning für die einzelnen Trust-Punkte gezeigt, und anschließend der Aspekt des gesamten FL-Prozessflusses diskutiert. Abschließend werden im Rahmen einer Conclusio & Weiterführende Arbeiten, die gewonnenen Erkenntnisse kurz zusammengefasst und potentiell-interessante weiterführende Richtungen aufgezeigt.

Zu jedem Trust-Punkt gibt es bereits vielversprechende individuelle Lösungsmethoden und teilweise konkrete Instanziierungen. Die aktuell gängigsten Methoden und konkreten Instanziierungen basieren auf dem privatsphären-bewahrenden kryptografischen Baustein Null-Wissen-Beweis („Zero-Knowledge Proof" / ZKP). Die verschiedenen individuellen Lösungs-Konzepte bieten unterschiedliche Trade-Offs; z.B. adressieren nicht alle Konzepte den Aspekt der Privatsphäre beim Aggregieren der retournierten ML-Parameter.

Ein Gesamtkonzept von Trusted Federated Learning im Rahmen des gesamten Prozessflusses, unter Wahrung der Privatsphäre, stellt die nächste wissenschaftliche und konstruktions-technische Herausforderung dar. Z.B. die effiziente Kombination von digitalen Signaturen und ZKPs. Und vor allem weil die meisten relevanten Werke von „zkFL" - Trusted FL via ZKPs - erst zwischen 2022 und 2024 publiziert worden sind, bleibt es spannend zu sehen wie sich dieses neue Feld weiterentwickelt, und wichtig für, z.B., etwaige praktische Anwender am Puls der Zeit bzw. Stand der Technik zu bleiben.

Article I. Inhaltsverzeichnis

0. Abkürzungsverzeichnis

FL	„Federated Learning" (Föderiertes Lernen)
ML	„Machine Learning" (Maschinelles Lernen)
MPC	„Secure Multi-Party Computation" (Sichere Mehrparteien-Berechnung)
SecAgg	„Secure Aggregation" (Sicheres Aggregieren)
Trust	Vertrauen
Trust-Punkt	Vertrauens-Aspekt bzw. -Punkt
VSS	„Verifiable Secret Sharing" (verifizierbares geheimes Teilen)
ZKP	„Zero-Knowledge Proof" (Null-Wissen-Beweis)
zkFL	„ZKP-enhanced Federated Learning" (Trusted FL via ZKPs)

1. Einleitung

Mithilfe von maschinellem Lernen (ML) - *bzw. oft als KI/AI bezeichnet* - und einer Vielzahl an Daten von verschiedensten Nutzern, können nahezu alle Anwendungen/Services verbessert werden; z.B. Analyse von Vitalaktivitäten via SmartWatches bzw. „Wearables" generell[1], Aufzeichnung vom Fahrverhalten[2] oder verbesserte Krebsanalysen via MRI-Bilder[3]. Einerseits werden solche Szenarien immer attraktiver, weil die entsprechenden ML-Technologien mittlerweile praktisch relevant sind, und andererseits ist dabei die Wahrung der Privatsphäre eine der größten Herausforderungen. Weiters können Anwender eines ML-Modells die Frage stellen, ob und inwiefern einem ML-Modell vertraut werden kann?

Warum Federated Learning (FL)? Um ein globales ML-Modell basierend auf Daten von vielen End-Nutzer-Geräten zu trainieren, und auch die Privatsphäre der Nutzer zu bewahren, hat Google 2016/17 FL ins Leben gerufen (McMahan et al. 2017). Bei FL trainiert jeder Nutzer lokal das entsprechende ML-Modell, und sendet „lediglich" die aktualisierten ML-Modell-Parameter – *oft Gewichte genannt* – zurück an den Server. Dieser Prozess bezeichnet eine FL-Epoche. Um solch ein ML-Modell optimal zu trainieren, kann es mehrerer solcher FL-Epochen bedingen. Jedoch hat es in den letzten Jahren praktisch-relevante Angriffe auf „Standard FL" gegeben, in denen, z.B., die Trainingsdaten von den retournieren ML-Modell-Gewichten rekonstruiert wurden.

Um einen guten Trade-Off zwischen ML-Modell-Genauigkeit und Wahrung der Privatsphäre zu erreichen, bietet sich föderiertes Lernen („Federated Learning" / FL) mit sicherer Aggregation („Secure Aggregation" / SecAgg) basierend auf, z.B., sicherer Mehrparteien-Berechnung („Secure Multi-Party Computation" / MPC) an. Siehe, z.B., (Bonawitz et al. 2017), (Bonawitz et al. 2022) oder (Koch 2024) für einen initialen Überblick von privatsphären-bewahrendem FL.

1.1. Motivation & Trust-Punkte von Trusted Federated Learning

Vertrauen in FL-trainierte ML-Modelle? Zwar ist nun die Privatsphäre der Trainings-Daten von Teilnehmern:innen gewahrt, jedoch kann sich der Anwender eines solchen ML-Modells fragen, ob den verwendeten Trainings-Daten bzw. aktualisierten ML-Parametern vertraut werden kann? Zum Beispiel im Kontext von virtuellen Tastaturen hat ein App-Provider – wie etwa Google's Gboard[4,5] - wahrscheinlich großes Interesse, dass das gelernte ML-Modell für die Vorhersage des nächsten Wortes bzw. Satzes adäquate Resultate liefert, und somit die Wahrscheinlichkeit erhöht, dass die App weiterhin verwendet wird. Oder im Kontext von Gesundheits-Diagnosen, wo z.B. ein Spital bzw. Doktor:in großes Interesse daran hat, dass das gelernte ML-Modell adäquate Diagnosen von gezeigten MRI-Bildern liefert. Zwar gibt bei solchen Szenarien *üblicherweise* der:die Dokotor:in die finale Diagnose, jedoch werden ML-Modelle immer häufiger als Unterstützung eingesetzt[6,7,8]. Und somit hat ein gut trainiertes ML-Modell das Potential schon frühzeitig einen etwaigen Krebs zu erkennen, und dies dem:der jeweiligen Doktor:in umgehend weiterzuleiten. Z.B. Abbildung 1 zeigt ein *grob-schematisches* Spital-übergreifendes FL-Szenario, um ein effizienteres & effektiveres *globales* ML-Modell, für die Analyse von MRI-Bildern, zu konstruieren.

[1] mdpi.com/1424-8220/23/5/2821
[2] link.springer.com/article/10.1007/s11227-023-05364-3
[3] nypost.com/2024/07/20/health/ai-detects-cancer-with-17-more-accuracy-than-doctors-ucla-study
[4] play.google.com/store/apps/details?id=com.google.android.inputmethod.latin
[5] aclanthology.org/2023.acl-industry.60.pdf
[6] en.wikipedia.org/wiki/Computer-aided_diagnosis
[7] mdpi.com/2673-9909/4/3/59
[8] arxiv.org/abs/2311.08908

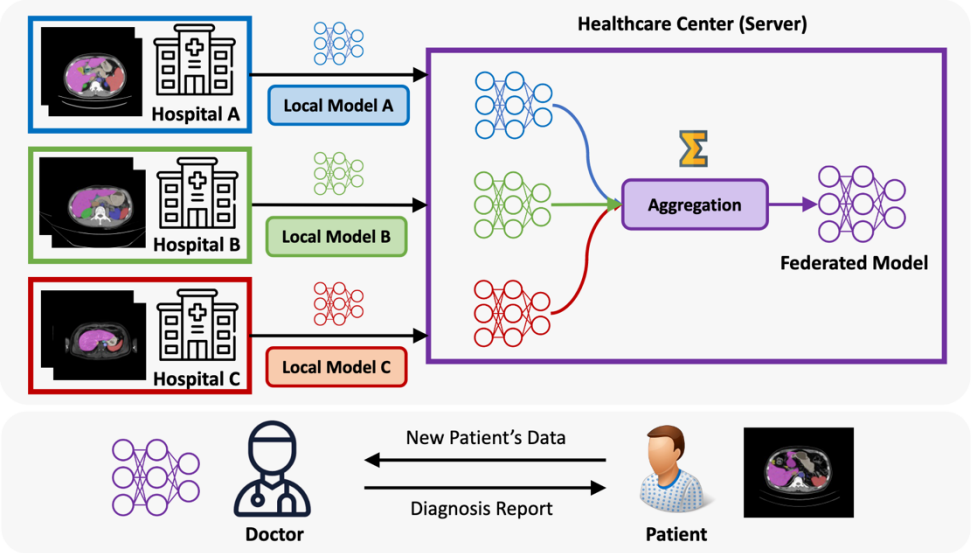

Abbildung 1 – *Grob-schematischer Überblick von Spital-übergreifendem FL, um ein effizienteres & effektiveres globales ML-Modell zu konstruieren. Dabei könnten, z.B., die 3 Spitäler (A,B,C) international verteilt sein und jeweils seltene und signifikante Datensätze herbergen; welche zwar nicht per se geteilt werden können, aber in das Training von einem ML-Modell fließen können. Was zu einer verbesserten Diagnoseunterstützung für Doktoren führt, und schlussendlich zu mehr Gesundheit für Patienten, da etwaige Risiken schon potentiell frühzeitiger entdeckt werden.*

Quelle: (Zhu et al. 2024).

6 Trust-Punkte. Somit wird es stets wichtiger, dass keine „Fake-Daten" bei FL bzw. ML verwendet werden, sondern adäquate Trainings-Daten, ML-Trainings-Funktionen und ML-Modell-Parameter zum Einsatz kommen. In Bezug auf das Vertrauen („Trust") in FL-trainierte ML-Modelle ergeben sich dabei **6 führende Fragen bzw. Trust-Punkte im FL-Prozess.** Jeweils 3 Fragen bzw. Trust-Punkte im Aspekt der Authentizität und Korrektheit:

(0) (Authentizität-Server) **ML-Modell vom Server** stammt? Und auch nicht während des Transfers modifiziert wurde?

(1) (Authentizität-Client) **Trainings-Daten von einer vertrauenswürdigen Quelle** stammen? Und somit keine „Fake-Daten" verwendet wurden. Z.B. Zeugnisse von einer Universität, Gesundheitsdaten von einem Spital oder Fitnessaktivitäten von einer verifizierten bzw. vertrauten SmartWatch.

(2) (Authentizität-Client) **Authentische Daten für das Trainieren** eines ML-Modells verwendet worden sind?

(3) (Korrektheit-Client) **Korrekte ML-Trainings-Funktionen** ausgeführt wurden?

(4) (Korrektheit-Client) **Korrekte resultierende ML-Parameter** bzw. -Gewichte an den Server retourniert worden sind? Bzw. im SecAgg-Prozess verwendet wurden.

(5) (Korrektheit-Server) **Korrekte ML-Modell-Aggregation** – mit den korrekten resultierenden ML-Parametern bzw. -Gewichten – ausgeführt wurde?

Um ein zuverlässigeres bzw. vertrauenswürdiges aktualisiertes ML-Modell zu erzielen, müssen daher **Authentizitäts- und Korrektheitsgarantien für Trainings-Daten und ML-Gewichte im FL-Prozess** hinzugefügt werden. Um nun beides zu erreichen **– Schutz der Privatsphäre und Vertrauenswürdigkeit („Trust") –** muss FL abermals upgegradet werden. Hierfür können z.B. privatsphären-bewahrende kryptografische Bausteine, wie etwa Null-Wissen-Beweise („Zero-Knowledge-Proofs" / ZKPs) für die Korrektheit oder Gruppen-Signaturen für die Authentizität, eingesetzt werden. Dabei kann man den FL-Prozessfluss näher betrachten und relevante Bereiche für Authentizität und Korrektheit beleuchten. Abbildung 2 zeigt den FL-Prozessfluss mit den sechs erwähnten Trust-Punkten, für das „Vertrauens Upgrade".

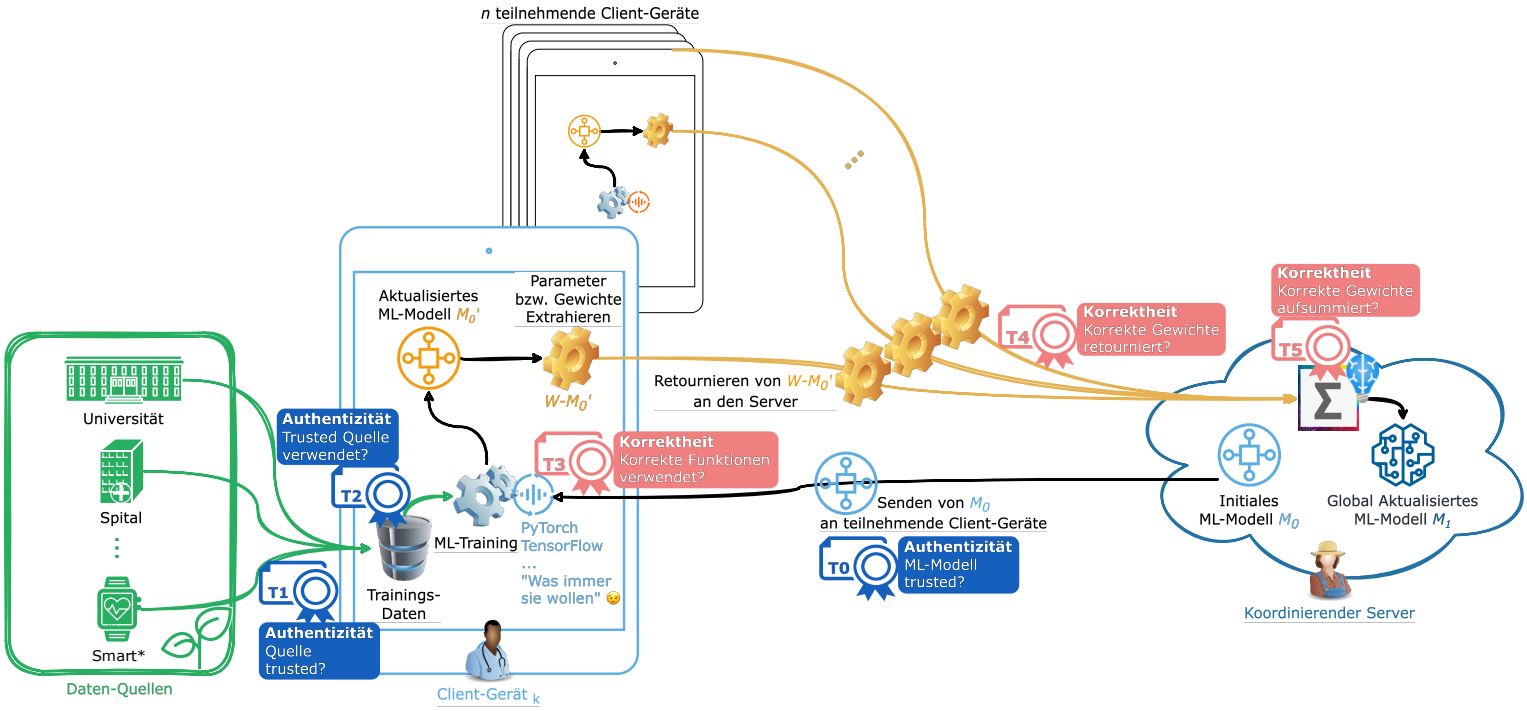

***Abbildung 2** - Prozess-Fluss einer Federated-Learning-Epoche inklusive der vier erwähnten Trust-Punkte. Wovon jeweils zwei Trust-Punkte den Aspekt der Authentizität und der Korrektheit adressieren. Der Authentizitäts-Aspekt bezeichnet – in diesem Szenario – ob die Trainings-Daten (T1) von einer vertrauenswürdigen bzw. trusted Quelle stammen und (T2) diese auch tatsächlich verwendet worden sind. Der Korrektheits-Aspekt bezeichnet – in diesem Szenario ob (T3) die korrekten Funktionen beim ML-Training ausgeführt worden sind und (T4) die korrekten resultierenden ML-Parameter bzw. -Gewichte an den Server retourniert werden.*

Gezeichnet mit draw.io. Das SmartWatch-Symbol stammt von Vecteezy.com, und das Blatt-Symbol von Flaticon.com.

Ziele & Ausblick. Zwar adressieren, z.B., die zwei erwähnten kryptografischen Bausteine – *isoliert betrachtet* – die jeweiligen Trust-Punkte *relativ einfach und geradeaus*, jedoch ist es nicht geradlinig wie diese Bausteine bzw. Methoden für alle sechs Trust-Punkte gleichzeitig - von den Trainings-Daten bis zum aktualisierten ML-Modell – eingesetzt werden können. Oder zumindest für mehr als einen Trust-Punkt, in einem ersten Schritt. Deshalb werden im folgenden Abschnitt (2) Methoden für Trusted Federated Learning für die einzelnen Trust-Punkte gezeigt, und anschließend der Aspekt des gesamten FL-Prozessflusses - das (2.7) Σ Summa Summarum ("Putting it all Together") - diskutiert. Abschließend werden im Rahmen einer (3) Conclusio & Weiterführende Arbeiten, die gewonnenen Erkenntnisse kurz zusammengefasst und potentiell-interessante weiterführende Richtungen aufgezeigt.

2. Methoden für Trusted Federated Learning

In diesem Abschnitt werden Methoden gezeigt, um „Standard FL" zu vertrauenswürdigem bzw. „Trusted FL" upzugraden. Bei einem FL-Framework, welches praktische Relevanz aufweist, muss man auf Trust im ganzen Zyklus einer FL-Epoche achten. Wie in der Einleitung beschrieben, gibt es sechs wichtige Vertrauenspunkte bzw. Trust-Punkte (TPs) in Trusted FL. Zu jedem TP wird eine Methode - inklusive Referenz auf mindestens ein konkretes Protokoll – beschrieben, um den jeweiligen TP zu erfüllen. Abschließend wird auf den konvergierenden Aspekt aller TPs im ganzen FL-Prozess-Fluss eingegangen.

2.1. (TP0-Server) Authentizität des verwendeten ML-Modells

Abbildung 3 zeigt den Ausschnitt von Trust-Punkt 0 im Prozess-Fluss einer FL-Epoche, inklusive ⭐Trust-Upgrade mit Signaturen (siehe Abbildung 13 für den Gesamt-Überblick aller ⭐Trust-Upgrades).

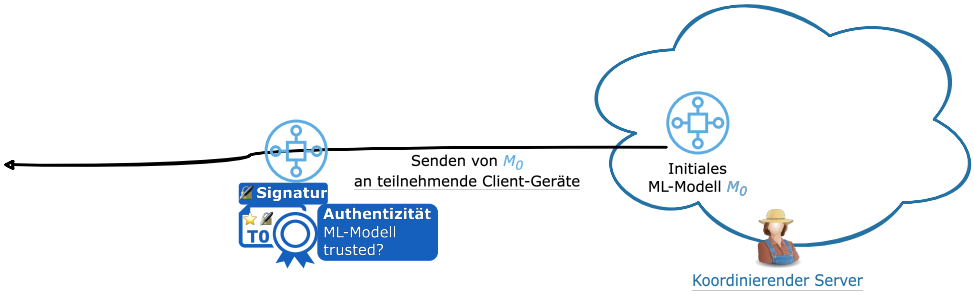

***Abbildung 3 - (Authentizität)** Fokus auf Trust-Punkt 0 im Prozess-Fluss einer FL-Epoche: ob Trainings-Daten von einer vertrauenswürdigen bzw. trusted Quelle stammen? Inklusive dem ⭐Trust-Upgrade mit Signaturen🖋.*

🎨 *Gezeichnet mit draw.io.*

Am Beginn des FL-Prozess-Flusses müssen Clients verifizieren ob dem verwendeten bzw. zugesandten ML-Modell vertraut werden kann. Wie generell bei Dokumenten bzw. Daten bietet sich hierbei eine digitale Signatur an. D.h. der Server signiert das zugesandte ML-Modell🖋 mit seinem privaten Signaturschlüssel 🔑sk. Clients können dann die Signatur🖋 mit dem zugehörigen öffentlichen Signaturschlüssel 🔑pk verifizieren. Neben der Authentizität, kann hierbei auch die Integrität des ML-Modells überprüft werden; d.h., dass das Modell unverändert transferiert wurde.

Je nach Trust-Modell im jeweiligen FL-Szenario, könnte ein Server – als Zusatz zur Signatur🖋 - das verwendete ML-Modell (semi-)öffentlich auf ein Archiv hochladen, zu welchem Clients Zugriff haben. Clients könnten dann zusätzlich abgleichen, ob das zugesandte ML-Modell der Version auf dem Archiv gleicht. Ist der Archiv-Zugriff mit einem zusätzlichen Trust-Mechanismus gesichert – wie ein weiteres Password oder Passkey – stellt das Archiv eine erweiterte Sicherheitsmaßnahme dar.

2.2. (TP1-Client) Authentizität der Daten-Quelle

Abbildung 4 zeigt den Ausschnitt von Trust-Punkt 1 im Prozess-Fluss einer FL-Epoche, inklusive ⭐Trust-Upgrade (siehe Abbildung 13 für den Gesamt-Überblick aller ⭐Trust-Upgrades).

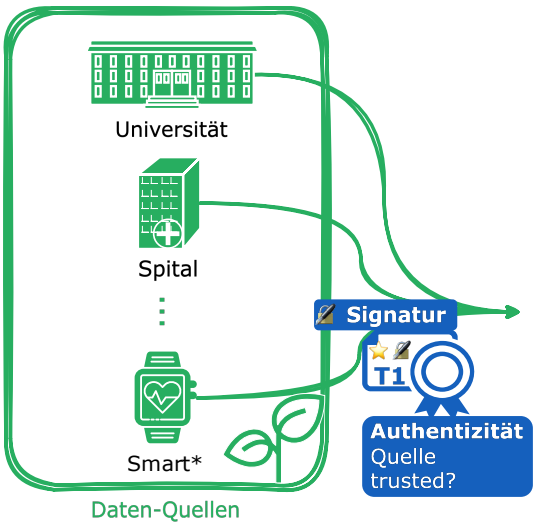

Abbildung 4 - (Authentizität) Fokus auf Trust-Punkt 1 im Prozess-Fluss einer FL-Epoche: ob Trainings-Daten von einer vertrauenswürdigen bzw. trusted Quelle stammen? Inklusive dem ⭐Trust-Upgrade mit Signaturen🖋.

🎨 *Gezeichnet mit draw.io. Das SmartWatch-Symbol stammt von Vecteezy.com, und das Blatt-Symbol von Flaticon.com.*

Eine der gängigsten Methoden um Authentizität der Daten-Quelle zu garantieren sind **digitale Signaturen**. Damit kann z.B. ein Spital einen Datensatz – mit dem privaten Signaturschlüssel 🔑sk - signieren und dies dem jeweiligen Client zur Verfügung stellen, welcher dann den signierten Datensatz🖋 fürs ML-Modell-Trainieren verwendet. Der signierte Datensatz🖋 kann dann vom Client, der verifizierenden Entität („Verifier"), etc. - mithilfe des öffentlichem bzw. public Signaturschlüssels 🔑pk – verifiziert werden, ob er von einer trusted Quelle (z.B. Spital in diesem Falle) signiert worden ist. Eine trusted Quelle kann ihren 🔑pk z.B. auf einem zugänglichen Archiv – wie ein GitHub-Repository – hochladen.

Dabei kann es Szenarien geben, wo es gewünscht ist im Zweifel nachzusehen wer genau (Arzt, Abteilung, etc.) signiert hat. Um aber die Privatsphäre der jeweiligen signierenden Personen, Abteilungen, etc. zu wahren bietet sich der privatsphären-bewahrende kryptografische Baustein der digitalen **Gruppen-Signaturen** an. (Chaum and van Heyst 1991) waren eine der ersten, welche Gruppen-Signaturen konzipiert hatten. Gruppen-Signaturen haben den Vorteil, dass man innerhalb der jeweiligen Organisation mehrere 🔑sk_i ableiten kann, und trotzdem nur einen 🔑pk benötigt um die Signaturen zu verifizieren. Zusätzlich kann man eine autorisierte Entität aufsetzen, welche im Zweifel die Identität der jeweiligen Signatur erfassen kann (wie ein Richter👨‍⚖️). Ein weiterer Vorteil der Gruppen-Signaturen besteht darin, dass man eine höhere Schlüssel-Agilität erreicht. So kann z.B. ein einzelner 🔑sk_i zurückgezogen werden, weil z.B. ein Arzt das Spital wechselt, sodass die anderen Schlüssel unberührt bleiben. Ansonsten müsste man ein neues digitales Schlüsselpaar generieren, und dies wieder an alle Ärzte, Abteilungen, etc. neu verteilen.

Z.B. (Koch et al. 2022) verwendeten digitale Gruppen-Signaturen🖋 um in einem ähnlichen Szenario („🐙KRAKEN") die Authentizität der Daten-Quelle zu verifizieren. In dem „🐙KRAKEN"-Szenario werden die jeweiligen signierten Datensätze für einzelne dedizierte MPC-Berechnungs-Knoten aufgeteilt und für eine spätere privatsphären-bewahrende Analyse zur Verfügung gestellt. Dabei führen die MPC-Berechnungs-Knoten die eigentlichen Berechnungen nur aus, wenn die Signatur🖋 als gültig verifiziert wurde.

2.3. (TP2-Client) Authentizität der Daten beim ML-Lernen

Abbildung 5 zeigt den Ausschnitt von Trust-Punkt 2 im Prozess-Fluss einer FL-Epoche, inklusive ⭐Trust-Upgrade (siehe Abbildung 13 für den Gesamt-Überblick aller ⭐Trust-Upgrades).

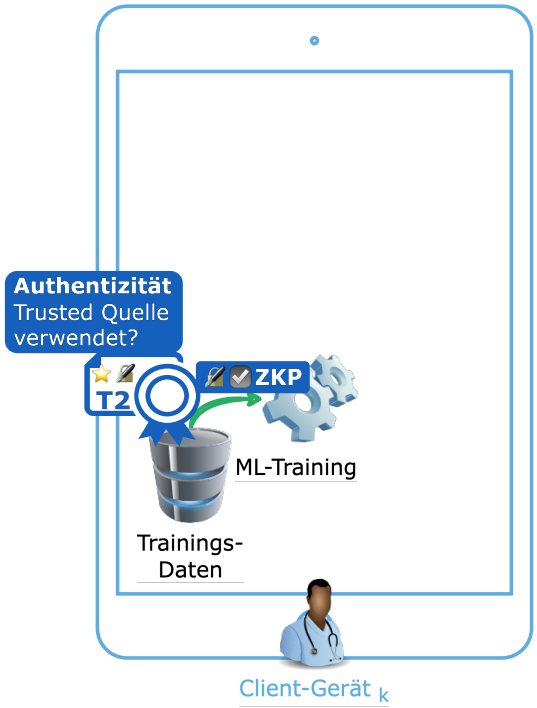

Abbildung 5 - (Authentizität) *Fokus auf Trust-Punkt 2 im Prozess-Fluss einer FL-Epoche: ob Trainings-Daten einer vertrauenswürdigen bzw. trusted Quelle verwendet worden sind? Inklusive dem ⭐Trust-Upgrade mit Signaturen🖋 und ZKPs ☑.*

🎨 *Gezeichnet mit draw.io.*

Um zu garantieren, dass authentische Daten tatsächlich beim Lernen des ML-Modells verwendet werden, muss die Signatur der Daten – unter der Annahme, dass auf digitale Signaturen aufgebaut wird - mit dem verwendeten Datensatz verknüpft werden.

Z.B. (Koch et al. 2022) verwenden den privatsphären-bewahrenden kryptografischen Baustein „**Pederson Commitment**" um die Signatur von Daten an den geheim aufgeteilten („secret-shared") Datensatz „anzuhängen". Mit dem Pederson Commitment und dem public 🔑pk, können die dedizierten MPC-Berechnungs-Knoten die Signatur auf den „secret-shared"-Daten verifizieren.

Im Bereich von FL bauen, z.B., (Zhu et al. 2024) ebenfalls auf solch ein hybrides Schema auf. Sie verwenden Pederson Commitments und verifizierbares geheimes Teilen („Verifiable Secret Sharing" / VSS) basierend auf Shamir und nennen ihr Konzept RiseFL. Dabei garantiert RiseFL, dass die resultierenden ML-Parameter bzw. -Gewichte innerhalb eines gültigen Wertebereichs liegen. Zusätzlich erhält der Server beim finalen Aggregieren keinen Einblick in die einzelnen retournierten Gewichte, sondern nur die Summe aller Gewichte. Abbildung 6 zeigt die verschiedenen Phasen von RiseFL.

Zwar gibt RiseFL keine Garantie, dass die verwendeten Trainings-Daten von einer trusted Quelle stammen, jedoch bietet RiseFL einen potentiellen Nährboden um ein FL-Framework dahingehend zu erweitern. Die Verifizierbarkeit von Trainings-Daten einer trusted Quelle, gegeben die retournierten Gewichte bzw. nur die finale Summe der Gewichte, ist eine komplexe Herausforderung für weiterführende Forschungs- und Konstruktions-Arbeiten.

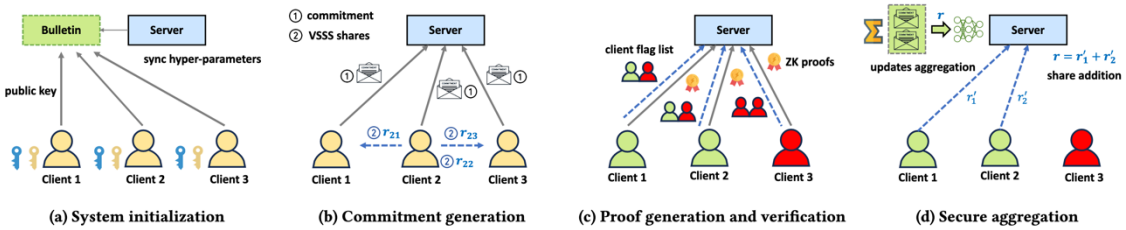

Abbildung 6 - *Die vier Phasen des FL-Konzepts RiseFL. (a) Zuerst werden relevante Meta-Parameter ausgetauscht (z.B. Anzahl an Teilnehmern und wie viele korrupt werden dürfen). (b) Clients commiten sich zu den retournierten ML-Parametern bzw. -Gewichten. (c) Die anderen Clients verifizieren die Korrektheit des Commitments, und auch ob die Gewichte in einem zulässigen Bereich liegen. (d) Der Server berechnet die finale Aggregation der Gewichte.*

🎨 *Quelle: (Zhu et al. 2024).*

2.4. (TP3-Client) Korrektheit der berechnenden ML-Funktionen

Abbildung 7 zeigt den Ausschnitt von Trust-Punkt 3 im Prozess-Fluss einer FL-Epoche, inklusive ⭐Trust-Upgrade (siehe Abbildung 13 für den Gesamt-Überblick aller ⭐Trust-Upgrades).

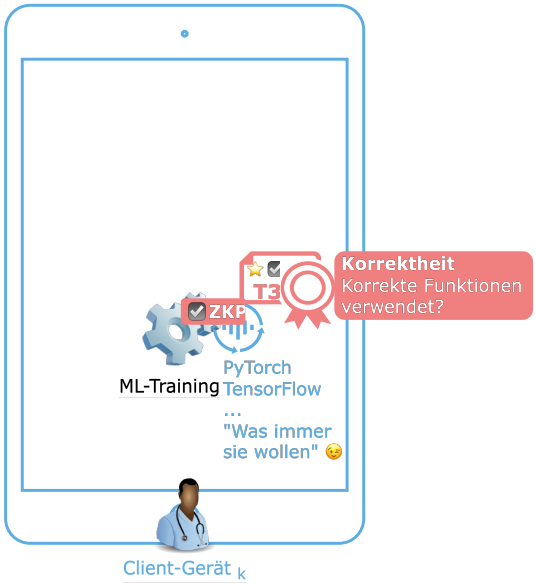

Abbildung 7 - *(Korrektheit) Fokus auf Trust-Punkt 3 im Prozess-Fluss einer FL-Epoche: ob korrekte ML-Trainings-Funktionen ausgeführt worden sind? Inklusive dem ⭐Trust-Upgrade mit ZKPs ☑.*

🎨 *Gezeichnet mit draw.io.*

Um zu garantieren, dass die korrekten ML-Funktionen während des ML-Trainings ausgeführt worden sind, bietet sich der privatsphären-bewahrende kryptografische Baustein der Null-Wissen-Beweise („Zer-Knowledge Proofs" / ZKPs) an. Z.B. (Xing et al. 2023) unterteilt die Herausforderung in mehrere ZKPs für die unterschiedlichen Berechnungsschritte während des ML-Trainings, und nennen ihren Ansatz ZKP-FL. Bzw. PZKP-FL für den praktischeren Aspekt. Abbildung 8 zeigt einen grob-schematischen Überblick von ZKP-FL.

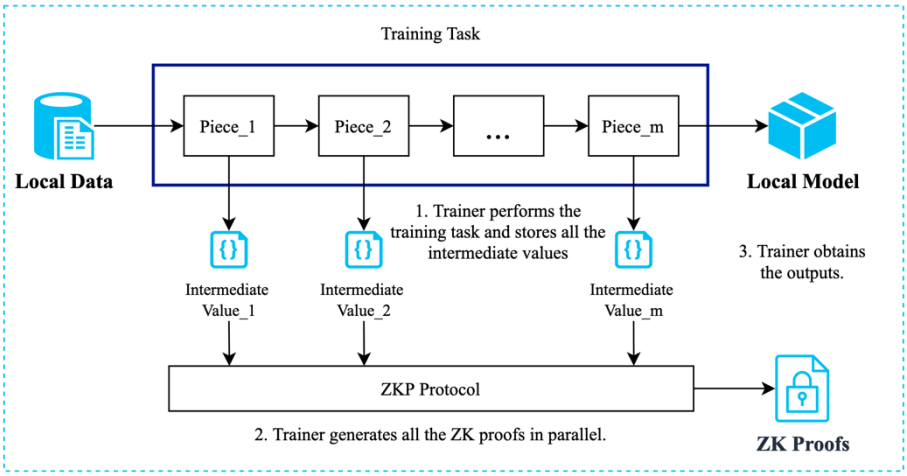

Abbildung 8 - *Grob-schematischer Überblick der verifizierten ML-Trainingsphase von ZKP-FL. Dabei werden für die verschiedenen Berechnungsschritte während der Trainings-Phase einzelne ZKPs erzeugt.*

Quelle: (Xing et al. 2023).

2.5. (TP4-Client) Korrektheit der retournierten Gewichte

Abbildung 9 zeigt den Ausschnitt von Trust-Punkt 4 im Prozess-Fluss einer FL-Epoche, inklusive ⭐Trust-Upgrade (siehe Abbildung 13 für den Gesamt-Überblick aller ⭐Trust-Upgrades).

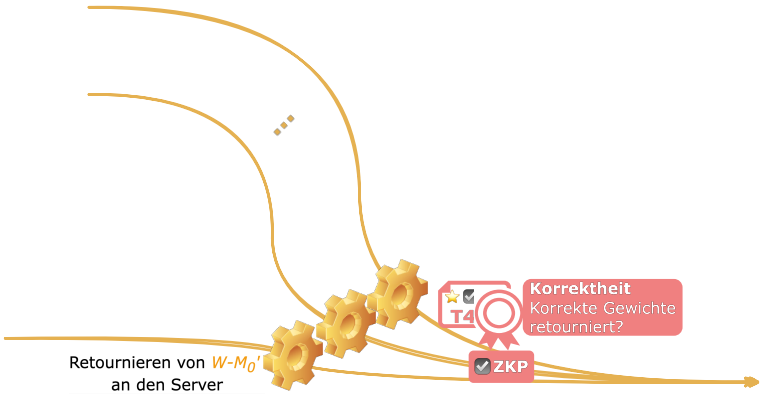

Abbildung 9 - *(Korrektheit) Fokus auf Trust-Punkt 4 im Prozess-Fluss einer FL-Epoche: ob korrekte resultierende ML-Parameter an den Server retourniert worden sind? Bzw. im SecAgg-Prozess verwendet wurden. Inklusive dem ⭐Trust-Upgrade mit ZKPs ☑.*

Gezeichnet mit draw.io.

Wie in 2.3 (TP2-Client) Authentizität der Daten beim ML-Lernen gezeigt, gibt RiseFL von (Zhu et al. 2024) Garantien, dass die retournierten ML-Parameter bzw. -Gewichte in einem gültigen Bereich liegen. RiseFL baut auf Pederson Commitments und VSS. Weiters konzipieren, z.B., auch (Lycklama et al. 2023) (RoFL) via additiven homomorphen Commitments, (Chowdhury et al. 2022) (EIFFeL) via VSS und geheim-geteilten nicht-interaktiven Beweisen („secret-shared non-interactive proofs") oder (Bell et al. 2023) (ACORN) via Bulletproofs – der Erweiterung vom traditionellen privatsphären-bewahrendem FL SecAgg bzw. SecAgg+ - ein FL-System mit Garantien der retournierten Gewichte. Von den genannten Konzepten berichtet RiseFL die effizientesten FL-Epochen. Abbildung 10 gibt einen Überblick der Laufzeiten unter verschiedenen Längen der ML-Parameter bzw. -Gewichte *(d)*.

Die Verknüpfung zum direkten Ergebnis des ML-Trainings ist eine komplexe Herausforderung und der nächste Schritt für diesen Trust-Punkt. Dabei stellt Trust-Punkt 3 (der berechnenden ML-Funktionen) einen potentiellen Startpunkt dar. D.h. eine Verknüpfung zu den ZKPs, dass die korrekten Funktionen ausgeführt worden sind.

#Param.	Approach	Client Computation (seconds)				Server Computation (seconds)				Comm. Cost per Client (MB)
		commit.	proof gen.	proof ver.	total	prep.	proof ver.	agg.	total	
d = 1K	EIFFeL	0.865	3.63	11.7	16.2	-	-	0.182	**0.182**	125
	RoFL	0.051	4.43	-	4.5	-	91.2	0.040	91.3	0.37
	ACORN	0.076	2.49	-	2.6	-	58.9	0.040	58.9	**0.004**
	RiseFL (ours)	0.054	1.48	0.08	**1.6**	1.17	75.6	0.071	76.8	0.44
d = 10K	EIFFeL	8.38	36.8	115	161	-	-	1.81	**1.81**	1250
	RoFL	0.51	46.4	-	46.9	-	860	0.41	860	3.66
	ACORN	0.75	24.5	-	25.3	-	522	0.41	523	**0.03**
	RiseFL (ours)	0.49	1.8	0.08	**2.3**	8.61	82.5	0.71	91.8	0.71
d = 100K	EIFFeL	84.7	382	1070	1536	-	-	18.8	**18.8**	12500
	RoFL	5.1	496	-	502	-	8559	4.1	8563	36.6
	ACORN	7.6	253	-	261	-	5087	4.1	5091	**0.3**
	RiseFL (ours)	4.8	4.5	0.08	**9.3**	73.3	139	7.2	219	3.5
d = 1M	EIFFeL	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
	RoFL	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
	ACORN	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
	RiseFL (ours)	48.0	31.2	0.08	**79.3**	653	612	72.1	**1338**	**30.9**

Abbildung 10 – Überblick der Laufzeiten unter verschiedenen Längen der ML-Parameter bzw. -Gewichte *(d)*. Dabei werden Berechnungs-Laufzeiten am Client und Server in Sekunden, sowie Kommunikations-Kosten per Client in MB gezeigt.

🎨 Quelle: (Zhu et al. 2024).

2.6. (TP5-Server) Korrektheit der aggregierenden Gewichte

Abbildung 11 zeigt den Ausschnitt von Trust-Punkt 5 im Prozess-Fluss einer FL-Epoche, inklusive ⭐Trust-Upgrade (siehe Abbildung 13 für den Gesamt-Überblick aller ⭐Trust-Upgrades).

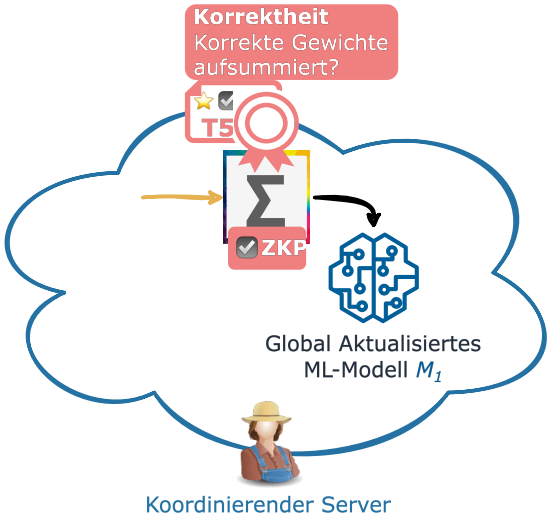

***Abbildung 11 -** (Korrektheit) Fokus auf Trust-Punkt 5 im Prozess-Fluss einer FL-Epoche: ob korrekte resultierende ML-Parameter bzw. – Gewichte vom Server bzw. im SecAgg-Prozess verwendet worden sind. Inklusive dem ⭐Trust-Upgrade mit ZKPs ☑.*

🎨 *Gezeichnet mit draw.io.*

Um die Korrektheit der Aggregation der ML-Gewichte auf Seiten des Servers zu verifizieren, muss dieser beweisen können, dass tatsächlich all die – von den Clients - retournierenden ML-Gewichte verwendet worden sind. Hierfür bietet sich abermals der privatsphären-bewahrende kryptografische Bausteine des **Null-Wissens-Beweis („Zero-Knowledge Proof" / ZKP)** an. Mithilfe solch eines ZKPs kann der Server den entsprechenden – auf Kryptografie- bzw. Mathematik-basierten – Beweis erstellen.

Z.B. (Wang et al. 2024) haben solch ein Konzept erstellt. Dabei konzipieren Wang et al. zwei Varianten im Hinblick auf die Verifikationsentitäten: (i) die Clients und (ii) Blockchain-Knoten, welche bei Verifikations-Erfolg den Hash-Wert des verschlüsselten resultierenden ML-Modells auf eine Blockchain speichern. Abbildung 12 zeigt einen grob-schematischen Überblick der ersten Variante von Wang et al., in welcher die Clients die ZKPs überprüfen.

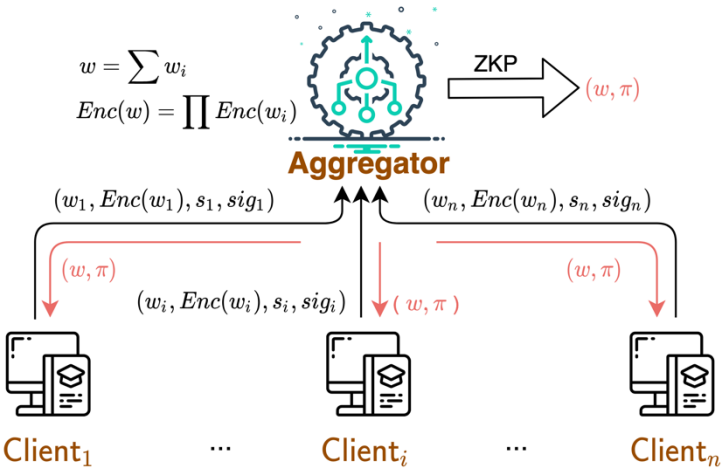

***Abbildung 12** - Grob-schematischer Überblick von dem Server-Verifikationsansatz der ML-Gewichte-Aggregation von (Wang et al. 2024). Wang et al. konzipieren auch eine Variante, in der Blockchain-Knoten – anstatt der einzelnen Clients – den ZKP-Beweis verifizieren und dann einen Hash in einer Blockchain speichern, welcher von den Clients verifiziert werden kann.*

🎨 *Quelle: (Wang et al. 2024).*

2.7. Σ Summa Summarum ("Putting it all Together")

Zu jedem Trust-Punkt gibt es bereits vielversprechende individuelle Lösungsmethoden und teilweise konkrete Instanziierungen. Jedoch gibt es – bis dato – kein Konzept welches alle Trust-Punkte gleichzeitig in einem gesamten FL-Prozess-Fluss adressiert. Dieses Gesamtkonzept von ganzheitlichem Trusted Federated Learning stellt die nächste wissenschaftliche und konstruktions-technische Herausforderungen dar. Abbildung 13 zeigt den gesamten FL-Prozess-Fluss inklusive ⭐Trust-Upgrades für jeden Trust-Punkt.

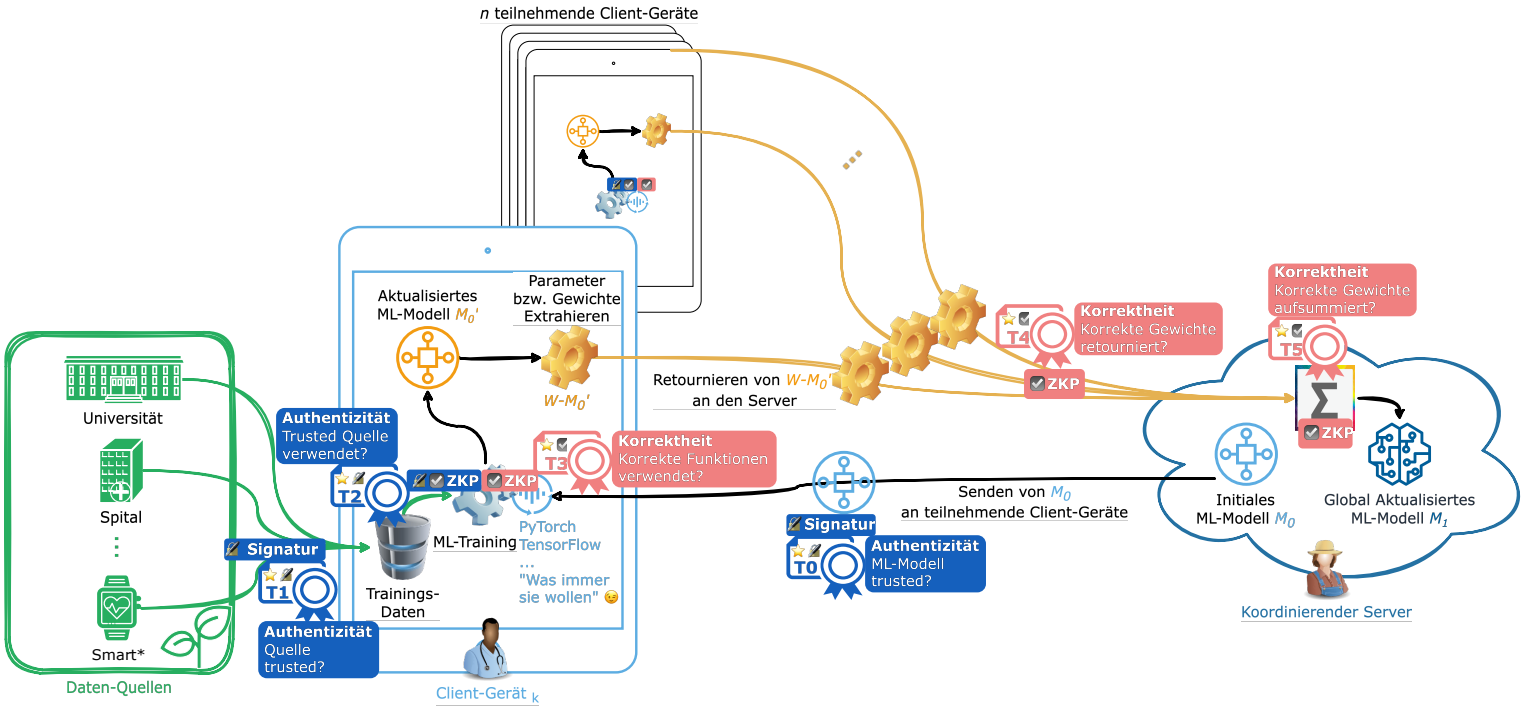

Abbildung 13 *- Prozess-Fluss einer Federated-Learning-Epoche inklusive der sechs erwähnten Trust-Punkte (0—5) und dem gesamtheitlichen ⭐Trust-Upgrade für jeden Trust-Punkt. Dabei kommen primär zwei Technologien zum Einsatz. Für den Authentizitäts-Aspekt digitale Signaturen ✒️. Für den Korrektheits-Aspekt ZKPs ☑️.*

🎨 *Gezeichnet mit draw.io. Das SmartWatch-Symbol stammt von Vecteezy.com, und das Blatt-Symbol von Flaticon.com.*

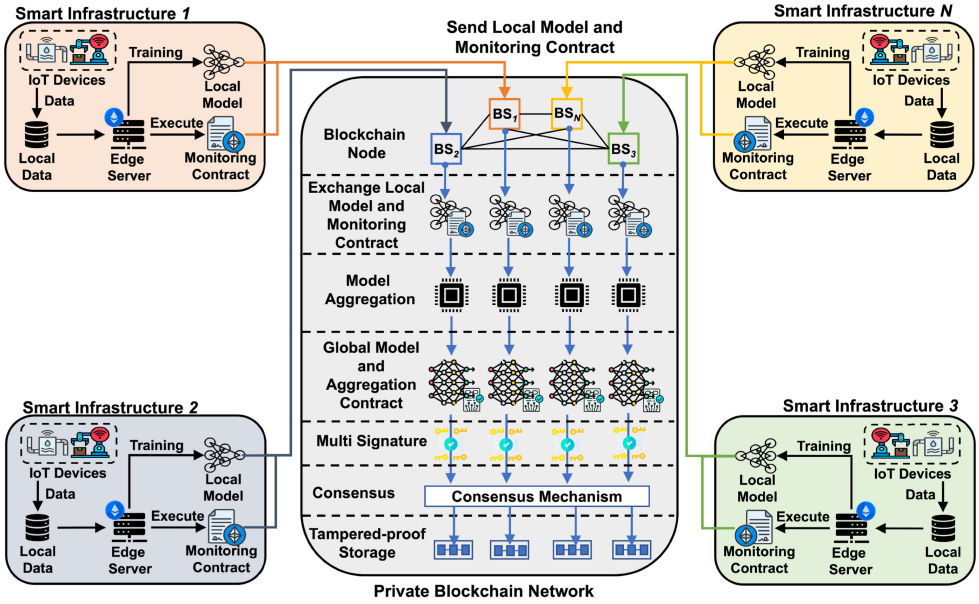

Abbildung 14 – Das FL-Framework von (Kalapaaking et al. 2024), um Korrektheit des ML-Trainings und des ML-Modell-Aggregierens mithilfe von „Smart Contracts" und einer Blockchain-Architektur zu realisieren.

🎨 *Quelle: (Kalapaaking et al. 2024).*

3. Conclusio & Weiterführende Arbeiten

Federated Learning (FL) ist ein mittlerweile praktisch-relevantes Konzept um ein ML-Modell lokal auf den jeweiligen Geräten zu trainieren. Bei einem FL-Framework, welches praktische Relevanz aufweist, muss man auf Vertrauen bzw. Trust im ganzen Zyklus einer FL-Epoche achten. Wie in der Einleitung beschrieben, gibt es sechs wichtige Vertrauenspunkte bzw. Trust-Punkte (TPs) in Trusted FL. Die ersten drei TPs adressieren primär den Aspekt der Authentizität („Authenticity"), und die letzten drei TPs adressieren primär den Aspekt der Korrektheit („Verifiability"). Zu jedem Trust-Punkt gibt es bereits vielversprechende individuelle Lösungsmethoden und teilweise konkrete Instanziierungen. Die aktuell gängigsten Methoden und konkreten Instanziierungen basieren auf dem privatsphären-bewahrenden kryptografischen Baustein Null-Wissen-Beweis („Zero-Knowledge Proof" / ZKP). Die verschiedenen individuellen Lösungs-Konzepte bieten unterschiedliche Trade-Offs; z.B. adressieren nicht alle Konzepte den Aspekt der Privatsphäre beim Aggregieren der retournierten ML-Parameter bzw. -Gewichte (siehe hierfür z.B. (Koch 2024)).

Bis dato gibt es kein Konzept welches alle Trust-Punkte gleichzeitig in einem gesamten FL-Prozess-Fluss adressiert. Dieses Gesamtkonzept von ganzheitlichem Trusted Federated Learning stellt die nächste wissenschaftliche und konstruktions-technische Herausforderungen dar. Vor allem die effiziente Kombination von digitalen Signaturen und ZKPs. Abbildung 13 zeigt den gesamten FL-Prozess-Fluss inklusive ⭐Trust-Upgrades für jeden Trust-Punkt.

Weitere bzw. andere Methoden für Trusted FL. Der Bereich von „zkFL", welcher Trusted FL via ZKPs realisiert, ist relativ neu. Die meisten relevanten Werke sind erst zwischen 2022 und 2024 publiziert worden. Es bleibt spannend zu sehen wie sich dieses neue Feld weiterentwickelt, und wichtig für, z.B., etwaige praktische Anwender am Puls der Zeit bzw. Stand der Technik zu bleiben.

Neben den Authentizitäts- und Korrektheits-Garantien via ZKPs, gibt es auch Lösungen für die verschiedenen Trust-Punkte via z.B. Blockchain. Bei der Blockchain werden Daten dezentralisiert und unwiderrufbar auf einer „Datenkette" gespeichert. Weiters ist es mit der Blockchain auch möglich, dass auf der „Datenkette" - „on-chain" via „Smart Contracts" – Berechnungen durchgeführt werden.

Z.B. (Dong et al. 2024) benutzen Blockchain-basierende „Smart Contracts" für eine sichere und verifizierte Aggregation der ML-Gewichte. In solchen Szenarien erhöht sich grundsätzlich jedoch auch der Berechnungsaufwand signifikant im FL-Prozess-Fluss. Weiters bedingt die Blockhain weitere Entitäten – wie die Blockchain-Knoten. Generell bieten Lösungsansätze via Blockchain ihre eigenen Trade-Offs. Weiters benutzen, z.B., (Kalapaaking et al. 2024) „Smart Contracts" und eine Blockchain-Architektur, um die Korrektheit (i) des ML-Trainings und (ii) des ML-Modell-Aggregierens, in einem dezentralisierten FL-Szenario zu realisieren.

Abbildung 14 zeigt die (Blockchain-)Architektur des FL-Frameworks von Kalapaaking et al. in einem FL-Szenario für smarte Infrastrukturen (wie IoT-Geräte in „Smart Factories").

Literaturverzeichnis

Bell, James, Adrià Gascón, Tancrède Lepoint, Baiyu Li, Sarah Meiklejohn, Mariana Raykova, and Cathie Yun. 2023. "{ACORN}: Input Validation for Secure Aggregation." Pp. 4805–22 in, *Usenix'23*.

Bonawitz, Kallista, Peter Kairouz, Brendan Mcmahan, and Daniel Ramage. 2022. "Federated Learning and Privacy." *Communications of the ACM* 65(4):90–97. doi: 10.1145/3500240.

Bonawitz, Keith, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2017. "Practical Secure Aggregation for Privacy-Preserving Machine Learning." Pp. 1175–91 in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, *CCS '17*. New York, NY, USA: Association for Computing Machinery.

Chaum, David, and Eugène van Heyst. 1991. "Group Signatures." Pp. 257–65 in *Advances in Cryptology — EUROCRYPT '91*, *EC'91*, edited by D. W. Davies. Berlin, Heidelberg: Springer.

Chowdhury, Amrita Roy, Chuan Guo, Somesh Jha, and Laurens van der Maaten. 2022. "EIFFeL: Ensuring Integrity for Federated Learning."

Dong, Nanqing, Zhipeng Wang, Jiahao Sun, Michael Kampffmeyer, William Knottenbelt, and Eric Xing. 2024. "Defending Against Poisoning Attacks in Federated Learning With Blockchain." *IEEE Transactions on Artificial Intelligence* 5(7):3743–56. doi: 10.1109/TAI.2024.3376651.

Kalapaaking, Aditya Pribadi, Ibrahim Khalil, Xun Yi, Kwok-Yan Lam, Guang-Bin Huang, and Ning Wang. 2024. "Auditable and Verifiable Federated Learning Based on Blockchain-Enabled Decentralization." *IEEE Transactions on Neural Networks and Learning Systems* 1–14. doi: 10.1109/TNNLS.2024.3407670.

Koch, Karl. 2024. "MPC-Based Secure Aggregation in Federated Learning: Overview, Protocols, & Google's Gboard – A-SIT Technologie."

Koch, Karl, Stephan Krenn, Tilen Marc, Stefan More, and Sebastian Ramacher. 2022. "KRAKEN: A Privacy-Preserving Data Market for Authentic Data." Pp. 15–20 in *Proceedings of the 1st International Workshop on Data Economy*, *DE '22*. New York, NY, USA: Association for Computing Machinery.

Lycklama, Hidde, Lukas Burkhalter, Alexander Viand, Nicolas Küchler, and Anwar Hithnawi. 2023. "RoFL: Robustness of Secure Federated Learning." Pp. 453–76 in *2023 IEEE Symposium on Security and Privacy (SP)*, *S&P'23*.

McMahan, Brendan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. "Communication-Efficient Learning of Deep Networks from Decentralized Data (FL Intro Paper)." Pp. 1273–82 in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. PMLR.

Wang, Zhipeng, Nanqing Dong, Jiahao Sun, William Knottenbelt, and Yike Guo. 2024. "zkFL: Zero-Knowledge Proof-Based Gradient Aggregation for Federated Learning."

Xing, Zhibo, Zijian Zhang, Meng Li, Jiamou Liu, Liehuang Zhu, Giovanni Russello, and Muhammad Rizwan Asghar. 2023. "(PZKP-FL) Zero-Knowledge Proof-Based Practical Federated Learning on Blockchain."

Zhu, Yizheng, Yuncheng Wu, Zhaojing Luo, Beng Chin Ooi, and Xiaokui Xiao. 2024. "(RiseFL) Secure and Verifiable Data Collaboration with Low-Cost Zero-Knowledge Proofs." *Proceedings of the VLDB Endowment* 17(9):2321–34. doi: 10.14778/3665844.3665860.