國立臺北科技大學
資訊工程系碩士班

碩士學位論文

漸進式委員會佔領攻擊與激勵相容防禦：區塊鏈聯邦學習的安全性研究

Progressive Committee Capture Attack and Incentive-Compatible Defense: Security Analysis for Blockchain-based Federated Learning

研究生：陸紀霖

指導教授：張世豪博士

中華民國一百一十五年一月

國立臺北科技大學
資訊工程系碩士班

碩士學位論文

漸進式委員會佔領攻擊與激勵相容防禦：區塊鏈聯邦學習的安全性研究

Progressive Committee Capture Attack and Incentive-Compatible Defense: Security Analysis for Blockchain-based Federated Learning

研究生：陸紀霖

指導教授：張世豪博士

中華民國一百一十五年一月

「學位論文口試委員會審定書」掃描檔

審定書填寫方式以系所規定為準，但檢附在電子論文內的掃描檔須具備以下條件：

1. 含指導教授、口試委員及系所主管的完整簽名。
2. 口試委員人數正確，碩士口試委員至少 3 人、博士口試委員至少 5 人。
3. 若此頁有論文題目，題目應和書背、封面、書名頁、摘要頁的題目相符。
4. 此頁有無浮水印皆可。

審定書不用頁碼

摘要

關鍵詞：區塊鏈、聯邦式學習、委員會佔領、驗證者共謀

基於區塊鏈的聯邦式學習 (BCFL) 透過去中心化共識機制解決了信任與隱私問題。現有的 BCFL 系統依賴基於委員會的驗證機制，並假設委員會成員是誠實的或擁有誠實多數。此假設容易受到驗證者共謀的威脅，攻擊者可透過累積權益 (Stake) 來主導委員會。我們識別出一種新型威脅——漸進式委員會佔領攻擊 (PCCA)，理性攻擊者利用激勵機制逐步累積權益，並佔領足夠的委員會席次以發動協同攻擊。一旦攻擊者取得委員會多數席次，現有的委員會架構便無法偵測或防範此類攻擊。為防禦 PCCA，我們提出一種挑戰增強型委員會架構，將安全性與委員會組成解耦：由小型委員會負責例行驗證以提供活性 (Liveness)，而由全域共識支持的挑戰機制提供安全性保證。任何惡意聚合行為都將觸發密碼學驗證、罰沒懲罰，並立即移除惡意驗證者——無論其在委員會中的席次多寡。此機制將安全門檻從委員會多數轉移至全網共識，從而瓦解委員會佔領攻擊。實驗結果顯示，當攻擊發生時，本機制能完全清除惡意委員會成員，而現有最先進的方法則允許攻擊者取得委員會完全控制權並執行不受制衡的攻擊。我們的解耦設計亦允許更小的委員會規模，在不犧牲安全性的前提下提升運算效率。

ABSTRACT

Keyword: Blockchain, Federated Learning, Committee Capture, Verifier Collusion

Blockchain-based Federated Learning (BCFL) addresses trust and privacy concerns through decentralized consensus. Current BCFL systems rely on committee-based validation assuming honest or honest-majority committees. This assumption is vulnerable to verifier collusion, where attackers accumulate stake to dominate committees. We identify Progressive Committee Capture (PCC), a novel threat where rational attackers exploit incentive mechanisms to gradually accumulate stake and capture sufficient committee seats for coordinated attacks. Existing committee-based architectures cannot detect or prevent such attacks once attackers achieve committee majority. To defend against PCC, we propose an Challenge-Augmented Committee Architecture that decouples security from committee composition: a small committee provides liveness through routine validation, while a challenge mechanism backed by global consensus provides security guarantees. Any malicious aggregation triggers cryptographic verification, slashing penalties, and immediate removal of malicious validators—regardless of their committee representation. This shifts the security threshold from committee majority to global network consensus, neutralizing committee capture attacks. Experimental results demonstrate complete elimination of malicious committee members upon attack attempts, while state-of-the-art approaches allow attackers to achieve full committee control and execute unchecked attacks. Our decoupled design also enables smaller committee sizes, improving computational efficiency without compromising security.

誌謝

所有對於研究提供協助之人或機構，作者都可在誌謝中表達感謝之意。

目錄

圖目錄

表目錄

第一章　緒論 (Introduction)

隨著人工智慧與分散式運算技術的進步，區塊鏈賦能的聯邦學習 (Blockchain-based Federated Learning, BCFL) 已成為解決多方互不信任情境下協作機器學習的核心技術路徑。在諸如低軌衛星網路 (LEO) [1, 2, 3]、車聯網 (V2X) [4, 5, 6] 以及工業物聯網 (IIoT) [7, 8, 9] 等實際應用場景中，BCFL 展現了其不可替代的重要性。特別是以 LEO 衛星星座為代表的太空 AI 應用場景，星地通訊窗口通常僅約 5 分鐘，且下行頻寬受限於 8Mbps 左右 [2]，使得依賴地面站聚合的傳統模型訓練方案難以實施。BCFL 通過在異質衛星營運商間建立去中心化信任層，成功將收斂時間減少達 30 小時 [3]。同樣地，在工業 4.0 的背景下，BCFL 允許協作工廠在不洩露商業機密的前提下進行預測性維護，實驗資料顯示其通訊開銷可較集中式架構減少約 41% [7]。這些場景共同呈現出「無可信中心」、「資源受限」與「資料高度異質」的特徵，促使 BCFL 成為通用去中心化學習架構的首選方案。

然而，BCFL 在邁向大規模部署時面臨著嚴峻的效率瓶頸，這在業界被稱為「可擴展性兩難」。目前絕大多數 BCFL 系統採用 PBFT (Practical Byzantine Fault Tolerance) 或其變體作為共識機制，其 $O(n^2)$ 的訊息複雜度在節點數增加時會導致效能急劇下降。根據 FLCoin [10] 的實證研究，當參與節點數達到 100 個時，單輪共識產生的訊息量將超過 20,000 條，導致共識延遲攀升至 25 秒以上，此延遲水平已達到模型訓練時間的量級。在極端的車載網路 (VANET) 實測中，100 輛車進行 BCFL 協作會產生 360.57 MB 的巨大資料量，單輪訓練的總通訊開銷高達 19.51 秒 [11]。此外，區塊鏈節點對儲存的高需求 (如比特幣需 200GB，乙太坊超過 465GB) 與邊緣設備 KB 至 MB 級的有限記憶體形成強烈衝突 [12]。這種效能與資源的雙重束縛，使得全節點驗證的傳統架構在實際工業部署中顯得難以維繫。

為了解決上述可擴展性挑戰，學界近年來轉向研究「委員會機制 (Committee Mechanism)」，其核心思想是將驗證責任從全體節點縮減至一組小型驗證者委員會。目前主流的選拔機制包含基於雜湊環的隨機抽樣 [13]、基於幣齡或權益的權重選舉 [14, 10] 以及基於預言機 (VRF) 的 Sortition 機制 [15, 16]。委員會機制的引入立竿見影地改善了系統效能：FLCoin [10] 通過滑動窗口選舉將通訊開銷降低了 90%，並實現了 5.7 倍的訓練加速；BFLC [14] 則利用委員會驗證成功將共識延遲穩定在 3 秒以內。這些最佳化雖成功將通訊複雜度降至與委員會規模 C 相關的 $O(C^2)$ 或 $O(C)$。然而，這種為了效率而進行的「算力與權力集中」也同時引入了新的、尚未被充分探討的安全攻擊面。

最令學界擔憂的危機在於現有委員會防禦機制對「誠實多數假設 (Honest Majority Assumption)」的過度依賴。根據 2024 年針對拜占庭強健聯邦學習的全面調查 [17, 18]，目前超過 93.3% 的 BCFL 研究雖部署了 Krum、Trimmed Mean 或 Median 等防禦演算法，但皆隱含地假設執行這些演算法的實體 (即委員會成員) 是絕對誠實的。現有的威脅模型大多只考慮惡意客戶端上傳毒化梯度，卻忽略了「理性驗證者 (Rational Verifiers)」的危害。最新研究指出，理性對手可以先透過合法行為積累聲譽，一旦在委員中取得超過 33% (針對 BFT 系統) 或 50% (針對一般投票系統) 的主導權，即可輕易繞過所有強健

聚合演算法，甚至偽造聚合結果而不受懲罰。BlockDFL [13] 與 FedBlock [19] 等前沿工作亦坦言，現有機制無法抵禦具備長期策略的委員會共謀攻擊。

上述現象揭示了一個關鍵的「研究缺口 (Research Gap)」：現有 BCFL 缺乏應對「漸進式委員會佔領攻擊 (Progressive Committee Capture Attack, PCCA)」的自癒機制。在 PCCA 中，對手並非採取暴力破壞，而是實施「策略性餓死 (Strategic Starvation)」——即在掌控委員會後，優先打包與自身利益相關的更新，並拒絕為誠實參與者提供驗證服務，從而操縱獎勵分配與權益動態。由於缺乏事後的「可追溯審計」與「有效威懾」，一旦誠實多數假設在某一輪次被攻破，系統權力將產生雪崩式的中心化。現有的基於同態加密或權益證明的方案雖然能保護隱私，卻無法在委員會本身已不再可信的情況下，保證模型更新的正確性與資源分配的公平性。如何解耦安全性與共識節點集體信用，成為實現真正去中心化 AI 平台的最後一哩路。

針對這一挑戰，本文提出了一種「挑戰者增強委員會架構 (Challenge-Augmented Committee Architecture, CACA)」，旨在為 BCFL 引入一種全新的安全性保險機制。本研究提出的核心思想是「即時執行、異步審計、罰沒威懾」，這與傳統的「先驗證、後提交」模式有本質區別。我們的主要創新點在於將系統的「活性 (Liveness)」與「安全性 (Security)」進行解耦：即使在委員會不完全可信、甚至被捕獲的情況下，系統仍能通過去中心化的挑戰者網路來檢舉委員會的錯誤決策。具體貢獻概括如下：

- 我們首次定義並模擬量化了漸進式委員會佔領攻擊對 BCFL 長期激勵相容性的破壞力。
- 我們提出了一套基於博弈論設計的「內部罰沒 (Internal Slashing)」協議，確保審計成本低於作惡罰金，從而使得誠實行為成為理性節點的納什均衡。
- 實驗結果顯示，在 30% 惡意共謀的極端環境下，本框架仍能維持超過 98.6% 的模型準確率，並成功將受擊頻率降低約 80%。在 100 節點規模的實驗中，本機制在相同的安全性水平下將日常通訊開銷降低了 44.4%，並將系統最低不可用率從 20% 壓制至 5% 以下。

本論文的組織結構編排如下：第一章為緒論，闡明研究動機、目標與貢獻。第二章介紹聯邦學習、區塊鏈底層架構、拜占庭容錯技術等背景知識，並對現有的去中心化聯邦學習文獻進行分類與批判性評述。第三章定義本研究的系統模型與 PCCA 攻擊者的行為特徵，深入分析其威脅模型。第四章詳細描述 CACA 的具體設計流程、協議設計及安全分析。第五章呈現模擬實驗的參數設定與效能對比結果，驗證所提架構的有效性。第六章對全論文進行總結，並探討本研究在未來的應用前景。

第二章　背景知識與相關研究

本章旨在建立理解本研究所需之技術基礎，並對現有區塊鏈聯邦學習（BCFL）的驗證機制進行系統性回顧與批判性分析。首先介紹聯邦學習的基本原理及其面臨的拜占庭威脅，接著探討區塊鏈技術如何重塑聯邦學習的信任模型。在相關研究部分，本章將深入剖析當前主流的密碼學驗證（zkML）、樂觀驗證（opML）與委員會共識機制，運用具體數據論證現有方案在安全性、效率與通用性上的結構性權衡。最後，本章定義了本研究所採用的基準系統模型，為後續的威脅建模與防禦設計奠定基礎。

2.1　聯邦式學習與拜占庭威脅 (Federated Learning and Byzantine Threats)

聯邦學習（Federated Learning, FL）是由 McMahan 等人於 2017 年正式提出之分散式機器學習框架 [20]。其核心目標在於多個參與方（Clients）協同訓練模型，而無需將原始資料集中於中央伺服器，從而解決資料隱私與孤島問題。

2.1.1　聯邦式學習基礎 (Fundamentals of Federated Learning)

在標準聯邦學習架構中，目標是最小化全域損失函數 $F(w)$：

$$\min_{w} F(w) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(w) \tag{2.1}$$

其中 K 為參與客戶端總數，n_k 為第 k 個客戶端之本地樣本數，$F_k(w)$ 為其本地損失函數。

經典的 *FederatedAveraging* (FedAvg) 演算法透過週期性地收集客戶端模型更新 w_{t+1}^k，並在伺服器端進行加權聚合：

$$w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k \tag{2.2}$$

此方法相較於同步隨機梯度下降（SGD）可顯著減少通訊開銷，但其安全性建立在中央聚合器完全誠實且客戶端皆非惡意的假設之上。

2.1.2　拜占庭攻擊模型 (Byzantine Attack Models)

在分散式環境中，系統必須面對拜占庭故障（Byzantine Fault）。根據 Blanchard 等人的定義，拜占庭節點可發送任意、潛在惡意的更新，並可能與其他惡意節點共謀。聯

邦學習中的攻擊主要分為兩類：

1. **資料投毒 (Data Poisoning)**：攻擊者汙染本地訓練資料（如標籤翻轉），導致模型學習錯誤的特徵。

2. **模型投毒 (Model Poisoning)**：攻擊者直接操控上傳的梯度或模型參數。研究顯示，模型投毒比資料投毒更具威脅性。例如，Bagdasaryan 等人提出的模型替換攻擊（Model Replacement Attack）[21] 可在單一輪次內植入後門，並保持主任務的高準確率。

2.1.3 傳統拜占庭容錯聚合 (Traditional Byzantine-Robust Aggregation)

為抵禦拜占庭攻擊，學界提出多種強健聚合演算法（Robust Aggregation Rules）：

- **Krum 及其變體** [22]：基於幾何距離選擇最接近多數節點的更新。Krum 選擇一個更新 u^*，使得其與最近 $n - f - 2$ 個鄰居的歐式距離平方和最小。

- **裁剪均值 (Trimmed Mean)** [23]：在每個維度上移除最大與最小的 β 比例數值後取平均，能有效抵禦統計極端值。

- **座標中位數 (Coordinate-wise Median)** [23]：取每個維度的中位數，具有高崩潰點（Breakdown Point）。

然而，這些防禦機制存在一個關鍵的局限性：**誠實聚合者假設**。如果不誠實的聚合器（Server）控制了聚合過程，它可以故意忽略防禦規則，甚至與惡意客戶端共謀。這構成了傳統聯邦學習的單點信任危機。

2.2 區塊鏈聯邦式學習 (Blockchain-based Federated Learning, BCFL)

區塊鏈聯邦式學習（BCFL）透過將分散式帳本技術（Distributed Ledger Technology, DLT）引入聯邦學習架構，從根本上重塑了多方協作訓練的信任模型。

2.2.1 技術動機：從中心化信任到去中心化共識

2.2.1.1 傳統聯邦學習的信任集中化困境

儘管聯邦學習承諾「數據不出本地」，其標準架構仍高度依賴單一中央聚合器（Central Aggregator），這種中心化設計引入了三類關鍵的信任風險。首先是**聚合器的誠**

實性風險：由於缺乏外部監督，中央伺服器可能執行選擇性聚合，故意排除特定客戶端的更新以操縱模型表現，甚至如同 Geiping 等人 [24] 指出，惡意伺服器可對梯度執行反演攻擊（Gradient Inversion Attack），從更新中重建原始訓練影像。其次是**單點故障（Single Point of Failure, SPOF）**：中央伺服器的可用性直接決定了整個系統的穩定性。最後是**拜占庭容錯能力的缺乏**：在惡意客戶端佔比超過 50% 或中央伺服器本身遭入侵的情況下，傳統的穩健聚合演算法將失效。

2.2.1.2 區塊鏈技術的解決方案

區塊鏈技術的引入為上述問題提供了結構性的解方。**不可篡改性（Immutability）**確保了所有模型更新與聚合結果一旦上鏈便無法被回溯修改，為系統提供了可信任的審計軌跡。**智能合約的透明性（Transparency）**將聚合規則與客戶端選擇邏輯代碼化，使得所有參與者皆能驗證聚合過程的正確性。**去中心化架構（Decentralization）**則通過點對點網路（P2P）取代了中央節點，利用共識機制（Consensus Mechanism）確保在部分節點失效或作惡的情況下，系統仍能維持運作並達成資料一致性。

2.2.2 BCFL 架構的演進概況

BCFL 的發展歷程反映了學界在去中心化程度、通訊效率與安全性三者之間的權衡。早期的研究如 BlockFL [25] 採用工作量證明（PoW）與全節點驗證，雖實現了極致的去中心化但帶來了難以承受的計算負擔與延遲。隨後，為了克服擴展性瓶頸，基於委員會（Committee-based）的架構逐漸成為主流，如 BFLC [14] 與 BlockDFL [13]，透過選舉小型驗證委員會來降低通訊複雜度，這也是本研究主要的探討對象。

2.2.3 智能合約的功能職責

在 BCFL 系統中，智能合約扮演著自動化管理者的角色，其功能通常劃分為四個核心模組：**註冊模組**維護參與者身分與權益；**聚合模組**協調訓練輪次與參數收集；**驗證模組**執行校驗邏輯（如準確率測試）以過濾惡意更新；**獎勵模組**則依據貢獻度自動分配代幣激勵，並對被偵測到的惡意行為執行罰沒（Slashing）。

2.3 拜占庭容錯機制 (Byzantine Fault Tolerance Mechanisms)

分散式系統在面對節點故障時，必須具備持續運作的能力。當故障涉及節點發送矛盾訊息或與其他惡意節點共謀時，系統便需要更強韌的拜占庭容錯機制。

2.3.1 拜占庭將軍問題與容錯閾值

拜占庭將軍問題由 Lamport 等人 [26] 提出，定義了在存在叛徒的情況下如何達成一致性。Lamport 證明了在僅使用口頭訊息的情況下，系統若要達成拜占庭容錯，總節點數 n 與惡意節點數 f 必須滿足 $n \geq 3f + 1$。此理論極限確立了 PBFT 等共識協議的安全性邊界。

2.3.2 實用拜占庭容錯協議 (PBFT)

Castro 與 Liskov 於 1999 年提出的 PBFT 協議 [27] 首次將 BFT 共識的通訊複雜度從指數級降至多項式級別 $O(n^2)$。協議透過 Pre-prepare、Prepare 與 Commit 三個階段達成共識。當節點收集到 $2f + 1$ 個匹配的 Commit 訊息時，即確認達成共識。在聯盟鏈環境下，PBFT 提供了堅實的信任基礎，特別適用於本研究所探討的委員會共識場景。

2.4 相關研究與盲點分析 (Related Work and Blind Spot Analysis)

現有的區塊鏈聯邦學習驗證方法可大致分為兩大類：基於密碼學證明的驗證方法與基於委員會的共識方法。前者追求數學上可證明的計算正確性，但往往面臨嚴重的效能瓶頸；後者則透過經濟激勵與多數決達成共識，但高度依賴誠實多數假設。本節將系統性分析這些方法的技術原理、效能數據與固有局限，以釐清本研究的學術定位。

2.4.1 基於密碼學驗證的方法 (Cryptographic Verification Methods)

2.4.1.1 零知識機器學習 (zkML) 的計算瓶頸

零知識機器學習（zkML）試圖透過將機器學習計算轉換為算術電路，使驗證者無需重新執行模型即可確認計算的正確性與數據隱私 [28]。其技術堆疊通常包括 Groth16（以約 200 bytes 的極小證明著稱）、PLONK（具備通用可更新設置特性）[29] 與 zk-STARK（無需信任設置且具量子抗性）。轉換過程需經歷三個高成本階段：首先將浮點數量化為有限域整數，接著將每個運算分解為多項式約束，最後生成密碼學證明。

然而，約束數量隨深度學習模型的複雜度呈指數級膨脹。根據 ZEN 編譯器的基準測試 [30]，一個簡單的 ShallowNet-MNIST 模型需要 4.31 百萬（4.31M）個約束，而稍大的 LeNet-Face-large-ORL 則暴增至 2.63 億（263M）個約束。Chen 等人在 EuroSys 2024 的實驗數據進一步揭示了其效能瓶頸 [28]：ResNet-18 模型的證明生成需耗時 52.9 秒，VGG16 需要 637 秒，而 DistillGPT-2 更高達 3,651 秒（約一小時），且在生成過程中需要高達 1TB RAM 的高規格硬體支援。儘管框架效能持續優化——例如 EZKL 相比 RISC

Zero 快了 65.88 倍且記憶體使用減少 98.13% [31]——但對於需要頻繁迭代的聯邦學習而言，此類開銷仍難以負荷。

此外，zkML 的核心局限在於難以支援複雜的拜占庭容錯聚合演算法。Krum 與 Multi-Krum 等演算法需計算所有客戶端更新間的成對距離，這將產生 $O(n^2 \cdot d)$ 的約束爆炸；排序與中位數運算在零知識電路中亦極度昂貴。現有 zkFL 方案如 RiseFL [32] 僅能支援 L2-norm 等簡單有效性檢查，將密碼學成本從 $O(d)$ 降至 $O(d/\log d)$，但仍無法實現完整的距離計算防禦。因此，zkML 雖然提供了強大的密碼學安全性，卻以犧牲聚合演算法的通用性與系統效率為代價。

2.4.2 基於樂觀驗證的方法 (Optimistic Verification Methods)

2.4.2.1 opML 的架構限制與延遲挑戰

樂觀機器學習（optimistic Machine Learning, opML）採用「預設正確」的執行模式，僅在爭議發生時才啟動驗證程序 [33]。其運作流程為：服務提供者於鏈下執行機器學習推論並提交結果，驗證者在規定的挑戰期內可發起欺詐證明（Fraud Proof），透過二分協議（Bisection Protocol）逐步縮小爭議範圍至單一計算步驟，最終由鏈上的 FPVM（欺詐證明虛擬機）進行仲裁。ORA Protocol 作為首個開源 opML 實現，已能支援 LLaMA 2 等 70 億參數模型直接於以太坊環境下運行 [34]。

挑戰期的設計反映了安全性與效率的根本權衡。為了確保驗證者有充足時間偵測並提交證明，同時容納網路延遲與潛在的審查攻擊，主流 Optimistic Rollup 協議如 Optimism 與 Arbitrum 分別採用了 7 天與 6.4 天的挑戰期 [35]。這種長週期的確認機制與聯邦學習的需求存在本質衝突——聯邦學習依賴快速的迭代更新與聚合，若每一輪訓練皆需等待數天的挑戰期，將使模型訓練完全不可行。

此外，opML 基於 AnyTrust 假設（即系統中至少存在一個誠實驗證者），這與 BCFL 的多方共識需求存在差異。opML 設計初衷為處理單一提交者與單一挑戰者間的爭議，而非多方參與者間的複雜共識。加之 FPVM 的記憶體限制需採用延遲載入設計，當聯邦學習模型涉及大量參與者的梯度更新時，可能超出其實際處理能力。雖然部分方案試圖整合 zkML 元件以增強隱私，但仍受限於單一證明者架構，無法滿足 BCFL 對於多驗證者去中心化共識的需求。

2.4.3 基於委員會共識的方法 (Committee-Based Consensus Methods)

相較於上述驗證方法，基於委員會的共識機制在效率與實用性之間取得了較佳的平衡，因而成為當前 BCFL 的主流選擇。現有研究主要在委員會的「選舉機制」上進行創新。

2.4.3.1 FLCoin：基於滑動窗口的動態選舉

FLCoin 提出了一種基於滑動窗口（Sliding Window）的動態委員會選舉機制 [36]，將聯邦學習的貢獻歷史作為委員會成員資格的依據。每個有效的更新區塊代表一份成員資格，窗口大小固定為 s，隨著新區塊的生成而滑動更新。節點的貢獻值 C_k 由訓練數據規模 $|D_k|$ 決定，貢獻值最高者將成為委員會領導者。

根據其安全性分析，在網路規模 $n = 500$、惡意節點比例 $\leq 25\%$、窗口大小 $s = 100$ 的條件下，委員會內惡意節點數不超過容錯閾值的機率可達 98.4%；若將 s 提升至 150，此機率更可提升至 99.8%。在效能方面，FLCoin 透過此機制相較於傳統 PBFT 實現了 90% 的通訊開銷降低與 35% 的訓練時間縮短。在 100 個節點的配置下，其共識延遲僅需 3.05 秒。然而，FLCoin 的滑動窗口設計未建立權益衰減機制，且依賴「預定義的可信管理者群組」來維護身分鏈，引入了潛在的中心化風險。更關鍵的是，該研究未考慮惡意節點採取長期策略，透過持續參與逐步累積權益以控制委員會的情境。

2.4.3.2 BlockDFL：基於權益加權的雙層驗證

BlockDFL 採用了完全去中心化的點對點架構 [13]，透過區塊雜湊值與權益加權（Stake-weighted）實現委員會選舉的隨機性與可驗證性。其核心假設為「持有大量權益的參與者傾向誠實行為，因為他們能從貨幣獎勵中獲益更多」。該系統設計了雙層評分機制：第一層由聚合者進行本地推論篩選，第二層由驗證者使用 Krum 演算法過濾異常值。

BlockDFL 宣稱能容忍高達 40% 的惡意參與者，優於多數現有框架。其獎勵連鎖機制確保了權益能均等地分配給貢獻者。與 FLCoin 的關鍵差異在於，BlockDFL 的選舉基礎是經濟權益而非純粹的貢獻次數。這雖然提高了攻擊門檻，但也引入了新的攻擊面：國家級攻擊者或競爭對手可能願意承擔經濟損失（Sacrifice Stake）來破壞模型。此外，該機制同樣缺乏防止 Sybil 攻擊者跨多重身份逐步累積權益的防禦手段。

2.4.3.3 BFLC 與其他信譽機制

BFLC 開創性地將委員會共識引入 BCFL [14]，採用雙區塊儲存設計並利用 K-fold 交叉驗證評估更新品質。然而，BFLC 面臨嚴重的冷啟動問題（Cold Start Problem）——新加入節點因缺乏歷史數據而難以建立信任，導致系統容易被早期進入的惡意節點掌控。後續研究如 VBFL [37] 與 VFChain [38] 雖引入了準確度差異（VAD）指標與可審計的雙跳鏈結構來增強安全性，但仍共同面臨 50% 拜占庭閾值的硬性限制以及資源受限裝置的計算負擔問題。

2.4.4 現有方法的系統性局限與研究缺口

2.4.4.1 安全性、效率與通用性的三難困境

綜合上述分析，揭示了現有方法在「安全性-效率-通用性」三個維度上的 Pareto 前沿權衡（Trade-off）。如表 2.1 所示，zkML 提供了最強的密碼學安全性，無需依賴誠實假設，但其證明生成時間隨模型規模呈超線性增長，且無法支援 Krum 等複雜聚合算法。opML 透過經濟激勵大幅降低了計算成本，但其長達數天的挑戰期與單一證明者架構使其難以應用於即時性要求高的聯邦學習場景。委員會方法在效率與實用性間取得了折衷，但均依賴某種形式的誠實多數假設。

表 2.1: 現有 BCFL 驗證與共識機制之比較分析

方案類型	代表技術/系統	安全性保證來源	效率 (典型延遲)	聚合通用性
密碼學驗證	zkML (Groth16, PLONK)	數學證明 (無信任假設)	分鐘至小時 (極慢)	僅限簡單聚合 (如 FedAvg)
樂觀驗證	opML (Optimism, ORA)	經濟賽局 (AnyTrust)	7 天挑戰期	受限於 FPVM 實作
委員會共識	FLCoin	統計機率 (98.4%)	3.05 秒	支援多種聚合
	BlockDFL	經濟權益 (40% 容錯)	<3 秒	支援 Krum/Median
	BFLC	聲譽積累 (50% 閾值)	中等	支援

2.4.4.2 關鍵缺口：動態累積攻擊的防禦真空

更為關鍵的是，所有現有的委員會方案均未充分處理**長期權益累積（Long-term Stake Accumulation）**導致的委員會滲透風險。無論是 FLCoin 的滑動窗口或 BlockDFL 的權益加權，均採用靜態的安全性分析模型（如超幾何分佈），假設攻擊者的資源或權益比例是固定的。然而，在實際賽局中，理性的攻擊者會採取**漸進式委員會佔領（Progressive Committee Capture）**策略：先在潛伏期表現誠實以獲取合法獎勵，待累積了足夠的權益或聲譽後，再於關鍵時刻發動攻擊。

現有文獻存在三個核心缺口：

1. **缺乏動態安全性分析**：忽略了攻擊者透過時間維度累積影響力的動態過程。
2. **驗證效率與通用性的矛盾**：尚無方案能在保持聚合演算法通用性（如支援 Krum）的同時，提供接近密碼學等級的安全性。
3. **缺乏抗累積機制**：現有權益證明機制多設計為線性累積，賦予了早期參與者或長期潛伏者過大的系統控制權，缺乏類似「權益衰減」或「動態輪換」的防禦設計。

本研究旨在透過引入時間衰減權益函數與動態挑戰機制，填補上述缺口，在維持委員會機制的高效率同時，顯著提升對抗長期滲透攻擊的安全性。

2.5 系統模型與前置定義 (System Model and Preliminaries)

本節定義本研究所採用的基準系統模型。此模型基於 BlockDFL 委員會架構並進行擴展，作為後續威脅分析與防禦設計的基礎。

2.5.1 網路模型

本研究考慮一個去中心化的區塊鏈聯邦學習系統，由以下三種核心角色構成：

1. **Update Providers (UP)**：原為客戶端 (Clients)，集合記為 $\mathcal{U} = \{u_1, u_2, ..., u_N\}$。每個 Update Provider 持有本地私有資料集 $\mathcal{D}_i$，負責在本地進行模型訓練並提交更新。

2. **Aggregators (AG)**：集合記為 $\mathcal{A} = \{a_1, a_2, ..., a_K\}$。負責收集 UP 的更新，執行初步聚合生成提案。Aggregator 的選擇基於權益。

3. **Verifier Committee (VC)**：集合記為 $\mathcal{V} = \{v_1, v_2, ..., v_M\}$。Verifiers 組成委員會，負責驗證 Aggregator 的提案。委員會成員通過共識機制批准提案並上鏈。

2.5.2 聚合與共識流程

在每個訓練輪次 r，系統執行以下流程：

1. **本地訓練**：UP 訓練 model update Δw_i 並發送給 AG。

2. **初步聚合**：AG 生成聚合更新 Δw_{agg} 並提交提案交易。

3. **委員會驗證**：委員會 $\mathcal{V}_r$ 執行驗證邏輯（如 Krum 檢驗）。

4. **共識決策**：委員會通過 BFT 共識對提案投票。

5. **獎勵分配**：若提案通過，AG、UP 與投票贊成的 Verifiers 共同瓜分系統獎勵。

2.5.3 權益動態與攻擊面

權益（Stake）在系統中扮演核心角色，既是選擇權重的依據，也是經濟獎勵的來源。這種「贏家通吃」的正反饋特性雖然激勵了誠實行為，但也創造了攻擊面：若攻擊者能策略性地累積權益，便能逐步掌控委員會。與傳統 PoS 不同，BCFL 中的攻擊者不僅能破壞共識，還能透過投毒模型永久損害全域模型的效能，且難以被傳統 BFT 機制偵測。

第三章　威脅模型 (Threat Model)

基於第二章定義的委員會架構系統模型（詳見 2.5 節），本章分析該架構在面對理性攻擊者時的安全脆弱性。本章重點定義「漸進式委員會佔領攻擊」(Progressive Committee Capture Attack, PCCA)，揭示攻擊者如何利用權益機制的正反饋特性逐步實現網路控制權的轉移，並為後續章節的防禦機制設計提供明確的安全目標。

3.1　攻擊者模型

3.1.1　攻擊者類型：理性攻擊者

本研究考慮的攻擊者為理性攻擊者 (Rational Adversary)，而非傳統的拜占庭攻擊者。兩者的關鍵區別在於動機：

- 拜占庭攻擊者：以破壞系統為目標，可能採取任意惡意行為，即使損害自身利益也在所不惜。
- 理性攻擊者：以利益最大化為目標，僅在預期收益大於成本時才發動攻擊。如果攻擊的預期收益為負，理性攻擊者不會嘗試作惡。

這種區分至關重要，因為它為基於博弈論的防禦機制提供了理論基礎。如果能夠設計激勵機制，使得攻擊的預期收益為負，理性攻擊者將自發地選擇誠實行為，無需依賴傳統的多數誠實假設。

3.1.2　攻擊者目標

理性攻擊者的主要目標包括：

- 經濟利益：通過操縱委員會，獨佔訓練獎勵，排擠誠實節點的收益。
- 權益壟斷：通過阻止誠實節點的權益增長，逐步提高自身在系統中的權益佔比，最終控制委員會選擇過程。

- 網路控制：當攻擊者的權益佔比足夠高時，可以持續控制委員會，進而控制整個聯邦學習過程，包括模型更新的接受與拒絕。

值得注意的是，理性攻擊者的目標不僅僅是短期的經濟收益，更重要的是長期的網路控制權。這種攻擊不同於傳統的模型投毒攻擊，後者僅影響模型品質，而前者則從根本上顛覆了去中心化系統的安全假設。

3.1.3 攻擊者能力

本研究假設攻擊者具有以下能力：

- 節點控制：攻擊者可以控制系統中一定比例的驗證者節點，記為 f。在典型場景下，假設 $f \leq 0.3$，即攻擊者控制不超過 30% 的節點。
- 協同作惡：被攻擊者控制的節點可以相互協調，共同執行攻擊策略。例如，當多個惡意節點同時被選入委員會時，它們可以串通一致地投票。
- 策略性行為：攻擊者可以根據系統狀態動態調整策略。例如，在權益較低時表現誠實以積累信譽，在獲得委員會多數席位時發動攻擊。
- 觀察能力：攻擊者可以觀察區塊鏈上的所有公開資訊，包括其他節點的權益、歷史行為、委員會組成等，並據此制定攻擊策略。

3.1.4 攻擊者限制

同時，攻擊者受到以下限制：

- 密碼學限制：攻擊者無法破解密碼學原語，無法偽造其他節點的簽名或篡改已上鏈的資料。
- 網路限制：攻擊者無法控制全網多數節點，無法單獨發動 51% 攻擊。
- 經濟約束：攻擊者受經濟激勵約束，如果攻擊的預期成本大於收益，理性攻擊者不會嘗試攻擊。

- 可驗證性：攻擊者無法阻止其他節點驗證聚合結果的正確性。任何節點都可以重新運算演算法，檢測委員會是否正確執行協議。

3.2 攻擊向量分析

區塊鏈聯邦學習系統面臨多層次的安全威脅。本節分析不同層次的攻擊向量，並說明本研究的關注目點。

3.2.1 資料層攻擊：已有防禦

資料層攻擊主要指惡意客戶端通過投毒攻擊破壞模型品質：

- 資料投毒 (Data Poisoning)：惡意客戶端在本地訓練時使用被污染的資料集，導致訓練出的模型更新偏離正常分佈。
- 模型投毒 (Model Poisoning)：惡意客戶端直接構造惡意的模型更新，而非基於真實訓練過程。

針對這類攻擊，現有研究已提出多種拜占庭強健演算法，如 Krum、Trimmed Mean、Median 等。這些演算法通過統計方法識別並過濾異常更新，在一定比例的惡意客戶端存在時仍能保證模型收斂。

然而，這些防禦方法存在一個關鍵假設：執行演算法的驗證者是誠實的。如果驗證者本身是惡意的，它們可以選擇不執行這些防禦演算法，或者篡改演算法的執行結果，從而使資料層的防禦完全失效。

3.2.2 共識層攻擊：本研究重點

共識層攻擊針對的是執行聚合和驗證的委員會本身：

- 驗證者共謀 (Verifier Collusion)：多個惡意驗證者協同作惡，共同通過惡意的聚合結果。

- 委員會佔領 (Committee Capture)：攻擊者通過操縱委員會選擇機制，逐步增加惡意節點在委員會中的佔比，最終控制委員會。

如第三章文獻分析所示，現有區塊鏈聯邦學習研究存在系統性的「驗證層盲點」：約 93% 的研究假設驗證者是誠實的或滿足誠實多數，僅有極少數研究 (如 KFC) 明確考慮惡意驗證者的場景。

此外，現有的 BlockDFL 類論文雖然引入了 Verifier 機制，但大多假設 Aggregator 和 Verifier 之間是獨立的，或者假設 Verifier 是誠實的。本研究指出了 Verifier 和 Aggregator 可以是同一利益集團 (Cartel) 的風險，即攻擊者可能同時控制委員會與聚合節點，這是對現有 BlockDFL 架構安全分析的重要補充。

本研究聚焦於共識層攻擊，特別是委員會佔領攻擊。這種攻擊的危險性在於：

- 繞過資料層防禦：惡意委員會可以直接接受惡意更新，無需執行 Krum 等防禦演算法。
- 隱蔽性強：攻擊者在初期表現誠實，不易被檢測，等到權益足夠高時才發動攻擊。
- 自我強化：一旦攻擊成功，攻擊者的權益會進一步增加，形成正反饋，使得攻擊越來越容易。

3.2.3 攻擊層次對比

表 3.1 對比了不同層次攻擊的特徵與現有防禦情況。

表 3.1: 攻擊層次對比

攻擊層次	攻擊者	攻擊目標	現有防禦	防禦假設	本研究關注
資料層	惡意客戶端	模型品質	Krum, Trimmed Mean	驗證者誠實	否
共識層	惡意驗證者	網路控制	誠實多數假設	多數驗證者誠實	是

從表中可以看出，資料層攻擊已有成熟的防禦方法，但這些方法依賴於驗證者誠實執行的假設。相比之下，共識層攻擊的防禦仍依賴於誠實多數假設，缺乏針對理性攻擊者的激勵相容機制。

Algorithm 1 High-Level Strategy of Progressive Committee Capture Attack (PCCA)

Require: Current Committee $\mathcal{V}$, Adversary Controlled Nodes $\mathcal{C}_{adv}$
Ensure: Action for the current round
Check Phase: Calculate control ratio $r = \frac{|\mathcal{V} \cap \mathcal{C}_{adv}|}{|\mathcal{V}|}$
if $r \leq 2/3$ **then**
    // State 1: Shadow Mode (Lurking)
    Follow the protocol honestly to accumulate stake and await majority.
else
    // State 2: Capture Mode (Occupying)
    if Aggregator is Adversarial then
        Full Stack Poisoning: Force approve malicious proposal.
    else
        Strategic Starvation: Force reject honest proposal.
    end if
end if

3.3 漸進式權益佔領攻擊 (Progressive Committee Capture Attack)

本節詳細定義本研究針對的核心威脅：漸進式權益佔領攻擊 (Progressive Committee Capture Attack, PCCA)。這是一種針對基於權益的委員會選擇機制的隱蔽性攻擊，通過兩階段策略逐步實現網路控制。

3.3.1 攻擊定義

漸進式權益佔領攻擊是指攻擊者通過以下策略，逐步增加其在系統中的權益佔比，最終控制委員會選擇過程：

1. 潛伏階段：攻擊者在初期表現誠實，提交正常的模型更新，積累權益與信譽。
2. 佔領階段：當攻擊者在委員會中獲得超過 2/3 席位時，啟動「戰略性餓死」策略，拒絕打包誠實節點的更新，獨佔獎勵。
3. 權益壟斷：由於誠實節點無法獲得獎勵，其權益停滯；而惡意節點持續獲得獎勵，權益呈指數增長，進一步提高其在未來委員會中的佔比。

這種攻擊的關鍵在於利用了權益機制的正反饋特性：權益高的節點更容易被選入委員會，獲得更多獎勵，進而權益更高。攻擊者通過操縱這一循環，實現權益的指數增長與網路控制權的轉移。

3.3.2 攻擊階段詳述

3.3.2.1 階段一：潛伏階段 (Latent Phase)

在潛伏階段，攻擊者的目標是積累初始權益並建立信譽，具體策略包括：

- 誠實行為 (Honest Behavior)：攻擊者控制的節點無論是作為 UP、Aggregator 還是 Verifier，均嚴格遵守協議規則，提交高質量的模型更新與正確的驗證結果。
- 權益積累 (Stake Accumulation)：通過誠實參與，攻擊者節點獲得系統獎勵，權益逐漸增加。
- 等待時機 (Waiting)：攻擊者持續觀察委員會組成，等待多個惡意節點同時被選入委員會，形成超過 2/3 席位的時機。

潛伏階段的持續時間取決於攻擊者的初始權益佔比與委員會大小。假設攻擊者控制 $f = 0.3$ 的節點，委員會大小 $C = 7$，則攻擊者需要至少 5 個節點被選入委員會才能形成超過 2/3 的優勢。根據超幾何分佈，這種情況發生的機率為：

$$P(\text{超過 2/3}) = \sum_{k=\lfloor 2C/3 \rfloor + 1}^{\min(fM, C)} \frac{\binom{fM}{k}\binom{(1-f)M}{C-k}}{\binom{M}{C}} \tag{3.1}$$

當 $f = 0.3, C = 7$ 時，這一機率約為 2.4%，意味著攻擊者平均需要等待約 42 輪才能獲得一次攻擊機會。

3.3.2.2 階段二：佔領階段 (Capture Phase)

當攻擊者在系統中累積了足夠的權益並控制了委員會的超過 2/3 席位時，PCCA 進入佔領階段。不同於傳統攻擊單一的破壞模式，本研究根據攻擊者對系統組件 (Verifier 和 Aggregator) 的控制程度，定義了兩種層次的攻擊場景：戰略性餓死與全棧投毒。

A. 場景一：戰略性餓死 (Strategic Starvation via Committee Capture) 在此場景中，攻擊者控制了 Verifier 委員會的超過 2/3 席位 ($|\mathcal{V}_{mal}| > \frac{2}{3}|\mathcal{V}_{committee}|$)，但當前輪次的 Aggregator 為誠實節點或未受攻擊者完全控制。

攻擊者的目標是最大化相對權益增益。基於 BlockDFL 的獎勵連鎖機制，只有當提案被委員會批准時，相關聯的 Aggregator 和 Update Providers 才能獲得獎勵。利用這一點，惡意委員會採取以下策略：

- **拒絕誠實提案**：惡意委員會投票否決由誠實 Aggregator 提交的高質量聚合結果。這導致誠實 Aggregator 及其背後的誠實 Update Providers 無法獲得本輪獎勵，造成「零收益」懲罰。
- **批准次優更新**：如果存在一個包含較多惡意 Update Providers 的 Aggregator (即使其聚合結果為次優，Sub-optimal)，惡意委員會會優先批准該提案。

後果分析：這種攻擊雖然在短期內僅導致模型收斂速度減緩 (因為接受了次優而非最優更新)，但其主要破壞力在於經濟層面。誠實節點的權益因持續被「餓死」而停滯，而惡意節點的權益則持續增長，導致攻擊者的權益佔比 (Stake Ratio) 在下一輪選擇中進一步擴大，形成正反饋循環。

B. 場景二：全棧投毒 (Full Stack Poisoning) 在此場景中，攻擊者同時實現了對共識層和聚合層的滲透，即同時控制了委員會超過 2/3 席位以及當選的 Aggregator。這是 PCCA 最危險的形態。

攻擊者的目標轉變為直接破壞模型性能。由於 Aggregator 和 Verifier 均被攻陷，現有的防禦機制 (如聚合層的 Krum 演算法或驗證層的準確率檢查) 將完全失效：

- **惡意聚合**：惡意 Aggregator 接收來自惡意 Update Providers 的「標籤翻轉 (Label Flipping)」更新，或者直接構造被污染的全域模型更新。
- **強制共識**：儘管該更新包含明顯的錯誤或惡意特徵，惡意委員會成員仍會協同投出贊成票，強制達成共識並將毒化模型寫入區塊鏈。

後果分析：全棧投毒繞過了系統所有的檢測機制。由於惡意 Aggregator 和 Verifier 瓜分了系統獎勵，攻擊者不僅成功破壞了全域模型 (Global Model) 的準確率，還進一步鞏固了其權益優勢，使得系統難以通過正常的選舉機制自我修復。

3.3.3 權益增長動態分析 (Stake Growth Dynamics Analysis)

在沒有外部干預的情況下，PCCA 會導致惡意節點的權益呈指數增長。我們可以通過數學模型來量化這種權益壟斷的過程：

- 初始階段：假設攻擊者初始權益佔比為 $f_0 = 0.3$。
- 首次攻擊：當攻擊者首次獲得委員會超過 2/3 席位時，獨佔獎勵 R，權益增加至 $S_{mal}(1) = S_{mal}(0) + R$。
- 循環攻擊：隨著權益增加，攻擊者獲得委員會超過 2/3 席位的機率提高，攻擊頻率增加。假設每 k 輪成功攻擊一次，則經過 t 輪後，惡意節點的平均權益為：

$$S_{mal}(t) = S_{mal}(0) + \frac{t}{k} \cdot R \tag{3.2}$$

而誠實節點的權益保持 $S_{hon}(t) = S_{hon}(0)$，導致權益比例為：

$$\frac{S_{mal}(t)}{S_{hon}(t)} = \frac{S_{mal}(0) + \frac{t}{k} \cdot R}{S_{hon}(0)} \tag{3.3}$$

隨著 t 增加，這一比例趨向無窮，意味著攻擊者最終將完全控制系統。

3.3.4 攻擊效果與影響

PCCA 對系統造成多層次的破壞：

- 模型品質下降：由於惡意委員會可能接受次優更新或排除部分誠實更新，模型收斂速度變慢，最終準確率下降。在極端情況下，如果惡意委員會完全拒絕誠實更新，模型將無法收斂。
- 網路控制權轉移：隨著惡意節點權益佔比的提高，它們在委員會中的佔比也持續上升。最終，攻擊者可以持續控制委員會，完全掌握聯邦學習過程。

- 去中心化假設崩潰：區塊鏈聯邦學習的核心價值在於去中心化，避免單點故障與中心化信任。然而，PCCA 通過權益壟斷，實質上將系統重新中心化至攻擊者手中，違背了去中心化的初衷。
- 經濟激勵扭曲：誠實節點發現無論如何努力，都無法獲得獎勵，可能選擇退出系統。這進一步降低了誠實節點的佔比，加速了系統的崩潰。

3.3.5 與傳統攻擊的區別

PCCA 與傳統的拜占庭攻擊或資料投毒攻擊有本質區別，如表 3.2 所示。

表 3.2: 與傳統攻擊的區別

特徵	傳統攻擊	PCCA
攻擊目標	模型品質	網路控制權
攻擊者動機	破壞	利益最大化
攻擊策略	直接投毒	漸進式滲透
隱蔽性	低 (立即可檢測)	高 (初期表現誠實)
自我強化	無	有 (權益正反饋)
防禦方法	資料層防禦	需要激勵相容機制

傳統攻擊可以通過 Krum 等資料層防禦方法應對，但 PCCA 繞過了這些防禦，直接攻擊共識層。這種攻擊的隱蔽性與自我強化特性，使得傳統的誠實多數假設不再可靠。

3.4 安全目標

基於上述威脅模型，本研究的防禦機制需要達成以下安全目標：

3.4.1 防止委員會被惡意節點控制

核心目標：即使攻擊者在某一輪獲得委員會超過 2/3 席位，也無法持續控制委員會。

具體要求：

- 攻擊者無法通過單次成功攻擊獲得長期優勢。
- 系統能夠檢測並懲罰惡意委員會的行為。

- 懲罰機制足以剝奪攻擊者的作惡能力，防止其再次獲得委員會超過 2/3 席位。

3.4.2 確保誠實節點的權益公平增長

核心目標：誠實節點通過正常參與系統，能夠持續獲得獎勵，權益穩定增長。

具體要求：

- 惡意委員會無法阻止誠實節點獲得應得的獎勵。
- 即使在攻擊發生時，誠實節點仍有機制保障其權益不受損害。
- 長期來看，誠實節點的權益佔比應保持穩定或增長，而非下降。

3.4.3 維持模型收斂性與準確性

核心目標：在存在 PCCA 攻擊的情況下，系統仍能保證模型正常收斂，達到預期準確率。

具體要求：

- 防禦機制能夠識別並拒絕次優更新。
- 即使部分輪次受到攻擊影響，整體訓練過程仍能收斂。
- 最終模型準確率與無攻擊場景相當。

3.4.4 保持系統的去中心化特性

核心目標：防禦機制本身不應引入新的中心化風險或信任假設。

具體要求：

- 不依賴可信第三方或中心化仲裁者。
- 不依賴誠實多數假設，而是基於激勵相容的博弈論機制。
- 任何節點都能參與驗證與挑戰，無需特殊權限。

3.4.5 激勵相容性

核心目標：使得理性攻擊者的最優策略是誠實行為，而非發動攻擊。

具體要求：

- 攻擊的預期收益必須為負，即 $E[\text{Payoff}] = P_{success} \cdot G_{attack} - P_{caught} \cdot L_{slash} < 0$。
- 懲罰機制 L_{slash} 必須遠大於潛在收益 G_{attack}，使得即使攻擊成功機率較高，預期收益仍為負。
- 獎勵機制應激勵誠實行為，使得誠實節點的長期收益高於攻擊者。

3.5 本章小結

本章定義了本研究針對的威脅模型，重點聚焦於區塊鏈聯邦學習系統中的「漸進式權益佔領攻擊」(PCCA)。與傳統的資料層攻擊不同，PCCA 針對的是共識層的驗證者，通過兩階段策略 (潛伏 → 佔領) 逐步實現網路控制權的轉移。

PCCA 的核心機制包括：

- 次優更新：惡意委員會提交次優聚合結果，隱蔽性強。
- 戰略性餓死：通過排他性獎勵分配，阻止誠實節點權益增長。
- 權益指數增長：利用權益機制的正反饋特性，實現權益壟斷。

這種攻擊的危險性在於其隱蔽性、自我強化性，以及對去中心化假設的根本性顛覆。現有的資料層防禦方法 (如 Krum) 無法應對這種攻擊，因為它們依賴於驗證者誠實執行的假設。

基於這一威脅模型，本研究提出了五個安全目標：防止委員會控制、確保權益公平增長、維持模型收斂、保持去中心化特性，以及實現激勵相容性。下一章將介紹本研究提出的防禦機制，展示如何通過激勵相容的挑戰與罰沒機制，在不依賴誠實多數假設的前提下，有效防禦 PCCA 攻擊。

第四章　挑戰增強型委員會架構

(Challenge-Augmented Committee Architecture)

為達成第三章提出的五項安全目標，本章提出「挑戰增強型委員會架構」(Challenge-Augmented Committee Architecture, CACA)。此架構在第二章 2.5 節定義的基準委員會模型上，新增異步審計機制與內部罰沒協議，將安全性從「門檻安全性」轉向「經濟安全性」。

本章首先概述 CACA 的設計哲學與整體架構（4.1 節），接著詳細描述各組件的設計，並透過通訊複雜度與概率模型分析論證其在維持高效率的同時如何提供強大且具備激勵相容性的安全保障。

4.1　系統架構概覽

本系統旨在建立一個具備經濟安全性與執行效率的去中心化學習平台。圖 4.1 展示了本研究所提出的 Challenge-Augmented Committee Architecture (CACA) 整體運作流程，涵蓋了從角色分配、模型訓練、聚合驗證到潛在挑戰仲裁的完整生命週期。

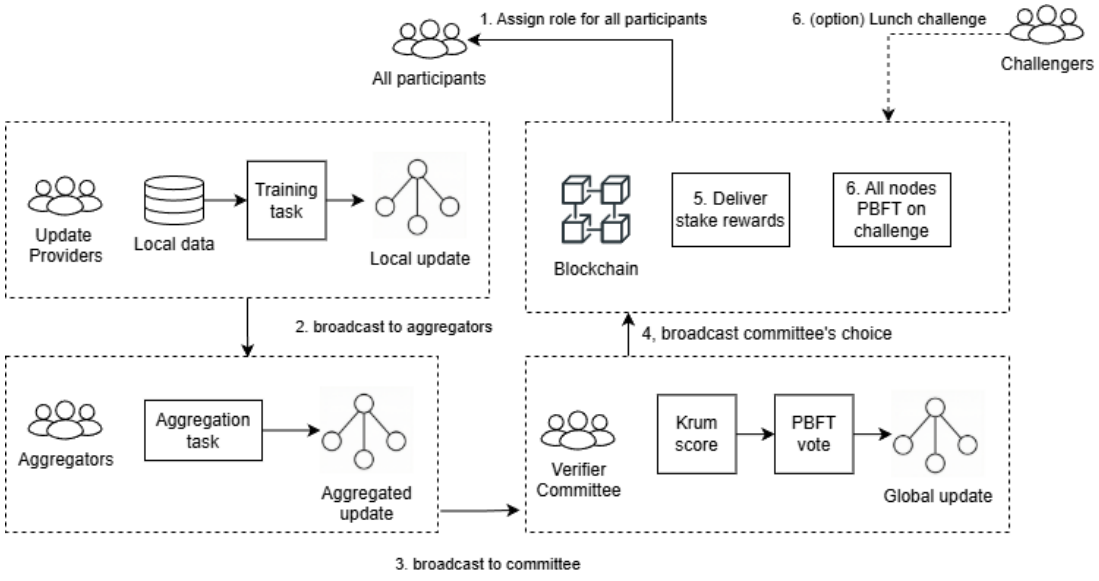


圖 4.1: Challenge-Augmented Committee Architecture (CACA) 系統架構與工作流程圖

4.1.1 核心角色定義

本系統包含四個核心角色，各自承擔不同的職責：

- **訓練者 (Update Provider, UP)**：持有本地私有資料的參與節點。負責在本地進行模型訓練，並將運算出的本地更新 (Local Updates) 提交給選定的聚合者。
- **聚合者 (Aggregator, AG)**：負責收集來自多個訓練者的本地更新，執行初步彙整並生成聚合更新 (Aggregated Updates)，隨後將其作為「提案 (Proposal)」提交給驗證委員會。
- **驗證委員會 (Verifier Committee, VC)**：由質押權重選出的小型委員會。其核心職責是針對多個聚合者提交的提案運行 Krum 評分，並透過 PBFT 共識投票決定其中哪一份定為該輪之全域更新 (Global Update)，隨後將其上鏈。
- **挑戰者 (Challenger)**：任何持有足夠質押的節點均可擔任。挑戰者在背景異步監聽鏈上資料，重新運算 Krum 演算法；若發現委員會選定的全域更新與正確之 Krum 計算結果不符，則發起挑戰。

4.1.2 工作流程

本系統之工作流程分為以下階段，旨在兼顧訓練效率與共識公正性：

1. **動態角色抽選**：在每一輪次起始時，區塊鏈根據前一區塊的哈希值 (Hash) 分配該輪角色。抽選機率與節點權益 (Stake) 成正比，且嚴格按照 **Verifier** → **Aggregator** → **Update Provider** 的順序進行抽選。
2. **本地訓練與廣播**：被選定為 UP 的節點使用本地資料進行模型訓練，並將結果傳遞給當輪選定的 Aggregators。
3. **提案彙整與提交**：Aggregators 整合接收到的多個 Local Updates，計算出初步的 Aggregated Updates 並提交給 Verifier Committee。

Algorithm 2 CACA Execution Protocol (Instant Update)

Require: Current Round r, Total Stake Weighted Nodes $\mathcal{N}$
Ensure: Updated Global Model w_{r+1}

Role Assignment:
Blockchain selects $\mathcal{V}$ (Committee), $\mathcal{A}$ (Aggregators), $\mathcal{U}$ (Update Providers) from $\mathcal{N}$ based on stake and randomness.
Training & Aggregation:
Each $u \in \mathcal{U}$ trains using w_r, broadcasts updates to $\mathcal{A}$.
Each $a \in \mathcal{A}$ aggregates updates into proposal p_a, sends to $\mathcal{V}$.
Consensus & Update:
$\mathcal{V}$ runs Krum on all proposals $\{p_a\}$.
$\mathcal{V}$ votes on the best proposal via PBFT.
Commit w_{r+1} to blockchain **immediately**.
Distribute rewards to $\mathcal{U}, \mathcal{A}, \mathcal{V}$.

Algorithm 3 Asynchronous Challenge Mechanism (Slash-Only)

Require: Challengers $\mathcal{C}$
Ensure: Punishment for Malicious Acts

for each Challenger $c \in \mathcal{C}$ **do**
    c retrieves committee inputs and re-executes Krum.
    if c detects outcome mismatch with w_{r+1} **then**
        c posts **Challenge Transaction** with deposit.
        Arbitration Triggered: All nodes re-verify.
        if Malicious Consensus Confirmed **then**
            Burn/Slash stake of malicious $\mathcal{V}$.
            Reward Challenger c and all nodes.
            // Note: Model w_{r+1} is NOT reverted.
        end if
        Exit Loop.
    end if
end for

4. **委員會驗證與投票**：驗證委員會針對所有收到的聚合提案運行 Krum 演算法進行評分。委員會成員隨後透過 PBFT 共識投票，選出評分最優的提案作為最終的 Global Update。

5. **即時更新與獎勵分發**：區塊鏈更新全域模型，並根據貢獻度向 UP、AG 與 VC 成員發放獎勵。此過程為非阻塞式，下一輪訓練將立即基於新模型開始。

6. **異步挑戰 (選用階段)**：若挑戰者發現委員會選定的結果與 Krum 運算答案不一致，可質押押金發起挑戰。隨後區塊鏈發起全參與者的 PBFT 仲裁，重新執行 Krum 運算以判定正確答案。

4.2 異步審計與究責機制

4.2.1 即時執行策略

設計哲學上，本系統區別於金融交易系統對「強一致性 (Strong Consistency)」的追求。聯邦學習作為一種機器學習過程，具有天然的「抗噪性 (Noise Tolerance)」。模型參數的微小偏差通常不會導致災難性後果，且可透過後續訓練修正。因此，本系統優先保證「活性 (Liveness)」：

- **機制**：只要驗證委員會達成共識，更新即視為有效。模型參數立即更新，所有訓練者基於新模型進行下一輪訓練。
- **優勢**：端到端延遲 (End-to-End Latency) 降至最低，系統運作效率與無防禦的中心化系統幾乎一致。

4.2.2 異步挑戰流程

挑戰流程之核心在於「數學確定性」，確保委員會無法利用資訊不對稱來操縱模型聚合結果。

1. **觸發條件**：挑戰者監控鏈上資料，發現委員會選定的全域更新與對該輪所有聚合提案執行 Krum 運算所得之結果不一致。
2. **挑戰發起**：挑戰者提交挑戰交易並繳納質押金，以防止濫用挑戰機制造成的 DoS 攻擊。
3. **仲裁執行 (Arbitration)**：
    - 智能合約鎖定相關質押金，並調取該輪鏈上緩存的所有聚合提案。
    - 觸發全網仲裁，全網驗證者重新運算 Krum 演算法並通過 PBFT 共識對仲裁結果進行投票判決。

4.2.3 處置決策：僅懲罰不回滾 (Slash-Only Policy)

當仲裁認定委員會作惡時，系統採取「僅懲罰不回滾」的處置策略。本設計不採用傳統分散式系統的回滾 (Revert) 機制，主要基於以下三點學術考量：

- **決策依據**：
  1. **算力效率與自癒特性**：回滾模型將導致該輪次後的所有訓練失效，造成嚴重的算力與資源浪費。考量到聯邦學習具備顯著的自我修復能力 (Self-healing Capacity)，誠實節點的後續更新能逐步抵銷惡意梯度帶來的噪音。
  2. **仲裁延遲與模型時效**：全參與者的 PBFT 仲裁機制具備較高的通訊複雜度與延遲。在高度動態的訓練過程中，當仲裁判定成立時，模型往往已透過後續輪次完成了初步自癒；此時若強行回滾，不僅不具時效性，反而會破壞系統的連續性。
  3. **正規化效應**：從機器學習角度分析，少許非最佳的次優選擇 (Sub-optimal Updates) 可視為向全域模型引入隨機噪音，在特定情境下有助於避免過擬合 (Overfitting) 問題，提升模型的泛化能力。
- **處理方式**：
  - **執行懲罰**：系統將立即罰沒 (Slash) 惡意委員會成員與聚合者的全額質押金。該筆資金除了作為挑戰者的賞金外，其餘部分將分配給全體誠實參與者作為補償獎勵，以維持激勵相容性。
  - **模型處理**：保留受影響的更新紀錄，不執行狀態回退。系統依靠 FL 演算法自身的強健性，由後續輪次的誠實更新逐步覆蓋並修正其影響。
- **威懾力**：透過將安全性由「事前預防」轉向「經濟制裁」，即便攻擊者成功注入一次毒化更新，其代價將是損失巨額資金並被永久移出治理委員會。這種極高且不可逆的作惡成本，足以中斷漸進式委員會佔領攻擊 (PCCA) 的權益正反饋循環，達成長期治理的穩定。

4.3 安全性保證

本節分析系統的安全性來源，提出雙層信任模型並分析攻擊成本。

4.3.1 雙層信任模型 (Two-Tier Trust Model)

本系統採用混合信任假設，將效率與安全性職責分層：

- **檢測層 (Detection Layer)**：採用 **1-of-N 誠實假設**。只要全網 N 個節點中，有一個誠實節點 (無論是委員會外的閒置節點還是候補節點) 願意擔任挑戰者，攻擊行為就會被揭露。這極大降低了監督門檻。
- **仲裁層 (Arbitration Layer)**：採用 **全網 2/3 誠實假設**。當挑戰發起後，最終判決權回歸全網 (或大型陪審團)。假設 $N_{total} > 3f$，即全網誠實節點佔多數。這是區塊鏈系統的標準安全假設。

邏輯總結：小委員會 (Small Committee) 負責效率，容忍其可能被短暫收買；大網路 (Full Network) 負責最終安全與仲裁，因其規模巨大而難以被收買。

4.3.2 攻擊成本分析

在此雙層模型下，攻擊者若想成功發動攻擊且不被懲罰，必須同時滿足以下條件：

1. 收買當前輪次的委員會超過 2/3 成員，以通過惡意更新。
2. 收買全網超過 $1/3$ 的節點，以在仲裁階段阻擋共識達成或扭曲判決。

結論：這將攻擊成本從單純收買小委員會的 $O(C)$ 提升到了收買全網節點的 $O(N_{total})$，實現了安全性的顯著擴展。

4.4 效率分析

本節透過通訊複雜度比較與概率模型分析，論證本系統的高效性與安全性平衡。

4.4.1 通訊複雜度公式

對比三種模式的訊息複雜度 (Message Complexity)：

- **傳統 PBFT (全網驗證)**：需要全網廣播與確認，複雜度為 $O(N^2)$。
- **BlockDFL (固定小委員會)**：僅在委員會內共識，複雜度為 $O(C^2)$，但安全性隨 C 減小而降低。
- **本方案**：
  - **正常情況**：僅需委員會共識，複雜度為 $O(C^2)$。由於有威懾機制，可安全使用極小的 C。
  - **挑戰情況**：委員會共識加上全網仲裁，複雜度為 $O(C^2) + O(N^2)$。

設挑戰發生概率為 p。在理性假設下，由於高額懲罰的存在，攻擊者傾向於不攻擊，故 $p \to 0$。期望通訊複雜度為：

$$E[Comm] = (1-p) \cdot O(C^2) + p \cdot (O(C^2) + O(N^2)) \approx O(C^2) \tag{4.1}$$

這表明在絕大多數時間，系統運行效率與輕量級的小委員會方案一致。

4.4.2 委員會大小的概率分析

為了進一步證明小委員會的安全性，我們使用超幾何分佈 (Hypergeometric Distribution) 進行分析。目標是運算最小委員會大小 C，使得惡意節點佔據委員會超過 2/3 ($> 2C/3$) 的機率 P_{mal} 低於特定閾值 (如 1%)。

參數定義：

- N: 驗證者總池大小。
- f: 網路中惡意節點的比例 (例如 30%)。
- X: 委員會中惡意節點的數量。

數學模型：委員會選舉屬於無放回抽樣，服從超幾何分佈。惡意節點數量 X 的概率質量函數為：

$$P(X=k)=\frac{\binom{fN}{k}\binom{(1-f)N}{C-k}}{\binom{N}{C}} \tag{4.2}$$

惡意節點佔據超過 2/3 (即攻擊成功) 的概率 P_{mal} 為 $X \geq \lfloor 2C/3 \rfloor + 1$ 的累積機率：

$$P(X \geq \lfloor 2C/3 \rfloor + 1)=\sum_{k=\lfloor 2C/3 \rfloor+1}^{C}\frac{\binom{fN}{k}\binom{(1-f)N}{C-k}}{\binom{N}{C}} \tag{4.3}$$

分析實例：設 $N = 100$, 惡意比例 $f = 0.3$ (即 30 個惡意節點)。不同 C 值下的風險如下：

- 若 $C = 5$，惡意佔領 ($X \geq 4$) 的機率約為 2.74%。
- 若 $C = 7$，惡意佔領 ($X \geq 5$) 的機率約為 2.42%。
- 若 $C = 9$，惡意佔領 ($X \geq 7$) 的機率驟降至 0.28%。
- 若 $C = 11$，惡意佔領 ($X \geq 8$) 的機率約為 0.25%。
- 若 $C = 13$，惡意佔領 ($X \geq 9$) 的機率約為 0.21%。

結論：即使在 N 較大時，只需要一個極小的 C (如 9) 即可將被惡意控制的風險控制在 1% 以下。配合異步審計機制，即使這 1% 的風險發生，攻擊者也會隨後面臨高額懲罰。這證明了使用小委員會兼顧效率與安全的可行性。

4.5 激勵機制

激勵機制是維持系統長期安全運行的動力核心。本系統維持基於 Slashing 的獎懲邏輯，但強調資金流向與即時執行的配合。

- **獎勵來源**：系統不依賴額外的增發來支付高額的審計費用，而是透過對違規者的資產罰沒 (Slashing) 來支付審計與仲裁成本。
- **動態調整**：若系統長期無挑戰發生，可適當降低挑戰者的質押門檻以鼓勵更多節點參與監聽；若挑戰頻發，則提高質押門檻與懲罰力度。

- **長期收益**：對於誠實節點，參與委員會獲得的區塊獎勵是穩定的預期收益；而對於潛在攻擊者，一次攻擊的收益是有限的(本次更新的控制權)，但損失是巨大的(全額質押金)。这种不對稱的風險收益比確保了誠實是經濟上的最優策略。

4.6 本章小結

本章提出了一種基於異步審計與即時執行的防禦框架。透過移除傳統的確認等待期，我們最大化了聯邦學習的訓練效率。同時，利用雙層信任模型與超幾何分佈分析，我們證明了小委員會配合異步挑戰機制，能夠在極低的通訊成本下實現等同於全網共識的安全性。這種設計成功解決了區塊鏈聯邦學習中效率與安全的兩難困境。下一章將通過多維度的模擬實驗，驗證本架構在面對漸進式委員會佔領攻擊時的有效性與穩健性。

第五章　實驗評估 (Experimental Evaluation)

本章旨在驗證所提出的「基於異步審計與即時執行的防禦架構」在防禦「權益佔領攻擊」方面的有效性，並評估其在維持去中心化安全性的同時，是否能顯著提升系統效率。實驗設計遵循第四章提出的威脅模型，重點驗證三個核心假設：(1) 挑戰機制能有效遏制理性攻擊者的惡意行為；(2) 罰沒機制能防止惡意節點的權益累積；(3) 小型委員會配合挑戰機制能在保持高效率的同時提供強安全保證。

5.1　實驗設置

為了公平比較，我們在相同的實驗環境下模擬了本研究提出的方法與目前主流的基於委員會的防禦方案。

5.1.1　資料集與模型

我們採用 MNIST 手寫數字資料集作為基準測試任務。模型架構為一個標準的捲積神經網路，包含兩個捲積層與兩個全連接層。

資料分佈設置：為了全面評估系統性能，本研究考量了獨立同分佈 (IID) 與非獨立同分佈 (Non-IID) 兩類環境。在 IID 設置中，資料被均勻地隨機分配給所有客戶端。而在 Non-IID 設置中，我們採用基於 Dirichlet 分佈 (Dir(α)) 的資料劃分，並將濃度參數設定為 $\alpha = 0.5$。這種設定會導致每個客戶端持有的類別分佈呈現高度異質性，模擬了真實場景中資料分佈極度不均的情況，從而增加模型聚合與抗攻擊的挑戰。

5.1.2　基準方法與攻擊場景

基準方法 (BlockDFL)：採用固定大小委員會的主流區塊鏈聯邦學習方案。該方案依賴誠實多數假設，使用 BFT 共識機制進行模型聚合驗證。委員會大小設定為 $C = 7$，這是 BlockDFL 論文中建議的配置，能在效率與基本安全性之間取得平衡。我們設定 BFT 的共識門檻為 $2/3$，即必須有超過 $2/3$ 的成員同意才能通過提案。

本研究方法 (Ours)：同樣採用 $C = 7$ 的委員會大小，但引入了事後挑戰機制。在正

常情況下，系統採用即時執行模式，僅由單一聚合器執行聚合；當檢測到異常時，任何節點都可以發起挑戰，觸發完整的 BFT 驗證流程。

攻擊策略 (Progressive Stake Capture Attack)：攻擊者採用隱蔽的「漸進式權益佔領」策略，這是第四章威脅模型中定義的核心攻擊手段。攻擊分為兩個階段：

1. 潛伏階段 (Latent Phase)：只要攻擊者尚未獲得委員會的控制權 (即未達 2/3 席位)，皆會維持潛伏狀態並表現誠實，透過提交正常的模型更新來穩定積累權益。此階段的目的是建立信譽並增加權益佔比，從而提升未來被選入委員會的機率，為發動攻擊奠定基礎。

2. 佔領階段 (Capture Phase)：一旦攻擊者在委員會中獲得超過 2/3 席位，立即根據控制情況啟動攻擊策略。具體包含兩種場景：

    - 場景一：戰略性餓死 (Strategic Starvation)。當攻擊者僅控制委員會超過 2/3 席位時，拒絕打包誠實節點的更新，僅接受包含攻擊者更新的提案，從而獨佔獎勵並使誠實節點權益停滯。
    - 場景二：全棧投毒 (Full Stack Poisoning)。當攻擊者同時控制委員會超過 2/3 席位與 Aggregator 時，直接繞過檢測機制提交「標籤翻轉」(Label Flipping) 的惡意更新，並利用委員會多數強制達成共識，從而直接破壞模型品質。

5.1.3 實驗參數

本研究實驗採用的系統參數配置如表 5.1 所示。這些參數的設定遵循了 BlockDFL [13] 等主流 BCFL 研究的標準配置，確保實驗結果的可比性。

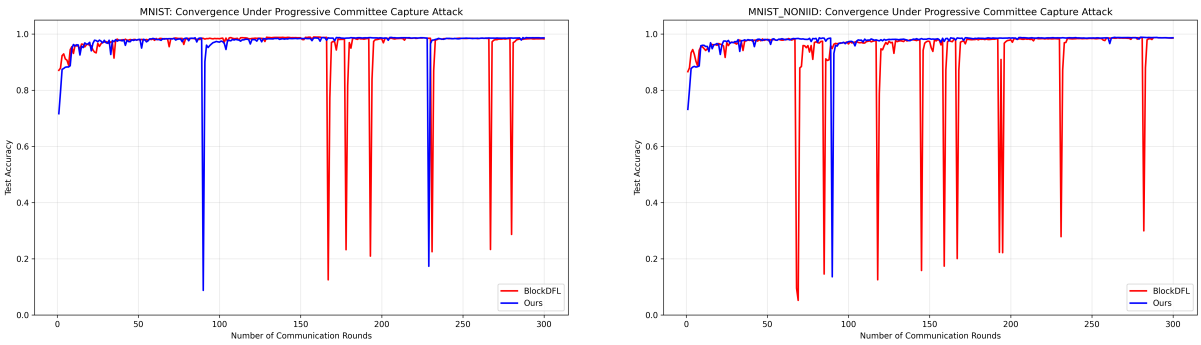
5.2 實驗結果與分析

5.2.1 模型效能與攻擊表現分析

本節針對系統在不同資料分佈下的收斂性與遭受攻擊的頻率進行量化分析。圖 5.1a 至圖 5.1b 分別展示了 IID 與 Non-IID 環境下，BlockDFL 與本研究方法（Ours）的表現。

表 5.1: 實驗參數配置 (Experimental Parameter Configurations)

參數名稱	**設定值**
訓練輪數	$R = 300$
客戶端總數	$N = 100$ (Verifier Pool Size)
委員會大小	$C = 7$
攻擊者數量	$M = 30$ (初始權益佔比 30%)
初始權益分配	所有參與節點初始均分配 100 單位
設備池分配	Aggregator: 4 位，Provider: 其餘節點
獎勵機制 (每輪)	Verifier: 1.0, Aggregator: 1.0, Provider: 0.05
罰沒機制	挑戰成功時，惡意委員全額罰沒 (Full Slashing)
學習率	$\eta = 0.01$ (衰減率 0.99)
本地訓練參數	Epochs = 1, Batch Size = 32
資料分佈環境	IID 及 Dirichlet-based Non-IID ($\alpha = 0.5$)



(a) IID 環境 (均勻分佈)　　(b) Non-IID 環境 ($\alpha = 0.5$)

圖 5.1: 模型準確率收斂比較。(a) 為 IID 環境，(b) 為 Non-IID 環境。

1) 顯性攻擊影響與收斂穩定性

實驗結果顯示，BlockDFL 在兩類環境下均展現出明顯的安全性漏洞。

攻擊頻率：在 300 輪訓練中，BlockDFL 分別遭受了 10 次（IID）與 12 次（Non-IID）成功的委員會佔領。相較之下，本研究方法透過異步審計機制，在 IID 中僅遭受 2 次佔領，在更具挑戰性的 Non-IID 環境中也僅遭受 3 次佔領，顯示出極強的韌性。

瞬時破壞力：以圖 5.1b（Non-IID）為例，BlockDFL 於第 68 輪遭受標籤翻轉（Label Flipping）攻擊時，準確度由正常水平瞬間崩潰至 9.55%。這證明了在傳統 BCFL 框架下，單次成功的委員會佔領即可對全球模型造成致命打擊。

顯性攻擊與聯邦學習的自癒性：觀察圖 5.1b（Non-IID）可以發現，BlockDFL 在第 68 輪遭受標籤翻轉攻擊後，準確度雖瞬間崩潰至 9.55%，但隨後幾輪呈現快速回升。這印證了聯邦學習具備顯著的自我修復能力（Self-healing capacity）：只要攻擊者無法持續佔領委員會，後續輪次的誠實更新即可逐步抵銷惡意梯度產生的噪聲。因此，單次的標籤翻轉攻擊雖會造成系統震盪，但通常不會導致模型不可逆的毀滅。

Non-IID 強健性解釋：值得注意的是，即便在 $\alpha = 0.5$ 的高度異質資料分佈下，本系統仍能維持與 IID 相似的收斂速度。此現象源於系統採用的「基於驗證的選優機制」（Selection-based mechanism），透過全局驗證集有效過濾了 Non-IID 引起的權重發散（Client Drift）。

2) 系統穩定性與最低不可用率分析

為了進一步量化攻擊對系統運行的實質衝擊，本研究定義「最低不可用率」（Minimum Unavailability Rate）為指標。我們保守地假設每次受擊後的恢復期僅需 5 輪（此為實驗觀測結果 5–25 輪之最小值），並據此運算系統處於效能崩潰狀態的比例。

下限估計與效能鴻溝：根據實驗資料的量化分析，在 Non-IID 環境下，BlockDFL 由於遭受了 12 次成功的委員會佔領攻擊，即便採用最為樂觀的 5 輪恢復期進行運算，系統在 300 輪的訓練過程中仍有至少 20%（即 60 輪）的時間處於不可用狀態。若進一步考量到實驗中實際觀察到的最大恢復期（25 輪），其實際癱瘓時間將遠超此比例。

相比之下，本研究提出的方法憑藉「異步審計機制」，將成功受擊次數大幅壓制在 3 次以內。在同樣的保守估計準則下，本系統的最低不可用率僅為 5%（15/300 輪）。這

項數據對比清晰地證明：儘管聯邦學習具有「自癒性」，但頻繁的受擊仍會使傳統框架在訓練過程中陷入極大的不穩定；而本方法則能確保系統在 95% 以上的訓練時間內，始終維持高品質的服務能力。

連續受擊的連鎖反應：此外，BlockDFL 的高受擊頻率（平均每 25 輪一次）與恢復期（5–25 輪）在時間軸上高度重疊。這意味著在 Non-IID 較複雜的收斂過程中，BlockDFL 極易在尚未從前次攻擊完全恢復時再次受擊，導致模型準確度長期在低位震盪，無法累積有效的全局知識。

3) 最終準確率對比

在經歷 300 輪的攻防博弈後，兩者的最終訓練結果如下：

- **IID 環境：**本研究方法最終準確率達到 98.63%，BlockDFL 為 98.26%。
- **Non-IID 環境：**本研究方法達到 98.67%，BlockDFL 為 98.57%。

誠然，BlockDFL 展現了聯邦學習的自癒特性，但在 Non-IID 環境下，每次受擊後的恢復期至少需要 5 輪。保守估計，BlockDFL 在訓練過程中有超過 20% 的時間處於不可用狀態。本研究方法透過異步審計將攻擊頻率降低了 80%，確保了模型在整個週期內維持高水準的服務能力。這種「過程穩定性」在需要實時部署的關鍵任務中，其價值遠超最終 0.1% 的準確率增益。

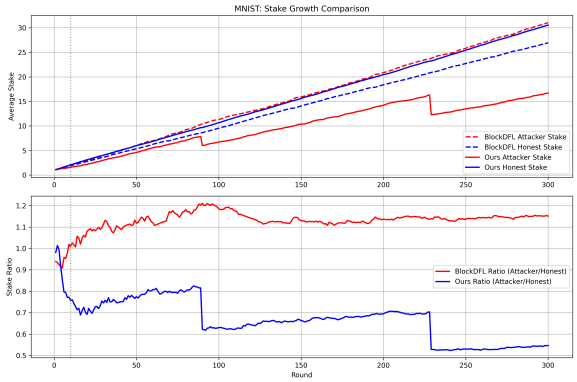
5.2.2 安全動態與治理風險深層分析

本節進一步探討權益演化與隱蔽攻擊的內在邏輯，揭示基於權益選拔（Stake-based Selection）系統中的固有治理風險。

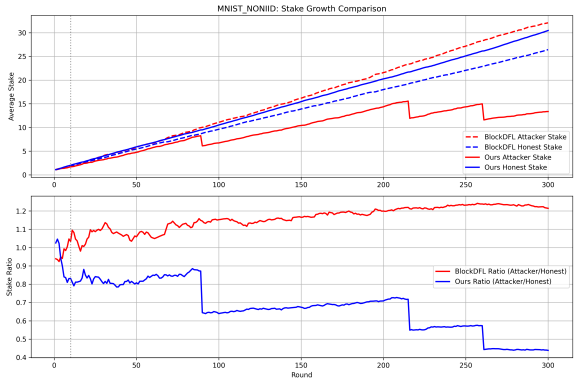
1) 權益優勢的建立與自我強化機制

透過對原始權益資料的追蹤發現，在 BlockDFL 中，攻擊者平均持有的權益穩定維持在誠實節點的 1.1 至 1.2 倍。這種優勢地位的建立具有其系統必然性：

任務價值差異：系統中執行運算量較大或關鍵性較高的任務（如 Aggregator 或 Committee 成員）所獲取的獎勵遠高於普通 Provider。



(a) IID 環境



(b) Non-IID 環境

圖 5.2: 權益演化比較。(a) 為 IID 環境，(b) 為 Non-IID 環境。

正向回饋循環：由於角色分配機制與權益掛鉤，一旦節點獲得初步權益優勢，其未來被選中擔任重要角色的機率隨之增加，進而獲得更多獎勵。

增長上限分析：攻擊者權益比未能呈現指數級成長，是因為其無法完全操控隨機的角色分配邏輯。即便惡意委員會策略性地選擇有利於惡意節點的更新，系統中仍有部分誠實節點（UP 或 AG）會獲得獎勵，從而形成了 1.1–1.2 倍的動態平衡區間。然而，只要「高貢獻任務獲得高獎勵」的分配邏輯不變，這種**領先者優勢（Leader Advantage）**便會轉化為長期的治理威脅。

2) 隱蔽投毒 (Covert Poisoning) 攻擊的普遍性與隱蔽性

進一步分析揭示，隱蔽投毒攻擊的隱蔽性並非僅限於 Non-IID 環境，而是系統層面的普遍風險。

模型指標的局限性：如實驗資料所示（例如 Non-IID 第 239 輪），即便委員會已被惡意佔領且正在執行隱蔽投毒攻擊，全球模型的準確度仍可能維持上升。這是因為攻擊者可透過保留部分高質量更新來偽裝其行為。

解耦威脅：這種現象顯示了**「模型效能」與「系統誠信」的解耦**。若缺乏本研究提出的罰沒機制（Slashing），攻擊者可以長期隱藏在系統中累積權益，直到達成「全棧共謀」（Full-stack Collusion）的條件。

3) 罰沒機制與權益抑制的動態演化

圖 5.2 記錄了 300 輪內節點權益的動態變化，這不僅反映了系統的獎懲邏輯，更揭示了惡意節點在攻擊過程中的資源損耗特徵。

1. 台階式下降的制裁特徵：觀察圖 5.2 可以發現，惡意節點的平均權益並非線性遞減，而是呈現顯著的「台階式下降」。這種現象對應了本研究異步審計機制觸發 Slashing 的具體時點：

- **IID 環境：**在第 90 輪與第 229 輪發生兩次大幅度的權益減損，最終降至誠實節點的 0.56 倍。
- **Non-IID 環境：**在第 90、216 與 261 輪分別觸發制裁，導致其權益在第 300 輪時僅剩誠實節點的 0.43 倍。

每一次「台階」的出現，都代表一次成功的惡意行為攔截與經濟懲罰。

2. 經濟資本的不可逆損耗：雖然在 300 輪的觀測期內，攻擊發生的頻率未呈現明顯的早晚期差異，但惡意節點的經濟資源（Stake）已處於持續枯竭狀態。由於本系統採用基於權益的角色選拔機制，攻擊者每次發動攻擊都面臨著喪失「治理資本」的風險。

3. 長期治理安全性的推論：儘管短期內攻擊者仍能憑藉剩餘權益參與競爭，但 0.43–0.56 倍的權益差距已構成實質性的進入門檻。

- **先行者優勢轉移：**誠實節點透過穩定訓練持續累積權益，擴大了與惡意節點的貧富差距。
- **攻擊難度遞增：**隨著訓練輪數繼續增加，惡意節點若要再次達成「委員會佔領」所需的席位，其權益權重將顯得捉襟見肘。

這種「台階式」的權益縮減證明了本機制能有效剝奪攻擊者的治理資源，從經濟層面限制了惡意行為的擴張潛力。

5.2.3 長期賽局中的經濟嚇阻力分析

為了驗證本研究提出的防禦機制在長期運作下的穩定性與嚇阻效果，我們將實驗模擬輪數擴展至 2000 輪。圖 5.3 展示了長期賽局下的權益動態變化，這些數據揭示了兩種機制在經濟誘因設計上的根本差異。

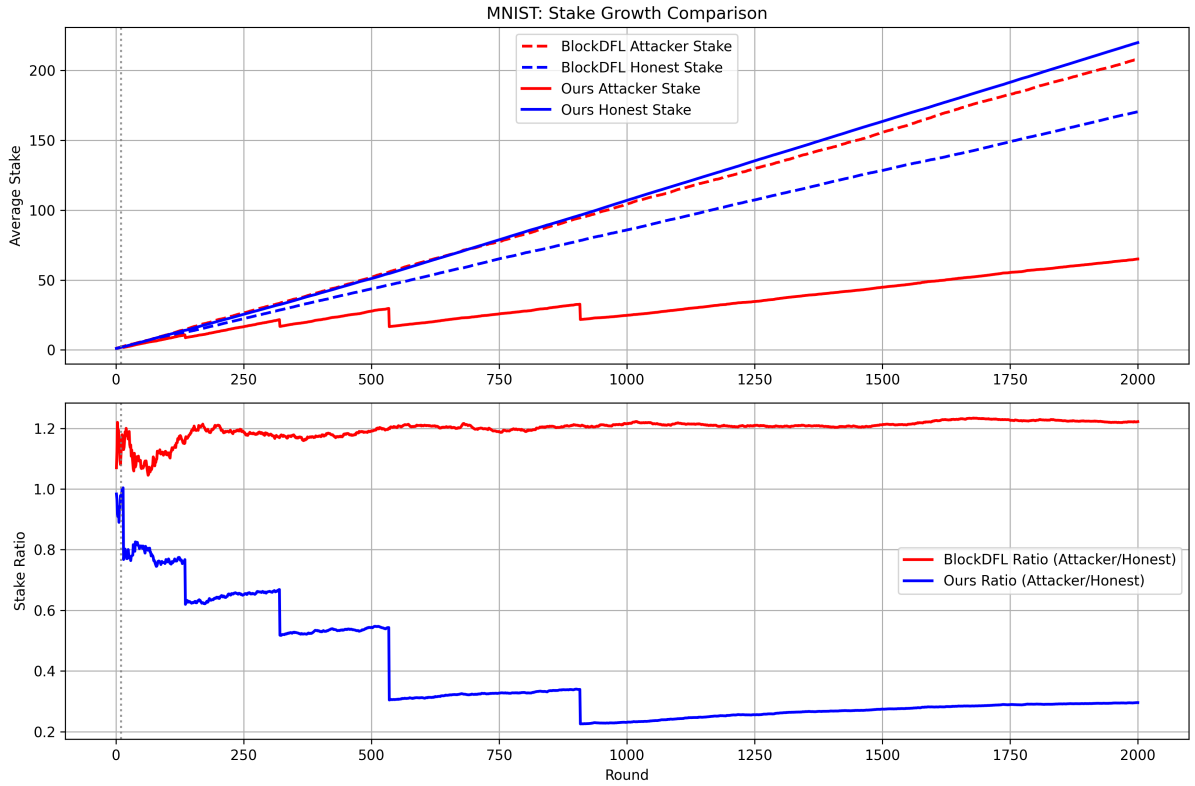


圖 5.3: 2000 輪長期模擬下的權益動態比較

1) BlockDFL 的財富固化與持續威脅

強者恆強的馬太效應：在 BlockDFL 的長期模擬中，我們觀察到顯著的財富固化現象。數據顯示，攻擊者的平均權益在約 250 輪後，便穩定維持在誠實節點的 1.2 倍左右。這種 20% 的權益優勢源於該機制缺乏有效的負向反饋迴路（Negative Feedback Loop）。一旦攻擊者透過初期優勢累積了較高的權益，其被選入委員會並獲得獎勵的機率便隨之提升，進而鞏固其經濟地位。

高頻率的治理失效：這種權益優勢直接轉化為對系統治理權的掌控。在總計 2000 輪的模擬中，惡意節點成功攻佔委員會多數高達 84 次。這意味著在 BlockDFL 架構下，攻擊者不僅能長期存活，更能平均每 24 輪就發動一次成功的委員會劫持，形成持續性的安全漏洞。

2) 本研究方法的經濟嚇阻與邊緣化效應

不對稱的攻擊風險：相較之下，本研究方法展現了極強的經濟嚇阻力。在相同的 2000 輪測試中，惡意節點僅成功佔領委員會 5 次。這巨大的差異（84 次對 5 次）證明了引入罰沒機制後，攻擊者的期望收益被大幅壓縮，迫使其在大部分時間必須保持誠實以避免資產歸零。

攻擊者的經濟致死螺旋：觀察圖 5.3 的第 909 輪可發現一個具決定性的轉折點：攻擊者在發動第五次攻擊後隨即受到異步審計機制的制裁 (Slashing)，導致其平均權益瞬間暴跌至誠實節點的 22.6%。

永久性的治理排除：這一經濟重創產生了長期的邊緣化效果。在隨後的 1091 輪（超過總時長的一半）中，攻擊者因權益基礎過低，徹底失去了競爭委員會席次的能力，再也無法成功發動任何一次佔領攻擊。這項結果有力地證實了本系統能有效將一次性的攻擊失敗轉化為永久性的治理排除，從而確保系統在長期演化中趨向於「誠實者主導」的穩定態。

5.3 效率與可擴展性分析

5.3.1 系統開銷與安全性需求對比

為了評估系統在極端壓力下的效能表現，我們設定基準安全性要求為「受攻擊頻率（成功被劫持機率 p）低於 1%」。在總節點數 $N = 100$、惡意節點佔比 $f = 30\%$ 的環境下，對比 BlockDFL 與本研究所提出的方案。

表 5.2: 不同防禦機制在相同安全性水平 ($p < 0.01$) 下的複雜度對比 ($N = 100, f = 30\%$)

評估維度	**傳統方案 (BlockDFL)**	**本研究方法 (CACA)**	**差異性質分析**
核心安全性模型	門檻安全性	經濟安全性	信任多數 vs. 激勵相容
設計哲學	悲觀併發控制	樂觀執行與異步審計	預防 vs. 治理
委員會大小 (c)	$c = 9$	$c = 5$	資源佔用降低 44.4%
常態通訊複雜度	$O(c^2) = 81$	$O(c^2) = 25$	顯著降低日常頻寬負載
安全維護成本	固定開銷	條件式開銷	靜態冗餘 vs. 動態防禦
挑戰觸發代價	無	$O(p \cdot N^2)$	僅在檢測異常時觸發
長期穩定狀態	固定於 $O(81)$	趨近於 $O(25)$	基於經濟嚇阻的博弈均衡

5.3.2 複雜度差異與經濟安全性分析

本小節針對表 5.2 中的複雜度模型進行深度分析，揭示兩者在處理安全性威脅時的本質差異：

1. 預防溢價與資源冗餘 (Pessimistic Overhead)

BlockDFL 採用的是一種「預防性策略」。為了將攻擊成功率壓制在 0.01 以下，系統必須在每一輪都維持高達 $c = 9$ 的大型委員會進行 BFT 共識。即便在系統完全誠實運行的狀態下，這份 $O(81)$ 的高昂通訊代價也是不可減免的「預防溢價」。這種設計雖然安全，但缺乏對實際威脅程度的自適應性，導致資源長期處於冗餘狀態。

2. 基於罰沒機制的博弈均衡 (Economic Deterrence)

相較之下，本研究方法將安全性保證由「事前攔截」轉為「事後追責」。透過引入罰沒機制 (Slashing Mechanism)，我們成功改變了攻擊者的收益預期：

- **激勵不相容 (Incentive Incompatibility)：**雖然本研究採用的 $c = 5$ 委員會在理論上被佔領的風險較高，但由於存在 $O(N^2)$ 的全量審計與全額罰沒風險，對於「理性攻擊者」而言，發動攻擊的期望收益將遠低於潛在的經濟損失。
- **p 值的動態演化：**雖然在模擬環境中我們考慮了 $p < 0.01$ 的頻率，但在真實部署環境中，一旦首位攻擊者遭到處罰並被剔除，後續節點將因「經濟致死螺旋」的威懾而選擇誠實策略。因此，實際的挑戰觸發頻率 p 將隨時間迅速遞減，使得系統的攤銷成本 (Amortized Cost) 極度趨近於 $O(c_{low}^2)$，從而在極低開銷下實現了與大型委員會等效的安全等級。

5.4 本章小結

本章透過多維度的實驗設計與複雜度建模，全面驗證了所提出的「挑戰增強型委員會架構」在動態攻防環境下的優越性。實驗結果不僅支持了本文的核心假設，更揭示了該架構在去中心化治理中的深層潛力，具體總結如下：

- **動態防禦的韌性：**在 MNIST 任務的 IID 與 Non-IID 環境下，本研究方法均展現了極強的抗攻擊能力。數據顯示，相較於傳統固定委員會方案（BlockDFL），本架構將受擊頻率降低了約 80%，並將系統的「最低不可用率」從 20% 大幅壓制至 5% 以下。這證明了挑戰機制能有效彌補小型委員會在即時防禦上的不足，確保模型訓練過程的連續性與穩定性。

- **經濟治理的有效性：**長期賽局實驗（2000 輪）證實，引入罰沒機制（Slashing）能對惡意行為產生實質性的經濟嚇阻。透過追蹤權益演化發現，攻擊者的治理資本會因挑戰觸發而陷入「致死螺旋」，最終其權益佔比降至誠實節點的 22.6%，達成永久性的治理排除。此結果說明了「挑戰增強」不只是技術層面的補救，更是一種從經濟誘因上根除惡意行為的治理手段。

- **效率與安全性的雙贏：**複雜度對比分析顯示，在相同的安全性邊界（$p < 0.01$）要求下，本架構成功打破了安全性與通訊開銷的強耦合關係。透過解耦共識流程，系統在常態下僅需維持 $c = 5$ 的輕量級運作（$O(c^2) = 25$），相較於必須維持 $c = 9$ 的傳統方案（$O(c^2) = 81$），顯著降低了系統整體的通訊冗餘。

綜上所述，實驗數據有力地支撐了本文論點：「挑戰增強型委員會架構」能以極低的常態通訊成本，換取等同甚至優於大型委員會的安全保證，為大規模區塊鏈聯邦學習的部署提供了一條具備高擴展性的技術路徑。

第六章　結論與未來展望 (Conclusion and Future Work)

6.1　研究總結 (Summary of Research)

本研究針對區塊鏈聯邦學習 (Blockchain-based Federated Learning, BCFL) 在委員會架構下過度依賴「誠實多數假設」的安全漏洞進行了系統性分析。我們識別出一種針對權益機制缺陷的「漸進式委員會佔領攻擊 (Progressive Committee Capture Attack, PCCA)」，揭示了理性攻擊者如何透過累積治理資源，規避傳統的資料層防禦。為了彌補這一安全性缺口，本論文提出「挑戰增強型委員會架構 (Challenge-Augmented Committee Architecture, CACA)」，其核心設計哲學在於安全性與治理規模的解耦。透過引入異步審計與內部罰沒協議，我們將系統的安全防禦從「門檻安全性 (Threshold Security)」轉向「經濟安全性 (Economic Security)」，確保系統在面對具備策略性的理性對手時，仍能維持高度的活性與模型聚合的正確性。

6.2　研究發現與貢獻 (Research Findings and Contributions)

本研究的主要發現與貢獻總結如下：

- 定義並驗證 PCCA 的威脅演化：本研究首次定義了漸進式委員會佔領攻擊的兩階段模型（潛伏與佔領），並量化了權益機制正反饋如何加速網路控制權的轉移。實驗證實，傳統架構（如 BlockDFL）在長期運行中存在顯著的財富固化與治理失效風險。

- 強化系統在極端環境下的服務能力：透過 CACA 的挑戰機制，系統在遭受 30% 惡意共謀的壓力下，能有效將成功受擊頻率壓制在極低水平。數據顯示，本架構不僅能將最低不可用率從 20% 降至 5% 以下，更能在 Non-IID 資料分佈下維持與 IID 環境相近的收斂穩定性。

- 重塑理性攻擊者的誘因結構 (Incentive Realignment)：長期賽局實驗顯示，罰沒機制能有效打破惡意節點的「權益累積循環」。數據指出，攻擊失敗導致的治理權益驟降（至誠實節點的 22.6%），實質上內部化了作惡的外部性成本，使得攻擊的預期收益遠低於潛在損失。這種經濟上的不對稱性，迫使理性節點趨向誠實策略，從而實現了無須依賴中心化仲裁的去中心化治理平衡。

- 打破安全性與通訊開銷的強耦合：本研究證明了「事前預防」轉向「事後追責」的效率優勢。在維持相同安全性邊界的前提下，CACA 允許系統在常態下僅維持

輕量級的小型委員會運作（如 $c = 5$），成功減少了約 44.4% 的通訊冗餘，為資源受限的邊緣運算場景提供具擴展性的防禦方案。

6.3 未來展望 (Future Work)

本研究提出的挑戰增強型委員會架構 (CACA) 在應對理性攻擊者時展現了優越的經濟防禦力。基於現有成果，未來研究可朝以下兩個方向進一步延伸：

6.3.1 聯邦學習自癒界限與災難性恢復機制

本研究目前仰賴聯邦學習本身的自癒能力來抵銷惡意梯度，並對攻擊者實施「僅懲罰不回滾」的策略以維持系統活性。然而，未來研究可進一步探討在更極端的攻擊行為（如旨在徹底毀滅模型的非理性拜占庭攻擊）下，自癒能力的失效界限。當「全棧投毒」場景注入的更新足以導致模型發生不可逆的發散時，如何設計一套高效的「模型回溯復原機制」將成為核心課題。此機制的挑戰在於，如何在偵測到災難性損害後，精準且低開銷地將模型狀態回溯至受攻擊前的檢查點，同時避免因頻繁回溯導致誠實節點的算力嚴重浪費。

6.3.2 針對多樣化應用情境之自適應委員會設計

本研究證實了小規模委員會配合挑戰機制能在常態下提供極高的效率。但在實際應用中，如低軌衛星網路 (LEO) 的通訊窗口限制、或是工業物聯網 (IoT) 中邊緣設備的異質資源約束，其面臨的威脅水平與環境壓力各不相同。未來研究可探討如何建構一套「自適應委員會」機制，根據當前網路的威脅監控數據與應用場景特徵，動態調整委員會的規模或選拔權重門檻。此方向的主要挑戰在於，如何在動態變化的環境中，始終維持足夠的經濟安全性 (Economic Security) 閾值，並確保效率優化不會因過度縮減委員會而產生不可預見的安全缺口。

參考文獻

[1] S. R. Pokhrel. “Blockchain Brings Trust to Collaborative Drones and LEO Satellites: An Intelligent Decentralized Learning in the Space”. In: *IEEE Sensors J.* 21.22 (2021), pp. 25331–25339.

[2] W. Wu, Z. Shen, et al. “A Sharded Blockchain-Based Secure Federated Learning Framework for LEO Satellite Networks”. In: *arXiv preprint arXiv:2411.06137* (2024).

[3] M. Elmahallawy and A. J. Akbarfam. “Decentralized Trust for Space AI: Blockchain-Based Federated Learning Across Multi-Vendor LEO Satellite Networks”. In: *arXiv preprint arXiv:2501.00000* (2025).

[4] Y. Lu et al. “Blockchain Empowered Asynchronous Federated Learning for Secure Data Sharing in Internet of Vehicles”. In: *IEEE Trans. Veh. Technol.* 69.4 (2020), pp. 4298–4311.

[5] H. Liu et al. “Blockchain and Federated Learning for Collaborative Intrusion Detection in Vehicular Edge Computing”. In: *IEEE Trans. Veh. Technol.* 70.6 (2021), pp. 6073–6084.

[6] S. R. Pokhrel and J. Choi. “Federated Learning With Blockchain for Autonomous Vehicles: Analysis and Design Challenges”. In: *IEEE Trans. Commun.* 68.8 (2020), pp. 4734–4746.

[7] Y. Lu et al. “Blockchain and federated learning for privacy-preserved data sharing in industrial IoT”. In: *IEEE Trans. Ind. Informat.* 16.6 (2020), pp. 4177–4186.

[8] Y. Qu et al. “Decentralized privacy using blockchain-enabled federated learning in fog computing”. In: *IEEE Internet Things J.* 7.6 (2020), pp. 5171–5183.

[9] W. Li et al. “EPP-BCFL: Efficient and Privacy-Preserving Blockchain-Based Federated Learning”. In: *Sci. Rep.* (2025).

[10] S. Ren, E. Kim, and C. Lee. “A scalable blockchain-enabled federated learning architecture for edge computing”. In: *PLoS One* 19.8 (2024), e0308991.

[11] M. Wang et al. “A Blockchain-Based Federated Learning Framework for Vehicular Networks”. In: *Sci. Rep.* (2024).

[12] J. Zhang et al. “FedChain: A blockchain-based federated learning framework with adaptive client selection”. In: *Proc. VLDB Endow.* (2024).

[13] Z. Qin et al. “BlockDFL: A blockchain-based fully decentralized peer-to-peer federated learning framework”. In: *Proc. Web Conf. (WWW)*. Singapore, 2024, pp. 2914–2925.

[14] Y. Li et al. “A blockchain-based decentralized federated learning framework with committee consensus”. In: *IEEE Netw.* 35.1 (2021), pp. 234–241.

[15] M. Shayan et al. “Biscotti: A Blockchain System for Private and Secure Federated Learning”. In: *IEEE Trans. Parallel Distrib. Syst.* 32.7 (2021), pp. 1513–1525.

[16] J. Weng et al. “DeepChain: Auditable and Privacy-Preserving Deep Learning with Blockchain-Based Incentive”. In: *IEEE Trans. Dependable Secur. Comput.* 18.5 (2021), pp. 2438–2455.

[17] X. Li et al. “Enhancing Byzantine robustness of federated learning via tripartite adaptive authentication”. In: *J. Big Data* (2025).

[18] Z. Xing et al. “Zero-Knowledge Proof-based Verifiable Decentralized Machine Learning: A Comprehensive Survey”. In: *arXiv preprint arXiv:2312.00000* (2023).

[19] D. H. Nguyen et al. “FedBlock: A Blockchain Approach to Federated Learning against Backdoor Attacks”. In: *Proc. IEEE Big Data*. 2024.

[20] B. McMahan et al. “Communication-efficient learning of deep networks from decentralized data”. In: *Proc. Int. Conf. Artificial Intelligence and Statistics (AISTATS)*. Fort Lauderdale, FL, USA, 2017, pp. 1273–1282.

[21] E. Bagdasaryan et al. “How to backdoor federated learning”. In: *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*. 2020, pp. 2938–2948.

[22] P. Blanchard et al. “Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent”. In: *NeurIPS*. 2017.

[23] D. Yin et al. “Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates”. In: *ICML*. 2018.

[24] Jonas Geiping et al. “Inverting Gradients - How easy is it to break privacy in federated learning?” In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. 2020, pp. 16937–16947.

[25] H. Kim et al. “Blockchained on-device federated learning”. In: *IEEE Commun. Lett.* 24.6 (2020), pp. 1279–1283.

[26] Leslie Lamport, Robert Shostak, and Marshall Pease. “The Byzantine generals problem”. In: *ACM Transactions on Programming Languages and Systems (TOPLAS)* 4.3 (1982), pp. 382–401.

[27] Miguel Castro and Barbara Liskov. “Practical Byzantine fault tolerance”. In: *OSDI*. Vol. 99. 1999. 1999, pp. 173–186.

[28] B. J. Chen et al. “ZKML: An Optimizing System for ML Inference in Zero-Knowledge Proofs”. In: *Proc. EuroSys*. 2024.

[29] Ariel Gabizon, Zachary Williamson, and Oana Ciobotaru. *PLONK: Permutations over lagrange-bases for oecumenical noninteractive arguments of knowledge*. Cryptology ePrint Archive, Report 2019/953. 2019.

[30] B. Feng et al. “ZEN: An optimizing compiler for verifiable, zero-knowledge neural network inferences”. In: *Cryptology ePrint Archive* (2021).

[31] EZKL. *Benchmarking ZKML frameworks*. EZKL Blog. 2024. URL: https://blog.ezkl.xyz/.

[32] Y. Zhu et al. "RiseFL: Secure and Verifiable Data Collaboration with Low-Cost Zero-Knowledge Proofs". In: *Proc. VLDB Endow.* 17.9 (2024), pp. 2321–2334.

[33] K. Conway et al. "opML: Optimistic Machine Learning on Blockchain". In: *arXiv preprint arXiv:2401.00000* (2024).

[34] ORA Protocol. *opML documentation*. docs.ora.io. 2024.

[35] Optimism Foundation. *Rollup protocol overview*. docs.optimism.io. 2024.

[36] S. Ren, E. Kim, and C. Lee. "A scalable blockchain-enabled federated learning architecture for edge computing". In: *PLoS One* 19.8 (2024), e0308991.

[37] H. Chen et al. "Robust blockchained federated learning with model validation and proof-of-stake inspired consensus". In: *arXiv preprint arXiv:2101.03300* (2021).

[38] Z. Peng et al. "VFChain: Enabling verifiable and auditable federated learning via blockchain systems". In: *IEEE Trans. Netw. Sci. Eng.* 9.1 (2022), pp. 173–186.