基於聯邦知識蒸餾與增量學習的入侵檢測系統

演講者: 北商大 資訊與決策科學研究所 廖文華 主任

日期: 2025/09/23

摘要 (Abstract)

本篇報告整理與分析了廖文華主任所主講的「基於聯邦知識蒸餾與增量學習的入侵檢測系統」，內容從人工智慧與機器學習的基礎概念出發，延伸至入侵檢測系統（Intrusion Detection System, IDS）的發展與挑戰，並進一步探討三項關鍵技術：聯邦學習（Federated Learning, FL）、知識蒸餾（Knowledge Distillation, KD）與增量學習（Incremental Learning, IL）。傳統 IDS 主要依賴特徵碼比對，對於未知攻擊的偵測能力有限，而資料孤島與隱私限制，更使得跨組織合作變得困難。在此背景下，演講提出將 FL 與 KD 結合，使系統能在不共享原始資料的情況下進行分散式協同訓練，並透過蒸餾壓縮模型，使其更適合部署於邊緣環境。另一方面，IL 的引入則有效解決了「災難性遺忘」問題，確保模型能在面對新型態攻擊時持續進化。

除了技術面向，報告亦涵蓋 IDS 在跨領域應用中的延伸價值，如醫療、金融與智慧城市的資安需求，並探討 AI 工具在數位轉型過程中的角色。最後，本報告指出該架構在實務應用中仍存在挑戰，包括通訊成本、效能下降、隱私與法規等問題。然而，隨著相關研究不斷推進，結合 FL、KD 與 IL 的新一代 IDS，無疑將成為未來網路防護體系中的重要發展方向。

前言 (Introduction)

在現今的數位時代，我們的生活中充滿了網際網路與資訊系統的各種應用，從金融交易、醫療服務、智慧城市基礎設施到日常通訊，皆須透過資訊技術的支援。而隨著應用的普及與應用規模日益擴大，導致網路安全議題也隨之加劇。根據國際網路安全研究指出，每年新型態的攻擊事件持續成長，駭客攻擊不僅具有更高的隱蔽性與智慧化，甚至能利用自動化工具發動大規模滲透。因此，如何有效防禦並及早偵測這些入侵行為，已成為資訊安全領域的重大挑戰 [1]。

入侵檢測系統（Intrusion Detection System, IDS）長期以來都是網路安全防護的重要基礎設施。傳統 IDS 主要依靠特徵碼比對（signature-based detection）與規則庫的持續更新，以偵測已知的惡意行為。這種方法雖然在處理既有攻擊樣本時相當有效，但對於**零日攻擊（zero-day attacks）**以及快速演化的攻擊手法，往往顯得力有未逮。此外，由於單一組織在資料蒐集上存在限制，傳統 IDS 缺乏足夠多樣化的攻擊樣本，難以全面應對各種複雜的網路威脅 [2]。

資料驅動的演算法近年來逐漸展現出新的可能性。透過在大量網路流量資料中自動萃取特徵，這類模型能辨識潛在的異常行為，而不必僅依賴人工設計的規則 [2]。相較於傳統方式，這樣的技術使 IDS 在面對未知攻擊時更具彈性與適應力。不過，此類方法也存在挑戰：一方面，不同機構之間因隱私與法規限制，往往無法直接共享原始數據；另一方面，這些模型通常需要龐大的運算資源，不易在邊緣環境中即時部署。再者，若缺乏持續更新，靜態模型也容易隨時間推移而失效，導致防禦效能下降。

在此脈絡下，「聯邦學習（Federated Learning, FL）」與「知識蒸餾（Knowledge Distillation, KD）」逐漸受到重視。前者允許多個節點在不交換原始數據的情況下進行協同訓練，達到隱私保護與模型共享的目的 [3]；後者則能將大型模型中所學得的知識壓縮至輕量化模型，使其更適合在計算能力有限的邊緣設備上部署，進一步提升即時入侵檢測的實用性 [4]。此外，「增量學習（Incremental

Learning）」則針對傳統模型容易出現的**災難性遺忘（catastrophic forgetting）**問題提出解方，使 IDS 能在面對新型攻擊樣本時，仍能兼顧既有知識的保留與新知識的吸收 [5]。

本次的演講核心重點在於如何將「聯邦學習」、「知識蒸餾」與「增量學習」結合，構建出新一代入侵檢測系統。此一架構不僅回應了資料隱私與跨組織合作的需求，也提供了模型持續演進的能力，讓 IDS 在面對不斷升級的資安威脅時，能具備更強的適應性與防護力。這樣的研究方向不僅具備理論價值，更對實務上的資安防護具有重要意義。

人工智慧與機器學習基礎

資訊科學領域中，人工智慧（Artificial Intelligence, AI）是極為重要的分支，其核心目標在於讓機器具備模擬甚至部分超越人類智慧的能力。廣義來說，人工智慧包括了規則式系統、機器學習（Machine Learning, ML）、深度學習（Deep Learning, DL）等不同層次的技術。其中，機器學習是一種透過數據進行模式學習的技術，能夠從歷史資料中歸納規律而進行預測或決策。而根據學習方式的不同，機器學習可被分為：監督式學習（Supervised Learning）、非監督式學習（Unsupervised Learning）與強化學習（Reinforcement Learning）[6]。

上述三者中，監督式學習最為廣泛應用，其訓練資料包含輸入與標籤，模型的目標是學習輸入與輸出的映射關係。常見應用包括分類（classification）、迴歸（regression）與語音辨識等 [7]。舉例來說，垃圾郵件過濾系統即屬於監督式學習的典型案例，它透過標記過的郵件資料，學習判斷新郵件是否屬於垃圾信。非監督式學習則是在沒有標籤的情況下進行，透過資料間的相似性或差異性來進行分群（clustering）與降維（dimensionality reduction）。例如，客戶市場分析中的分群方法，能夠協助企業更精準地進行行銷決策 [8]。強化學習則是一種與環境互動的學習方式，代理人（agent）在環境中透過不斷嘗試行動，並依據獲得的獎勵或懲罰

來調整策略。著名的 AlphaGo 即是強化學習的成功代表，它透過自我對弈與策略優化，達成擊敗世界圍棋冠軍的成就 [9]。

隨著演算法的演進與資料規模的擴大，深度學習在近十年成為 AI 的核心驅動力。深度學習利用多層神經網路自動學習資料中的高階特徵，突破了傳統機器學習在特徵工程上的瓶頸。其成功的原因主要來自於三大支柱：大數據（Big Data）、高效能運算（特別是 GPU/TPU 的普及）以及雲端計算的快速發展 [10]。這些因素共同促進了深度學習在語音辨識、自然語言處理、電腦視覺等領域的廣泛應用。此外在深度學習的基礎上，進一步衍生出多種創新技術。其中，生成對抗網路（Generative Adversarial Networks, GANs）是一種由生成器（Generator）與判別器（Discriminator）組成的架構，透過雙方對抗式訓練，生成器能不斷提升生成樣本的真實性，並被廣泛應用於圖像合成、影像修復與語音生成等場景 [11]。另一方面，Foundation Models 的出現成為了 AI 技術的一大轉捩點。這類模型通常擁有龐大的參數規模與跨領域的訓練資料，透過預訓練（pre-training）與微調（fine-tuning），能夠快速適應不同任務。例如，Transformer 架構支撐了 BERT、GPT 等自然語言處理模型，使其在語意理解與文本生成上展現卓越表現 [12]。

綜上所述，人工智慧與機器學習的基礎架構從監督式、非監督式與強化學習逐步發展到深度學習，再延伸至 GAN、Foundation Models 與 AIGC，已形成一條完整的技術演進脈絡。這些技術的進展不僅推動了影像處理、語音辨識與自然語言處理的突破，更為資訊安全帶來了新的契機。未來，隨著 AI 模型的持續演進與跨領域應用的拓展，AI 將成為資安防護體系中不可或缺的重要支柱。

入侵檢測系統 (IDS) 的演進

入侵檢測系統（Intrusion Detection System, IDS）自 1980 年代提出以來，一直是資訊安全體系中不可或缺的核心技術 [13]。其主要功能在於監控網路流量與系統行為，以發現潛在的惡意活動或異常事件。傳統 IDS 的運作方式大致可分為兩種：特徵碼式檢測（signature-based detection）與異常檢測（anomaly-based

detection）。其中，特徵碼式檢測仰賴既有的攻擊特徵庫進行比對，對於已知攻擊具有高度準確性。然而，此方法對於未知攻擊（zero-day attacks）或變異型惡意程式的偵測力相對不足，且需要持續更新特徵庫以維持效能 [14]。

為了彌補這些不足，研究逐漸轉向異常檢測方法，其核心理念是透過建立「正常行為基線」，當系統或網路流量偏離此基線時，即視為可能存在入侵。然而，早期異常檢測常受到誤報率過高（false positives）的困擾，導致在實務部署中成效有限。隨著人工智慧（AI）與機器學習（ML）的快速發展，IDS 的研究與應用逐漸進入新階段。

近年來，深度學習（Deep Learning, DL）技術被廣泛應用於 IDS 之中。透過卷積神經網路（CNN）、循環神經網路（RNN）、以及長短期記憶網路（LSTM）等模型，IDS 可以從龐大的網路流量資料中自動提取高維特徵，進而辨識複雜攻擊模式 [15]。相較於傳統依賴人工設計特徵的方式，AI 驅動的 IDS 更具泛化能力，能夠有效偵測未知攻擊，特別是零日攻擊。舉例來說，基於深度學習的 IDS 不僅能在傳統網路環境下發揮作用，也逐漸延伸至雲端與物聯網（IoT）環境，以因應更多元的資安威脅 [16]。

IDS 的發展歷程大致呈現由特徵碼比對 → 異常檢測 → AI 驅動的智慧檢測的演進脈絡。當前的趨勢是透過結合機器學習與深度學習技術，打造具有持續學習能力與更高適應性的 IDS，以對抗日益多樣化與隱匿性更強的網路攻擊。

聯邦學習 (Federated Learning, FL)

在現今高度互聯的數位環境中，資料已成為人工智慧模型訓練的核心資源。然而，實務中不同組織或機構往往因隱私保護、法規遵循（如 GDPR）以及競爭因素，而難以直接共享原始資料。傳統集中式的機器學習模式需要將所有資料收集至同一伺服器進行訓練，不僅存在潛在的隱私風險，也容易造成單點故障與資安漏洞 [17]。

聯邦學習（Federated Learning, FL）作為一種分散式機器學習架構，提供了解決方案。其核心概念是允許多個參與節點（clients）在不共享原始數據的情況下，於本地端進行模型訓練，僅將參數或梯度回傳至中央伺服器進行聚合 [18]。透過這種方式，敏感數據始終保留在本地端，顯著降低了隱私洩漏風險。

在資訊安全領域，聯邦學習的價值尤為凸顯。入侵檢測系統（IDS）需要依賴多樣化且龐大的網路流量資料進行訓練，才能有效辨識不同型態的攻擊。然而，單一組織的數據往往不足以涵蓋所有威脅樣本，若各機構能透過 FL 協作，便可建立更具泛化能力的 IDS 模型，同時滿足隱私與合規要求 [19]。舉例而言，不同企業或金融機構能透過 FL 共同訓練資安模型，而不必將客戶資料集中於單一平台，既保護了資料安全，也提升了威脅偵測效能。

聯邦學習還支援異質性環境（heterogeneous environments），能適應不同裝置、網路與計算能力的差異 [20]。這使得其特別適合應用於物聯網（IoT）與邊緣運算場景，例如在智慧城市或分散式感測網路中部署 IDS，達成跨域資安防護的效果。

聯邦學習不僅為 AI 模型提供了隱私保護機制，更為資安領域帶來合作共防的可能性，為新一代入侵檢測系統的建構奠定了基礎。

知識蒸餾 (Knowledge Distillation, KD)

知識蒸餾（Knowledge Distillation, KD）是一種常見的模型壓縮技術，其核心概念在於將大型模型（Teacher model）中學習到的知識轉移到較小的模型（Student model） [21]。透過這種方式，學生模型能夠模仿教師模型的行為與輸出特徵，在效能維持相近的情況下，有效降低參數數量與計算複雜度 [22]。

在傳統的機器學習與深度學習中，大型模型雖然具有良好的表現與泛化能力，但其龐大的計算需求，使得它們難以部署於計算資源有限的環境，例如行動裝置、物聯網設備或邊緣節點。知識蒸餾提供了一條可行的途徑：不僅能縮小模型規模，

還能在某些情境下提升小型模型的任務表現，使其接近甚至超越傳統訓練方式的水準 [23]。

在入侵檢測系統（Intrusion Detection System, IDS）的應用上，知識蒸餾展現出極高的潛力。由於 IDS 必須在即時環境下處理大量網路流量，模型除了要有足夠的準確性外，還需要兼顧低延遲與高效率。透過蒸餾，可以將在大型資料中心訓練的複雜模型壓縮成更精簡的版本，以便在邊緣設備上運行 [24]。這樣的設計使 IDS 能在智慧城市、工業控制系統及物聯網環境中，快速偵測異常行為或潛在攻擊，進一步提升資安防護能力。

知識蒸餾也可與其他技術結合，例如在聯邦學習（Federated Learning, FL）的架構下，各節點可透過分散式訓練獲得高效能的教師模型，並在確保資料隱私的前提下，將知識蒸餾至邊緣的學生模型 [25]。這樣的組合不僅能提高 IDS 的擴展性，還提供了一種同時兼顧效能與隱私的可行方案。

增量學習 (Incremental Learning)

在傳統的機器學習與深度學習中，模型的訓練往往假設所有資料一次性可得，並且在固定的資料集上完成訓練。然而，現實世界的應用場景，特別是資訊安全領域，資料通常是動態且持續生成的。例如，新的網路攻擊手法不斷湧現，若僅依靠靜態訓練的模型，便無法有效因應未見過的威脅。增量學習（Incremental Learning, IL）因此被提出，作為解決模型持續適應與演進需求的技術 [26]。

增量學習的核心在於使模型能夠在接收到新的資料或新類別時進行更新，而不需要從頭開始訓練整個模型。這不僅能大幅減少計算資源的消耗，也能確保模型在面對動態環境時維持穩定效能 [27]。然而，傳統模型在增量訓練時常遇到的挑戰是災難性遺忘（catastrophic forgetting），即模型在學習新知識時，會覆蓋掉原本已學會的知識，導致對舊任務的表現急遽下降 [28]。為了克服這個問題，研究者提出了多種方法，包括：

- **正則化方法**：在損失函數中加入正則化項，限制模型權重的更新幅度，以保留舊任務的重要知識。典型方法如 EWC（Elastic Weight Consolidation）[29]。
- **基於記憶的方法**：保留部分舊資料樣本，並與新資料一同訓練，以減少遺忘。這類方法被廣泛應用於持續學習框架中。
- **動態架構方法**：隨著新任務的到來，逐步擴展神經網路結構，確保模型有足夠的容量來容納新知識，而不影響舊知識的保存 [30]。

在資訊安全的應用場景中，增量學習的價值尤為突出。入侵檢測系統（IDS）需要持續監測龐大且多變的網路流量。由於攻擊手法經常快速演進，例如惡意程式的多型變化（polymorphic attacks）、零日攻擊（zero-day attacks）、以及針對 IoT 設備的新型滲透手段，若 IDS 缺乏持續學習的能力，將很快喪失防護效果 [31]。透過增量學習，IDS 能在不斷更新的資料環境下持續演進，避免因靜態模型過時而降低檢測能力。

比如說，基於增量學習的 IDS 可以在接收到新的攻擊樣本後，即時更新其檢測模型，而不需要重新蒐集並訓練龐大的資料集。這樣的特性不僅縮短了模型更新週期，還能降低運算資源消耗，特別適合部署於邊緣運算與資源有限的環境中 [32]。進一步而言，增量學習若與聯邦學習（FL）結合，則可讓不同組織在保障隱私的情況下共享新知識，並透過增量更新強化整體 IDS 的防禦能力，實現跨組織協作的資安防護。

增量學習為資訊安全領域提供了新一代的智慧防禦機制，使入侵檢測系統能夠持續進化，應對不斷變化的攻擊威脅。隨著相關技術（如正則化方法、記憶增強與動態架構）的成熟，增量學習將在資安場景中發揮更為關鍵的角色，並可能成為未來自適應型 IDS 的核心技術之一。

AI 在健康照護與情感陪伴的應用

隨著全球人口老齡化趨勢日益嚴峻，健康照護與長者陪伴已成為社會亟需解決的重要議題。根據世界衛生組織（WHO）的統計，2050 年全球 60 歲以上人口預計將超過 20 億人，這將對醫療資源、家庭結構以及社會照護體系帶來前所未有的壓力 [33]。在此背景下，AI 逐漸被視為協助解決健康照護挑戰的重要工具，特別是在智慧醫療（smart healthcare）與情感陪伴（emotional companionship）兩個層面。

AI 在健康照護的應用涵蓋廣泛，包括疾病診斷、健康監測、病歷管理以及個人化醫療建議等。其中，透過自然語言處理（NLP）與對話系統（chatbots），AI 能夠模擬醫護人員的部分角色，提供基礎的健康諮詢與數據紀錄。例如，AI 聊天系統可用於紀錄長者的日常健康資訊（如血壓、血糖、心率），並透過即時分析提供異常警示，協助醫護人員與家屬及時掌握狀況 [34]。

此外，AI 驅動的可穿戴裝置與感測器能夠持續收集生理訊號，並利用機器學習模型檢測異常模式，例如不規則心律、睡眠品質下降或跌倒事件 [35]。這些應用不僅能降低醫療系統的負擔，也能讓長者獲得更即時且個人化的照護，進一步提升生活品質。

除了身體健康，心理與情感照護同樣至關重要。孤獨感與憂鬱症是高齡人口常見的心理健康挑戰，而 AI 技術在此領域展現出相當潛力。透過對話式 AI（conversational AI）與生成式模型，智慧陪伴系統能與使用者進行自然互動，提供陪伴與情感支持 [36]。這類系統不僅能進行日常對話，還能辨識使用者的情緒狀態，進而調整回應策略，達到減緩孤獨感與提供心理慰藉的效果。

已有研究顯示，AI 聊天機器人能有效降低長者的孤獨感，並促進其社會參與度 [37]。部分先進系統甚至能與家庭成員連結，當 AI 偵測到長者情緒異常或健康數據異常時，能自動發送通知給家人或醫護人員，實現「健康監控」與「情感照護」的雙重功能。

AI 在健康照護與情感陪伴的應用不僅具有醫療價值，也體現了其跨領域的潛力。結合物聯網（IoT）、雲端運算與大數據分析，AI 能夠建構一個全方位的智慧健康管理生態系統。例如，智慧居家環境可透過感測器與 AI 系統協助監測長者行為模式，一旦發現異常活動（如跌倒、長時間未活動），系統即可自動發出警報[38]。

資料隱私與安全性是關鍵問題，如何在收集與分析健康數據的同時，保障個人隱私與防止資料外洩，是推動 AI 醫療應用的重要前提。其次，情感陪伴 AI 雖能在一定程度上緩解孤獨，但無法完全取代人際互動，因此如何定位 AI 在心理照護中的角色，亦需進一步探討 [39]。

AI 在健康照護與情感陪伴的應用在未來將更趨多元。隨著生成式 AI 與大型語言模型（LLM）的發展，智慧對話系統將能提供更自然、更人性化的互動體驗。另一方面，AI 與醫療專業人員的協作模式也將更加緊密，透過「人機協同」的方式，共同提升醫療效率與照護品質。

總而言之，AI 在智慧健康照護與情感陪伴領域的應用，展現了跨領域整合的巨大價值。從即時健康數據監測到情感支持，AI 不僅提升了長者的生活品質，也減輕了家庭與醫療系統的壓力。雖然仍面臨隱私與倫理等挑戰，但隨著技術成熟與制度完善，AI 有望成為未來健康照護與心理支持的重要支柱。

數位轉型與 AI 工具現況

數位轉型（Digital Transformation, DX）已成為組織提升競爭力的關鍵，其進程可分為三個階段：數位化（Digitization）、數位優化（Digitalization）與數位轉型（Digital Transformation） [40]。數位化強調資料的電子化，數位優化則利用資訊技術提升流程效率，而數位轉型則透過新技術徹底改變商業模式。

近年來，AI 工具在推動數位轉型上扮演重要角色。文字生成工具如 ChatGPT，能協助知識管理與自動化客服；影像生成工具如 Midjourney，促進設計

與創意產業革新；短影音生成技術如 Sora，則廣泛應用於行銷與教育領域 [41]。這些 AI 工具的滲透，正加速各行業的數位轉型進程，使組織能更有效率地回應市場需求，並創造新價值。

挑戰與限制

雖然聯邦學習（Federated Learning, FL）、知識蒸餾（Knowledge Distillation, KD）與增量學習（Incremental Learning, IL）作為核心所構建的入侵檢測系統（IDS）架構，能夠展現出兼顧隱私保護、模型輕量化與持續演進的前瞻性，但若是在實際應用過程中，仍存在許多的挑戰與限制，需進一步研究與克服。

聯邦學習的通訊與計算成本問題仍是推廣的一大障礙。由於 FL 採取分散式訓練，每個參與節點需在本地進行模型更新，並定期將梯度或模型參數回傳至中央伺服器進行聚合 [42]。在多節點環境下，這會導致通訊頻寬消耗過大，特別是當模型規模龐大或節點數量眾多時。此外，節點之間的異質性（heterogeneity），如資料分佈不均（non-IID data）、裝置計算能力差異，也會影響模型收斂速度與最終效能 [43]。因此，如何降低 FL 的通訊開銷並提升在異質環境中的適應性，是一項重要挑戰。

其次，知識蒸餾可能導致精度下降。KD 的主要目標是將大型 Teacher 模型的知識轉移到較小的 Student 模型，以利於部署於資源有限的邊緣設備。然而，蒸餾過程中，若教師模型與學生模型的結構差異過大，或蒸餾策略設計不當，往往會導致學生模型效能顯著低於教師模型 [44]。此外，蒸餾過程對溫度參數（temperature parameter）與軟標籤（soft labels）的敏感性，也可能造成模型性能不穩定 [45]。在 IDS 應用中，這樣的精度下降可能意味著對新型攻擊的檢測能力不足，帶來潛在的安全風險。

第三，增量學習仍受「災難性遺忘」（catastrophic forgetting）問題困擾。雖然 IL 的設計初衷是讓模型能夠持續學習新知識並保留舊知識，但在實務應用中，

模型在引入新資料時仍可能覆蓋已學習到的舊知識 [46]。雖然已有多種解決方案，如正則化技術（例如 Elastic Weight Consolidation, EWC）、記憶增強方法（replay-based methods）、以及動態架構方法（progressive networks），但這些方法仍存在限制，例如增加額外的計算成本或記憶體需求，難以在大規模實際部署中完全消除遺忘問題 [47]。

隱私與法規限制是實際應用中的一項挑戰，雖然 FL 的設計理念是避免共享原始數據，以保護使用者隱私，但在實作過程中仍可能面臨模型更新資訊被反向推導（model inversion attack）的風險 [48]。另一方面，在跨國與跨產業合作中，必須符合不同地區的隱私保護法規（如歐盟的 GDPR、美國的 HIPAA），這大大地增加了系統設計與落地實施的複雜性 [49]。因此，在確保隱私與合規的前提下，如何兼顧效能與安全，仍是未來研究的重要方向。

綜上所述，雖然結合 FL、KD 與 IL 的 IDS 架構展現出強大的潛力，但在推動實務應用之前，仍需針對通訊成本、精度下降、災難性遺忘以及隱私與法規挑戰提出更具體且有效的解決方案。只有在技術與制度雙方面取得突破，該架構才能真正發揮其價值，成為資訊安全防護的重要基石。

未來展望 (Future Work)

隨著人工智慧技術的快速演進，未來的入侵檢測系統（Intrusion Detection System, IDS）有望進一步結合大型語言模型（Large Language Models, LLMs），以提升對網路流量與系統日誌的語義分析能力 [50]。傳統 IDS 雖能透過統計特徵或深度學習模型進行異常偵測，但在複雜攻擊模式下，往往缺乏語境理解。LLMs 具備強大的自然語言處理與上下文建模能力，未來若能應用於資安領域，將有助於更準確地解析攻擊指令、異常通訊內容及惡意程式碼描述，進而提升對未知攻擊的檢測效能 [51]。

另一個重要的發展方向是跨領域應用。隨著數位化進程加快，AI 在醫療、金融以及智慧城市等關鍵領域的安全需求日益迫切。舉例來說，在醫療場域，AI 結合 IDS 可協助保護電子病歷與遠距醫療系統，防範資料外洩與勒索攻擊 [52]；在金融領域，AI 可用於檢測異常交易與網路詐騙，降低金融風險 [53]；在智慧城市中，AI 驅動的 IDS 更能保障智慧交通、能源管理與公共基礎設施的安全，確保城市運作不受網路威脅干擾 [54]。

除此之外，未來研究亦需著重於隱私保護與模型可信度。隨著聯邦學習、知識蒸餾與增量學習的應用，如何兼顧模型效能、資源效率與資料隱私，仍是研究挑戰之一。特別是在跨組織協作場景下，如何確保模型共享不導致資訊洩漏，將成為重要的發展議題 [55]。

未來 IDS 的發展將朝向智慧化、跨領域化與隱私保護並重的方向前進。透過結合 LLM 的語義理解能力，並將 AI 技術擴展至多樣化的應用場景，入侵檢測系統將不僅是單純的防禦工具，更將成為推動數位社會安全與永續發展的核心基石。

心得

此次參與由北商大資訊與決策科學研究所廖文華主任所主講的「基於聯邦知識蒸餾與增量學習的入侵檢測系統」演講，其具理論深度與實務價值。演講從人工智慧的基礎切入，並帶出深度學習、生成模型（GAN）、基礎模型（Foundation Models）與生成式 AI（AIGC）的發展脈絡，讓我更能理解這些技術如何推動人工智慧進入爆炸式成長的時代。尤其是這些技術被引入到資安領域時，展現出極高的應用潛力，也呼應了當前數位社會對於資訊安全的迫切需求。

在演講中，講者深入介紹了聯邦學習（Federated Learning）、知識蒸餾（Knowledge Distillation）以及增量學習（Incremental Learning）三者的結合如何為新一代入侵檢測系統帶來突破。過去的 IDS 受限於資料孤島與模型靜態性的挑戰，難以應對不斷變化的攻擊樣態。而透過聯邦學習，系統能在不同機構之間協作

訓練，同時避免資料外洩；知識蒸餾則能讓模型輕量化，使其適合部署於邊緣環境；增量學習更是確保模型能隨著時間持續進化，有效應對新興攻擊。這樣的設計思路，讓我深刻感受到 AI 不僅是學術研究的課題，更是資訊安全防護的重要武器。

除了技術層面的探討，演講也延伸到 AI 在跨領域的應用，例如智慧健康照護、情感陪伴與數位轉型。這部分讓我意識到，AI 並非僅侷限於效率提升或自動化，而是逐漸滲透到人類生活的方方面面，甚至在心理支持與社會照護領域扮演關鍵角色。當 AI 能透過對話陪伴長者、監測健康數據，並及時通知家人或醫療人員，這種「技術結合人文」的應用，讓我看見 AI 發展的社會價值。

當然，講者也提醒我們必須正視 AI 發展所面臨的挑戰與限制，包括聯邦學習的通訊成本、知識蒸餾可能造成的效能下降、增量學習的「災難性遺忘」問題，以及隱私與法規遵循的困境。這些挑戰提醒我，技術的進步並非萬能，我們需要在創新與責任之間找到平衡，才能真正發揮 AI 在資安與社會中的正向影響。

這場演講不僅讓我獲得了技術層面的新知識，也促使我思考未來 AI 在資訊安全、醫療、金融與智慧城市等領域的角色。我深切體會到，AI 的價值在於跨領域整合與持續進化，而我們能做的就是積極學習、勇於實踐，並在未來的研究或實務應用中，將這些理念落實下來，為更安全與智慧的社會盡一份心力。

參考文獻 (References)

[1] Denning, D. E. (1987). An Intrusion-Detection Model. *IEEE Transactions on Software Engineering*, SE-13(2), 222–232.

[2] Sommer, R., & Paxson, V. (2010). Outside the Closed World: On Using Machine Learning for Network Intrusion Detection. *IEEE Symposium on Security and Privacy*.

[3] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. Proceedings of AISTATS.

[4] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531.

[5] Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual Lifelong Learning with Neural Networks: A Review. Neural Networks, 113, 54–71.

[6] Mitchell, T. M. (1997). Machine Learning. McGraw Hill.

[7] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

[8] Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data Clustering: A Review. ACM Computing Surveys (CSUR), 31(3), 264–323.

[9] Silver, D., et al. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484–489.

[10] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.

[11] Goodfellow, I., et al. (2014). Generative Adversarial Nets. Advances in Neural Information Processing Systems (NeurIPS).

[12] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT.

[13] Denning, D. E. (1987). An Intrusion-Detection Model. IEEE Transactions on Software Engineering, SE-13(2), 222–232.

[14] Axelsson, S. (2000). Intrusion detection systems: A survey and taxonomy. Technical Report, Chalmers University of Technology.

[15] Kim, G., Lee, S., & Kim, S. (2016). A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Expert Systems with Applications, 41(4), 1690–1700.

[16] Shone, N., Ngoc, T. N., Phai, V. D., & Shi, Q. (2018). A deep learning approach to network intrusion detection. IEEE Transactions on Emerging Topics in Computational Intelligence, 2(1), 41–50.

[17] Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., & Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.

[18] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. *Proceedings of AISTATS*.

[19] Nguyen, T. D., Marchal, S., Miettinen, M., Fereidooni, H., Asokan, N., & Sadeghi, A. R. (2019). Federated Learning for Anomaly Detection in IoT. *arXiv preprint arXiv:1911.00600*.

[20] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Processing Magazine*, 37(3), 50–60.

[21] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*.

[22] Bucilă, C., Caruana, R., & Niculescu-Mizil, A. (2006). Model compression. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[23] Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6), 1789–1819.

[24] Xu, H., Li, X., Deng, R., & Chen, K. (2020). Lightweight Intrusion Detection for Edge Devices via Knowledge Distillation. *IEEE Access*, 8, 167856–167865.

[25] Li, Q., Wen, Z., & He, B. (2020). Federated learning systems: Vision, hype, and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering*, 35(8), 3347–3366.

[26] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys (CSUR)*, 46(4), 1–37.

[27] Chen, Z., & Liu, B. (2018). Lifelong machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 12(3), 1–207.

[28] French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4), 128–135.

[29] Kirkpatrick, J., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences (PNAS)*, 114(13), 3521–3526.

[30] Rusu, A. A., et al. (2016). Progressive neural networks. *arXiv preprint arXiv:1606.04671*.

[31] Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, 18(2), 1153–1176.

[32] Lin, W. C., Ke, S. W., & Tsai, C. F. (2015). CANN: An intrusion detection system based on combining cluster centers and nearest neighbors. *Knowledge-Based Systems*, 78, 13–21.

[33] World Health Organization. (2021). Ageing and health. *WHO*.

[34] Bibault, J. E., Chaix, B., Guillemassé, A., et al. (2019). A chatbot versus physicians to provide information for patients with breast cancer: Blind, randomized controlled non-inferiority trial. *Journal of Medical Internet Research*, 21(11), e15787.

[35] Majumder, S., Mondal, T., & Deen, M. J. (2017). Wearable sensors for remote health monitoring. *Sensors*, 17(1), 130.

[36] Laranjo, L., Dunn, A. G., Tong, H. L., et al. (2018). Conversational agents in healthcare: A systematic review. *Journal of the American Medical Informatics Association*, 25(9), 1248–1258.

[37] Khosravi, P., Rezvani, A., & Wiewiora, A. (2016). The impact of technology on older adults' social isolation. *Computers in Human Behavior*, 63, 594–603.

[38] Chen, L., Hoey, J., Nugent, C. D., Cook, D. J., & Yu, Z. (2012). Sensor-based activity recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6), 790–808.

[39] Sharkey, A., & Sharkey, N. (2012). Granny and the robots: Ethical issues in robot care for the elderly. *Ethics and Information Technology*, 14(1), 27–40.

[40] Vial, G. (2019). Understanding digital transformation: A review and a research agenda. *The Journal of Strategic Information Systems*, 28(2), 118–144.

[41] Dwivedi, Y. K., et al. (2023). So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI. *International Journal of Information Management*, 71, 102642.

[42] Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., & Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.

[43] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Processing Magazine*, 37(3), 50–60.

[44] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*.

[45] Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6), 1789–1819.

[46] French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4), 128–135.

[47] Kirkpatrick, J., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences (PNAS)*, 114(13), 3521–3526.

[48] Hitaj, B., Ateniese, G., & Perez-Cruz, F. (2017). Deep models under the GAN: Information leakage from collaborative deep learning. *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 603–618.

[49] Voigt, P., & Von dem Bussche, A. (2017). The EU General Data Protection Regulation (GDPR). Springer International Publishing.

[50] Brown, T., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 1877–1901.

[51] Wang, W., et al. (2023). Exploring the Use of Large Language Models for Cybersecurity Applications. *arXiv preprint arXiv:2306.10031*.

[52] Jiang, F., et al. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*, 2(4), 230–243.

[53] Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*.

[54] Zanella, A., et al. (2014). Internet of things for smart cities. *IEEE Internet of Things Journal*, 1(1), 22–32.

[55] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Processing Magazine*, 37(3), 50–60.