JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 24, No. 6, pp. 1317-1326, Jun. 2023

Check for updates

데이터 중독 공격 방어를 위한 신뢰도 점수 기반 연합학습

오 석 환[1] · 정 송 헌[2] · 김 경 백[3*]
[1]전남대학교 정보보안협동과정 석사 [2]전남대학교 정보보안협동과정 석사과정 [3*]전남대학교 인공지능융합학과 교수

Trust Score-based Federated Learning to Defend against Data Poisoning Attacks

Sukhuan Ou[1] · Songheon Jeong[2]· Kyungbaek Kim[3*]

[1]Master, Department of Information Securitynts, Chonnam University, Gwangju 61186, Korea
[2]Master's Course, Department of Information Securitynts, Chonnam University, Gwangju 61186, Korea
[3*]Professor, Department of Artificial Intelligence Converence Chonnam University, Gwangju 61186, Korea

[요 약]

연합학습은 로컬 데이터를 공유하지 않고 모델 업데이트만 공유하여 여러 클라이언트에서 분산된 개인 데이터 세트를 활용해 성능을 향상시키고 사용자에게 개인정보보호를 제공하는 머신러닝 기술이다. 하지만 연합학습은 모든 학습 클라이언트가 목적에 맞는 데이터를 보유하고 있으며, 학습 모델을 개선하는데 항상 긍정적인 기여를 한다는 가정으로 인해 데이터 중독 등의 공격을 통해 모델 학습에 부정적인 영향을 끼칠 수 있는 악의적인 클라이언트가 참여할 수 있다. 이 논문에서는 연합학습에 참여하고자 하는 클라이언트의 신뢰도 점수를 계산하고, 신뢰도 점수를 기반으로 참여 클라이언트를 선택하여 데이터 중독 공격 방어를 가능하게 하는 신뢰도 점수 기반 연합학습(Trust Score based Federated Learning, TSFL)알고리즘을 제안한다. 제안된 알고리즘을 구현하여 FedAVG 및 BlockFLow와 같은 선행연구와 비교 평가한 결과, 연합학습에 데이터 중독 공격이 진행되는 상황에서 제안 알고리즘이 최대 7%의 모델 성능 개선을 달성하는 것을 확인하였다.

[Abstract]

In machine learning, federated learning is used to improve performance and provide greater privacy to users by utilizing distributed personal datasets across multiple clients while sharing only updates to a learning model rather than any local data. However, federated learning is vulnerable to malicious clients who can negatively affect the learning process through attacks such as data addiction, because existing methods assume that all learning clients provide useful data and always make a positive contribution to improving the learning model. In this study, we propose a trust score-based federated learning (TSFL) algorithm that calculates a reliability score for clients in federated learning to provide a defense against data addiction attacks by selecting participating clients accordingly. The results of an experimental evaluation of the proposed algorithm and a comparison with existing methods such as FedAVG and BlockFLow confirmed that the proposed algorithm achieved up to 7% better performance in data addiction attacks on a federated learning system.

색인어 : 머신러닝, 연합학습, 데이터 중독 공격, 신뢰도, 알고리즘
Keyword : Machine Learning, Federated Learning, Data Poisoning Attack, Reliability, Algorithm

http://dx.doi.org/10.9728/dcs.2023.24.6.1317

Received 14 April 2023; **Revised** 17 May 2023
Accepted 02 June 2023

***Corresponding Author; Kyungbaek Kim**

Tel: +82 062-530-3438
E-mail: kyungbaekkim@jnu.ac.kr

Ⅰ. 서 론

기존의 중앙 집중식 머신러닝 방식에서는 사용자의 IoT 기기 및 스마트폰 등의 로컬 장치에서 수집한 데이터를 클라우드 가반 서버 또는 데이터 센터에서 중앙 집중식으로 업로드하고 처리한다. 특히 측정, 사진, 동영상, 위치정보 등의 데이터를 수집하고 이를 중앙 서버의 데이터 센터에서 관리 및 처리한다.

그러나 빅데이터 시대에 따른 소비자의 개인정보 보호에 대한 우려가 강조되는 추세이다. 이에 따라 유럽연합 집행위원회는 GDPR(General Data Protection Regulation)[1], 미국은 소비자 개인정보 보호[2]와 같은 데이터 개인정보 보호법을 시행하였다. 특히 GDPR의 경우 데이터 최소화 원칙(GDPR 5조) 및 동의(GDPR 6조)에 의하여 소비자 동의하에 데이터 수집 및 저장에 대한 처리를 절대적으로 필요한 것으로만 제한한다.

기존 중앙 집중식 학습 방식에서는 로컬 장치로부터 업로드되는 데이터에 개인정보가 포함되어있는 데이터가 존재할 경우 개인정보 보호에 대한 문제가 발생할 수 있다. 또한, 다수의 로컬 장치에서 중앙서버로 데이터를 전송하는 경우 대량의 네트워크 트래픽과 커뮤니케이션 비용이 발생한다. 이와 같은 문제점에 대해 구글에서 제안한 연합학습[3]을 이용하여 해결책을 제시한다.

연합학습은 학습을 위한 데이터를 중앙서버에서 관리하지 않으며, 다수의 로컬 클라이언트에서 로컬 데이터에 대한 학습을 진행한 로컬 업데이트를 집계하여 글로벌 모델을 학습하는 방식인 분산 데이터 학습 기술이다. 이 기술의 장점으로는 로컬 장치의 데이터를 중앙서버로 전송하지 않기 때문에 데이터에 대한 보안 보장 및 개인정보 보호가 가능하며, 모델의 업데이트만 전송하기 때문에 커뮤니케이션 비용이 줄어드는 이점이 있다. 하지만 연합학습은 학습 클라이언트가 목적에 맞는 충분한 데이터를 보유하고 있으며, 글로벌 모델을 개선하는데 정직하다고 가정하고, 기존 연합학습 방법에서는 학습을 진행하는 과정에 있어서 글로벌 모델에 위협이 될 수 있는 악성 클라이언트를 제재할 방법이 존재하지 않는 문제점을 가지고 있어 중독 공격, 추론 공격, Free-riding 등의 글로벌 모델을 손상시킬 수 있는 악성 클라이언트로 인한 공격 위협이 존재한다.

연합학습 시스템에서 신뢰할 수 있는 클라이언트를 선택하는 것은 해결해야 할 과제 중 하나이며, 대부분의 연합학습 시스템은 무작위 선택 또는 자원 조건을 통해 훈련에 참여할 클라이언트를 선택한다. 또한, 위 시스템은 훈련에 참여하는 클라이언트가 중앙서버의 권한에 의해 선택되거나 제약 조건 없이 자유롭게 모델 훈련에 참여하는 분산 방식으로 선택된다. 그러나 이러한 방법은 글로벌 모델을 손상시킬 수 있는 악의적인 클라이언트를 판단하기 위한 지표를 측정할 수 없기 때문에 악성 클라이언트에 대한 대처가 불가능한 문제가 존재하며, 학습 클라이언트에 대한 모니터링 방법이 따로 존재하지 않아 중앙 서버가 대규모의 동작을 실시간으로 모니터링할 수 없는 문제가 있다. 이에 따라 적시에 모니터링 방법이 없는 중앙 서버는 연합학습 시스템에서 신뢰할 수 없는 클라이언트나 악의적인 클라이언트를 탐지하는데 어려움이 존재한다. 결국 학습 과정에서의 악의적인 행위에 대한 정보와 악의적이거나 신뢰할 수 없는 클라이언트를 판단 가능한 정보가 부족하기 때문에 글로벌 모델 개선에 부정적인 영향을 끼칠 수 있는 악의적인 클라이언트가 지속적으로 학습에 참여할 수 있다.

본 논문에서는 연합학습 시스템에서 학습에 참여하는 클라이언트를 선택하기 위해 각 클라이언트마다 신뢰도 점수를 계산하고. 이를 기반으로 데이터 중독공격으로부터 모델의 성능을 방어하며, 신뢰할 수 있는 연합학습을 위한 클라이언트를 선택할 수 있는 알고리즘을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 배경과 관련 연구에 대해 설명하고, 3장에서는 제안하는 모델에 대해 설명한다. 4장에서는 제안하는 모델을 적용한 실험 결과를 제시하며, 마지막으로 5장에서는 결론을 맺는다.

Ⅱ. 배경 및 관련 연구

2-1 연합학습

머신러닝 기술이 발전하면서 개인정보 보호와 데이터 기밀유지에 대한 문제점이 부각되었다. 이에 대응하기 위해 2018년 유럽연합은 일반 데이터 보호규정(GDPR)을, 2020년 캘리포니아는 소비자 개인정보 보호법을 제정하여 데이터 보호와 보안을 강조하고 있다. 이에 따라 기존의 머신러닝 환경에서 발생할 수 있는 개인정보 보호 및 보안법률을 준수하는 방법인 데이터 통합 방법의 연합학습이 제안되었다.

연합학습[4]은 2016년 구글에서 처음 제안되어 여러 모바일 장치에서 공동으로 학습하기 위해 구글 키보드에 처음 활용하였고, 저장된 데이터를 공유하지 않고 인공지능 모델을 학습할 수 있는 분산형 기계 학습 기법이다. 연합학습의 과정은 중앙서버가 모델을 클라이언트에 보내면 각 클라이언트는 로컬 데이터로 이 모델을 훈련 시키고, 훈련된 로컬 모델의 파라미터를 중앙서버로 전송한다. 중앙서버는 이러한 과정을 반복하여 로컬 모델의 정보를 결합하여 글로벌 모델을 업데이트한다. 이렇게 개선된 글로벌 모델은 다시 로컬 장치에 전송되어 기존 모델을 업데이트함으로써 글로벌 모델의 성능을 점점 향상시킨다. 그림 1은 일반적인 연합학습에서 나타나는 프로세스이다.

1단계 : 서버는 학습에 참여하는 클라이언트에게 글로벌 모델을 전달한다.

2단계 : 각 클라이언트는 전달받은 모델을 로컬 데이터를 활용하여 학습시킨다.

3단계 : 클라이언트는 학습을 통해 업데이트된 모델의 파라미터만 서버로 전송한다.

4단계 : 서버는 각 로컬로부터 전달받은 파라미터를 집계하여 글로벌 모델을 업데이트한다.

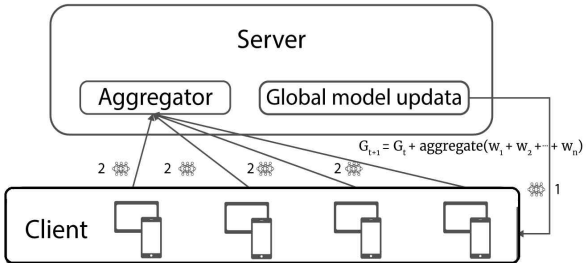


그림 1. 연합학습 프로세스
Fig. 1. Federated learning process

2-2 중독공격

연합학습은 기존 머신러닝과 비교했을 때 학습에 사용되는 데이터를 로컬 장치에서 관리하여 데이터를 공유하지 않는 점에서 개인 데이터 보호와 관련된 개인정보 보호 보장이 제공되는 이점이 있다. 그러나 반복적인 학습 과정과 모델 업데이트 교환 과정에서 악의적인 공격자에게 시스템이 노출될 수 있다. 이 과정에서 중독공격이 발생할 수 있는데, 중독공격은 연합학습에서 가장 일반적으로 사용되는 공격 기술 중 하나이다.

중독공격에서 악의적인 클라이언트는 조작된 데이터를 학습에 사용하여 글로벌 모델에 영향을 미치게 하며, 공격자의 목표에 따라 무작위 공격과 표적 공격으로 나뉜다[5]. 무작위 공격의 목표는 연합학습 모델의 정확도를 낮추게 하는 것이고, 표적 공격은 연합학습 모델이 공격자가 지정한 목표 레이블을 출력하도록 하는 것을 목표로 한다. 중독공격은 학습 단계 동안 모델 또는 데이터에 대해 수행될 수 있다. 중독공격은 공격자의 유형에 따라 데이터 중독공격과 모델 중독공격으로 분류될 수 있다[6].

데이터 중독공격은 공격자가 기존의 학습 데이터를 수정하거나 잘못된 레이블을 붙여서 새로운 중독 데이터 샘플을 만들어내는 공격이다[7]. 이를 통해 악의적인 클라이언트는 연합학습에서 공유되는 글로벌 모델의 결과를 왜곡하거나 특정한 방향으로 유도할 수 있다. 이와 같은 공격을 label-flipping이라 칭하며, 이는 학습 데이터의 레이블을 원래의 레이블 값이 아닌 다른 값으로 변경하는 데이터 중독공격이다.

[8]의 저자는 sybil 공격에 대응하여 FoolsGold라는 방법을 제안하였으며, sybil 기반 데이터 중독공격이 연합학습 시스템에 미치는 영향에 대해 연구를 진행하였다. 위 연구에서 악의적인 클라이언트는 학습 과정에서 중독된 데이터 샘플을 생성하여 공격을 시도한다. 실험 결과에 따르면 두명의 악의적인 클라이언트만으로도 95% 이상의 잘못된 분류를 달성할 수 있음을 보였다. 이에 따라 데이터 중독공격은 모든 클라이언트로부터 수행될 수 있기 때문에 시스템 전체를 위협할 수 있다는 결과를 보여준다.

모델 중독공격은 학습 프로세스 중에 클라이언트와 연합학습 서버 간에 공유되는 업데이트에 영향을 미치는 공격이다. 공격자는 로컬 모델의 업데이트를 수정하여 글로벌 모델의 업데이트를 방해하거나 왜곡시키는 것을 목표로 하여 데이터 중독공격과 달리 적은 수의 공격자만으로도 전체 글로벌 모델을 손상시킬 수 있다. 또한, 연합학습에서 모델의 업데이트만 공유하는 특성으로 인해 공격자를 특정하기가 어렵다. 이에 따라 모델의 참여자가 많은 대규모 연합학습에서 더 위협적일 수 있음을 확인하였다.

[9]의 연구에서는 탐지를 회피하기 위해 공격 목표과 학습 손실에 대해 번갈아 최적화하고, 공격 스텔스를 증가시키며, 정상 클라이언트의 업데이트에 대한 매개변수 추정을 사용함으로써 연합학습 환경에서 모델 중독공격이 데이터 중독공격보다 훨씬 더 효과적이라는 것을 입증했다. 따라서, 연합학습 환경에서는 데이터 중독공격과 모델 중독공격 모두 매우 심각한 보안 위협으로 인식되어야 하며, 이러한 공격에 대한 방어 및 대응 방법이 개발되어야 한다.

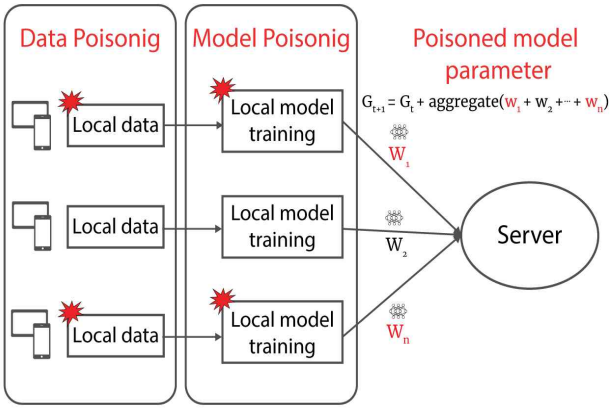


그림 2. 데이터 중독과 모델 중독
Fig. 2. Data poisoning and model poisoning

2-3 관련 연구

Liu 등[10]은 블록체인 기반 보안 연합학습 프레임워크를 제안하여 5G 네트워크에서 연합학습의 보안을 보장하기 위한 데이터 프라이버시 유출 문제에 대한 해결책을 제시하였다. 위 논문에 제안된 프레임워크에서는 블록체인의 스마트 컨트랙트를 실행하여 악의적이고 신뢰할 수 없는 참가자를 인식하여 중독공격을 방어하고, 클라이언트에 로컬 차등 프라이버시 기술을 적용하여 노이즈를 모델 매개변수에 추가 함으로써 공격자가 모델 매개변수에 접근 하더라도 원래의 모델 매개변수와 로컬 데이터를 복구할 수 없도록 하여 멤버십 추론 공격 방지를 시도하였다.

Mugunthan 등[11]은 블록체인과 스마트 컨트랙트의 분산 특징을 활용하였고, 개인정보 보호를 위해 각 클라이언트에 대한 평가를 진행하여 측정된 평가 점수를 기반으로 인센티브를 제공함으로써 책임있는 연합학습 프레임워크를 제안하였다. 해당 프레임워크의 목표는 로컬 데이터 세트의 개인정보 보호와 악의적인 공격자에 대한 대응 기여도에 비례하여 보상하는 것이며, 차등 프라이버시 도입과 모델 기여를 위한 새로운 감사메커니즘이 도입되었다. 평가점수는 클라이언트의 모델을 공유하여 정확도를 평가하고, 다른 클라이언트들로부터 측정된 점수의 중앙값을 구하여 평가점수를 설정한다. 만약 측정된 중앙값이 낮으면 해당 클라이언트는 성능이 낮은 모델을 등록했다고 간주되어 부정적인 영향을 받게 된다. 제안된 프레임워크는 중앙 집중식 검증 데이터 세트나 신뢰할 수 있는 감사자가 필요하지 않으며, 완전히 분산되어 있어 악의적인 공격자가 존재하는 상황에 대하여 최대 50%의 담합 공격에 대해 내성을 가진다.

LO Sin Kit 등[12]은 블록체인 기반의 연합학습 시스템을 제안하여 신뢰할 수 있는 연합학습의 책임성 문제를 해결하고자 했다. 블록체인의 투명성 특성과 불변성 특성을 활용하여 감사 가능성을 보장하고, 시스템의 책임성을 향상시켰다. 학습 라운드에서 클라이언트는 데이터 버전과 로컬 모델 매개변수을 데이터 모델 레지스트리 스마트 컨트랙트에 해시값으로 업로드하며, 중앙서버는 해시된 글로벌 모델 매개변수를 스마트 컨트랙트로 전송한다. 스마트 컨트랙트에서는 로컬 및 글로벌 모델 매개변수의 해시 값이 구조모델에 기록되며, 업로드한 정보는 수정할 수 없다. 블록체인을 사용하여 로컬 및 글로벌 모델 버전의 해시 값을 저장하면 데이터 모델의 출처를 얻을 수 있고, 사용자는 학습 모델 성능을 감사할 수 있다. 이러한 작업 로그는 블록체인의 고유한 특징인 변조방지 설계로 인해 수정이나 제거가 불가능하며, 이는 연합학습의 감사 추적에 대한 증거를 제공함과 동시에 블록체인 내 책임성을 보장하고 시스템의 신뢰성을 향상시킬 수 있음을 의미한다.

Yuzheng Li 등[13]은 악의적인 목적을 가진 클라이언트의 공격, 중앙 서버의 글로벌 모델에 대한 공격, 사용자 개인정보 데이터에 대한 지속적인 공격 등 기존 연합학습의 보안 문제를 해결하기 위해 위원회 합의가 있는 블록체인 기반 분산형 연합학습 프레임워크(BFLC)를 제안하였다. 이 프레임워크는 스마트 컨트랙트와 트랜잭션을 통해 중앙 서버의 기능을 대체하여 구현하였다. BFLC에서 블록체인은 글로벌 모델 블록과 로컬 모델 블록으로 구성되며, 학습 참여 클라이언트 중 평판 점수가 높은 소수의 클라이언트를 선출하여 로컬 모델을 검증하는 위원회를 구성하였다. BFLC의 학습 성능은 개인정보 보호 기능을 갖추면서 약 1% 이내로 근접하는 성능을 보이며, 기존 연합학습 방식과 비교하였을 때, 보안 문제를 효과적으로 해결할 수 있다는 장점이 있다.

현재 연합학습에서는 [10]-[13]의 선행 연구와 같이 블록체인의 신뢰성과 스마트 컨트랙트의 자동화를 이용하여 중앙 서버를 대체하는 연합학습 프레임워크에 대한 연구가 진행되고 있다. 그러나 중앙 서버가 없는 연합학습 시스템에서는 글로벌 모델을 업데이트하기 위해 모델이 블록체인에 업로드되고, 각 클라이언트 간 모델을 공유해야 한다. 이로 인해 악성 클라이언트는 쉽게 정상 모델을 획득할 수 있는 문제가 발생한다.

Ⅲ. 신뢰도 점수 기반 연합학습

3-1 신뢰도 점수 기반 연합학습 시스템 구조

그림 3은 최종 글로벌 모델을 업데이트하기 전 가중치를 취합하는 단계에서 그림 1의 일반적인 연합학습 프로세서와 달리 각 클라이언트에 대한 신뢰도 점수를 계산하고, 이는 클라이언트를 선택하는 과정이 존재함을 보여준다. 신뢰도 점수는 글로벌 모델 업데이트에 반영되는 클라이언트의 가중치 비율을 결정하는 데 사용된다. 이 신뢰도 점수는 각 학습 데이터의 품질과 성능 기여도를 고려하여 계산되며, 악성으로 판단되는 클라이언트나 성능이 낮고 기여도가 적은 클라이언트는 모델 학습에서 제외되어 모델의 성능을 향상시킬 수 있다.

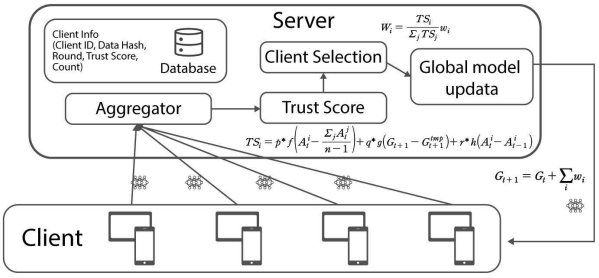


그림 3. 신뢰도 점수 기반 연합학습 시스템 구조
Fig. 3. Trust score based federated learning system architecture

3-2 신뢰도 점수

연합학습에 참여하는 클라이언트는 신뢰도 점수를 통해 선택된다. 신뢰도 점수는 각 라운드에서 평가 정확도를 사용하여 계산한다. 신뢰도 점수 계산은 식 (3), (4), (5)으로 구성되며 최종적인 신뢰도 점수는 수식 (6.a)과 (6.b)를 통해 계산된다.

$$f(A_t^i - \frac{\sum_{j}^{n-1} A_t^j}{n-1}) \tag{3}$$

$$f(x) = \begin{cases} x, & |x| > \epsilon_p \\ 0, & else \end{cases} \tag{3.a}$$

식 (3)는 다른 클라이언트의 평균 정확도와 비교하여 클라이언트의 성능을 평가하기 위해 사용된다. 식 (3)을 사용하여

계산된 값이 양수인 경우, 해당 클라이언트는 평균 이상의 성능을 보이는 정상 클라이언트로 간주 되고, 계산된 값이 음수인 경우에는 악성 클라이언트로 간주한다. 그러나 평가된 클라이언트는 다른 클라이언트의 평균 성능과 비교하기 때문에 정상 클라이언트 임에도 평균 성능 이하 가능성이 반드시 존재한다. 이에 따라 식 (3.a)를 통해 정상 클라이언트의 성능이 ϵ_p 이내의 오차범위 내에 있을 경우 그 값을 0으로 처리하여 해당 정상 클라이언트가 악성 클라이언트로 간주 되는 것을 방지한다.

$$g(G_{t+1} - G_{t+1}^{tmp}) \tag{4}$$

$$g(x) = \begin{cases} x, & |x| > \epsilon_q \\ 0, & else \end{cases} \tag{4.a}$$

식 (4)는 신뢰도 점수 계산의 대상이 되는 클라이언트를 제외하고 이를 취합한 임시 모델을 생성한다. 또한, 글로벌 모델의 업데이트를 위해 전체 클라이언트에 대한 모델을 취합하고 반영했을 때의 결과를 비교한다. 이 요소는 양의 값일 경우 글로벌 모델 개선에 긍정적인 영향을 주는 클라이언트라 판단할 수 있고, 최종 글로벌 모델을 생성할 때 각 클라이언트가 모델에 미치는 영향을 판단할 수 있다. 식 (4)의 값은 평균보다 높은 성능을 보이는 클라이언트의 가중치를 포함하고 있고, 취합된 글로벌 모델이 항상 더 나은 성능을 보이지 않는 경우가 존재하기 때문에 식 (4.a)와 같이 ϵ_q 이내의 오차 범위 안의 결과값을 0으로 처리하여 낮은 성능을 보이는 정상 클라이언트가 악성 클라이언트로 분류되는 상황을 방지한다.

$$h(A_t^i - A_{t-1}^i) \tag{5}$$

$$h(x) = \begin{cases} x, & |x| > \epsilon_r \\ 0, & else \end{cases} \tag{5.a}$$

식 (5)는 학습 클라이언트의 데이터가 변경된 상황(U =F*alse*)에서만 반영되고, 이전 라운드에서의 정확도와 비교한다. 식 (5)의 수치에 따라 모델 성능개선에 기여 가능한 데이터의 추가 여부를 판단할 수 있다, 그러나 정상 데이터가 추가되었더라도 추가된 데이터는 이전까지 학습된 글로벌 모델에 사용하지 않았던 새로운 데이터이기 때문에 정상 데이터임에도 성능이 감소하는 상황이 발생할 수 있다. 이에 따라 식 (5.a)를 통해 ϵ_r 이내의 오차 범위 안의 결과값을 0으로 처리하여 새로운 정상 데이터가 사용되는 상황을 고려하도록 한다.

$$TS_i = p^* f(A_t^i - \frac{\sum_j^{n-1} A_t^j}{n-1}) + q^* g(G_{t+1} - G_{t+1}^{tmp}) \tag{6.a}$$

$$TS_i = p^* f(A_t^i - \frac{\sum_j^{n-1} A_t^j}{n-1}) + q^* g(G_{t+1} - G_{t+1}^{tmp}) + r^* h(A_t^i - A_{t-1}^i) \tag{6.b}$$

식 (3), (4), (5)을 반영한 최종 신뢰도 점수 수식은 클라이언트의 데이터가 이전 라운드의 학습 데이터와 비교하여 동일한 경우(U =T*rue*)는 식 (6.a), 동일하지 않는 경우(U =F*alse*)는 식 (6.b)를 통해 계산된다. 각 클라이언트에 부여되는 신뢰도 점수는 라운드마다 학습을 마치고 글로벌 모델 업데이트를 위한 가중치 집계 과정에서 가중치 반영 비율을 결정하기 위해 사용된다.

Algorithm 1 Trust Score based Federated Learning (TSFL) Algorithm

Input:
Client Trust Score : TS
Client Set : C
Check Client Data Change : U
Client Accuracy : A^i
Training round : $t \in r$
Opportunity to engage in learning : c

$Count_i \leftarrow 0$
for $t \in r$ **do**
  for all client $C_i \in C$ **do**
    $TS_i \leftarrow 0$
    if U is $TRUE$ **then**
      $TS_i \leftarrow p * f(A_t^i - \frac{\sum_{j=1}^{n-1} A_t^j}{n-1}) + q * g(G_{t+1} - G_{t+1}^{tmp})$
    else
      $TS_i \leftarrow p * f(A_t^i - \frac{\sum_{j=1}^{n-1} A_t^j}{n-1}) + q * g(G_{t+1} - G_{t+1}^{tmp}) + r * h(A_t^i - A_{t-1}^i)$
    end if
    if $TS_i < 0$ **then**
      $Count_i \leftarrow Count_i + 1$
      if $Count_i > c$ **then**
        Eliminate C_i in C
      end if
    else
      if $Count_i > 0$ **then**
        $Count_i \leftarrow Count_i - 0.5$
      end if
      $w_i \leftarrow \frac{TS_i}{\sum_k TS_k} w_i$
    end if
    $G_{t+1} \leftarrow \sum_i w_i$
  end for
end for

그림 4. 신뢰도 점수 기반 연합학습 알고리즘
Fig. 4. Trust Score based Federated Learning(TSFL) algorithm

그림 4의 신뢰도 점수 기반 연합학습 알고리즘은 학습에 참여하는 클라이언트에서 가중치와 정확도를 기준으로 신뢰도 점수를 결정하고, 계산된 점수를 통해 가중치 취합 비율과 다음 학습 라운드에 참여할 클라이언트를 결정한다. 신뢰도 점수를 계산하고 학습하는 과정은 다음과 같다. 연합학습을 시작하기 전 참여하는 모든 클라이언트의 학습 참여 기회 C*ounti*와 신뢰도 점수 TS*i*는 0으로 초기화하며, TS*i*값은 매 라운드가 시작할때 0으로 초기화한다. 신뢰도 점수 TS*i*를 계산하기 전에 각 클라이언트의 이전 라운드와 현재 라운드 학습에 사용한 데이터의 변경 여부를 확인한다. 데이터의 변경 여부에 따라 식 (6.a)와 식 (6.b)의 수식을 반영하여 각 클라이

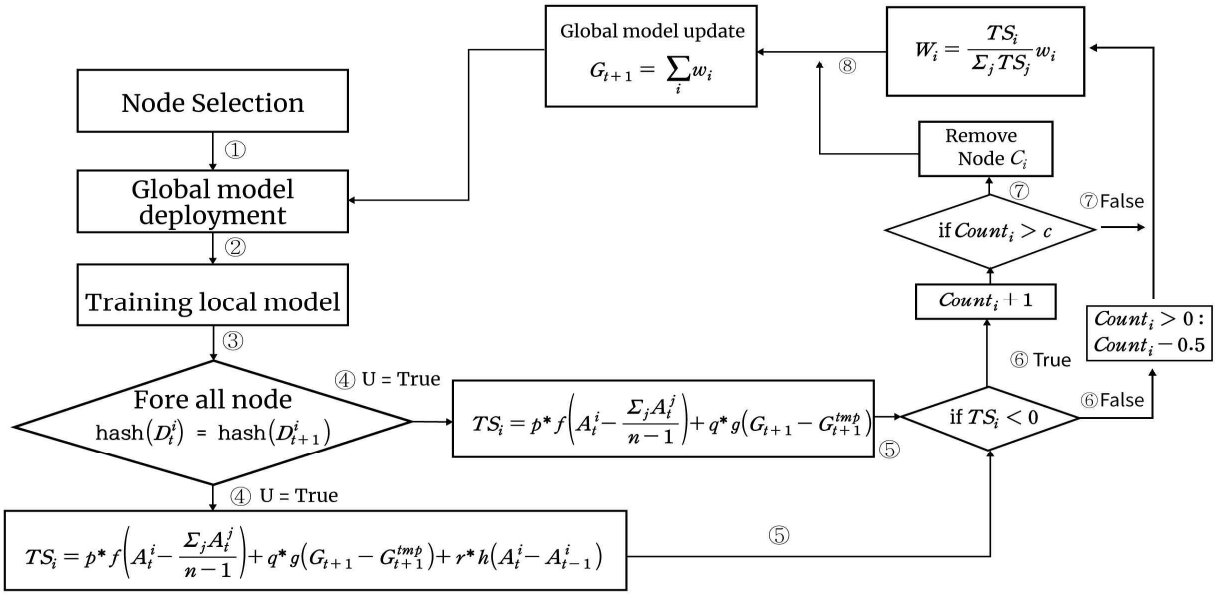


그림 5. 신뢰도 점수 기반 연합학습 프로세스
Fig. 5. Trust score based federated learning process

언트의 신뢰도 점수 TS*i*를 계산한다. 신뢰도 점수 TS*i*가 0보다 작은 값을 갖는 클라이언트는 C*ounti*값에 1을 더하고, C*ounti*값이 학습 참여 기회 임계치 *c*값을 넘어가는 클라이언트는 연합학습에 참여할 수 없도록 학습 참여 클라이언트 집합에서 제거한다. 신뢰도 점수 TS*i*가 0보다 큰 클라이언트 집합에서 C*ounti*값이 0보다 큰 클라이언트는 학습에 긍정적인 기여를 한 보상으로 C*ounti*값이 0보다 큰 경우 0.5만큼 차감시켜 정상적인 클라이언트임에도 악성 클라이언트로 분류되어 학습에서 제외되는 상황을 방지한다. 신뢰도 점수 TS*i*에 따라 각 가중치는 재조정되며, 최종 글로벌 모델은 재조정된 가중치를 모두 더한 값이 된다. 이때, 신뢰도 점수 TS*i*가 0보다 작지만 C*ounti*값이 학습 참여 기회 임계치를 넘지 않는 가중치는 글로벌 모델을 업데이트하는데 포함되지 않는다. 그러나 다음 학습 라운드에서 글로벌 모델을 받아 학습에 참여할 수 있다.

3-3 신뢰도 점수 기반 연합학습 프로세스

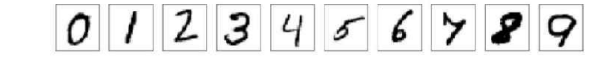
신뢰도 점수 기반 연합학습은 일반적인 학습 프로세스와 비교하여 로컬 모델을 학습하고 최종 글로벌 모델을 업데이트하는 과정 사이에 신뢰도 점수를 계산하여, 다음 학습 라운드에 참여할 클라이언트를 결정하는 과정이 존재한다. 그림 5의 신뢰도 점수 기반 연합학습의 프로세스는 다음과 같다.

① : 서버는 학습에 참여할 클라이언트를 선택한다.
② : ①에서 선택한 클라이언트에 글로벌 모델을 배포한다.
③ : 각 클라이언트는 ②에서 전달받은 모델을 보유한 데이터로 학습하고, 도출된 가중치를 서버에 전송한다.
④ : 서버는 클라이언트의 신뢰도 점수를 계산하기 전에 각 클라이언트의 데이터 변경 여부를 확인한다. 이를 위해 이전 라운드와 현재 라운드에서 학습에 사용된 데이터의 해시값을 비교한다.
⑤ : ④에서 확인된 데이터 변경 여부에 따라 식 (6.a)와 (6.b)를 적용하여 신뢰도 점수 TS*i*를 계산한다.
⑥ : TS*i* 가 0보다 작은 값을 가지면, 클라이언트 C*i*는 모델을 개선에 긍정적 기여를 하지 못하는 클라이언트로 판단하고 C*ounti*에 1을 더한다. TS*i*가 0보다 크거나 같은 값을 가질 때 학습에 긍정적 기여를 한 보상으로 클라이언트의 C*ounti*값이 0보다 큰 경우 0.5만큼 차감하여 누적된 C*ounti*값으로 인해 정상적인 클라이언트가 악성 클라이언트로 분류되어 학습에서 제외되는 상황을 방지한다.
⑦ : C*ounti*을 학습 참여 기회에 대한 임계치 값인 *c*와 비교한다. TS*i*가 0보다 작은 값을 갖더라도 C*ounti*값이 *c*값을 초과하지 않는 경우에 다음 학습 라운드에 참여할 수 있도록 하고, C*ounti*값이 *c*값을 초과하는 클라이언트는 라운드가 진행되는 동안 긍정적 기여를 하지 못했다고 판단하고 학습에 참여할 수 없도록 제거한다.
⑧ : 글로벌 모델을 업데이트할 때, TS*i*값이 임계치인 0보다 큰 클라이언트의 가중치만 집계한다. TS*i*는 글로벌 모델을 업데이트할 때, 각 클라이언트의 가중치 반영 비율을 결정하는데 사용한다.

Ⅳ. 실험 및 결과

4-1 실험 환경

블록체인을 활용한 평가점수 기반 연합학습 BlockFLow [13]과 FedAVG 기반의 연합학습 방법, TSFL 방법을 이용하여 데이터 중독공격을 수행하는 악성 클라이언트 여부에 따른 학습 정확도를 비교 및 평가하였다. 블록체인을 활용한 평가점수 기반 연합학습 BlockFLow에서는 모든 클라이언트가 자신이 보유한 데이터셋을 이용하여 다른 모델을 평가한다. 평가점수는 모델을 공유한 각 클라이언트의 로컬 모델에 대해 보고된 점수의 중앙값과 보고한 점수, 중앙값 사이의 최대 차이의 역수 중 더 작은 값으로 결정된다. 만약 클라이언트의 모델 성능이 낮으면 보고된 점수의 중앙값이 낮을 것이라 보인다. 또한, 거짓으로 부정확한 평가를 보고하는 경우에 보고한 점수와 중앙값의 차이가 큰 차이를 보이기 때문에 학습에 부정적인 영향을 미칠 수 있다. 이때 보고한 점수와 중앙값 사이의 역수는 역수를 구하는 대신 $\max(0, \frac{0.5-x}{0.5+x})$를 취해 보고된 점수의 중앙값과 보고한 점수가 0.5 이상 벌어지는 경우 0으로 만들어 올바른 평가를 제출하도록 유도한다. 연합학습 환경으로 학습에 참여하는 10개의 클라이언트 구성과 15번의 라운드를 진행하였다. 실험에는 Fashion-MNIST 데이터와 MNIST 데이터를 사용하였으며, 악성 데이터는 다른 레이블로 지정하는 label-flipping 방식으로 학습과 관계없는 데이터를 학습에 사용하였다. 또한, 모든 클라이언트는 3000개의 데이터를 보유한 상태로 학습에 참여하고, 특정 라운드에 각 500개의 데이터가 추가되여, 최종적으로 6000개의 데이터를 갖는 상황을 구성하였다.



정상 레이블	0	1	2	3	4	5	6	7	8	9
S1	0	7	2	3	4	5	6	1	8	9
S2	8	7	3	2	4	5	6	1	0	9
S3	8	7	3	2	9	6	5	1	0	4
S4	Fashion-Mnist Dataset									

그림 6. 시나리오별 데이터 및 레이블
Fig. 6. Scenario specific data and labels

그림 6의 시나리오 S1, S2, S3는 label-flipping 공격방법을 활용하였고, 공격에서 사용되는 데이터 수의 차이에 따른 모델 성능을 비교하기 위한 시나리오로 사용하였다. 반면에 S4는 label-flipping 공격방법이 아닌 학습과 관련 없는 데이터를 활용한 방법을 활용하였으며, 공격에서 사용되는 데이터의 개수가 S3과 동일하지만, 모델과 관련된 데이터인지에 대한 여부에 따른 성능 비교를 위한 시나리오로 사용하였다.

4-2 실험 결과 분석

1) 데이터 중독 공격 대응 성능 평가

그림 7은 그림 6의 시나리오별 악성 클라이언트의 비율에 따라 FedAVG, BLockFLow, 그리고 제안하는 TSFL모델의 성능을 비교한 그래프이다. 시나리오 별로 데이터 중독 공격의 비율이 증가할수록 연합학습 모델의 정확도가 떨어지는 것을 확인할 수 있었다. 그리고 시나리오 별로 공격방법 및 공격 비율이 다르지만, 전반적으로 제안하는 TSFL모델이 데이터 중독공격에 대해 효과적으로 대응할 수 있음을 확인할 수 있었다. 특히, 모든 시나리오(S1, S2, S3, S4)에서 악성 클라이언트의 비율이 전체 클라이언트의 30%인 경우에 제안하는 TSFL 방법은 FedAVG방법에 비해 최대 5%이상의 성능 향상을 보임을 확인할 수 있었다.

FedAVG 방법의 경우 글로벌 모델을 업데이트하는 과정에서 학습에 참여하는 모든 클라이언트의 가중치를 아무 조치 없이 모두 반영하기 때문에 실험에서 글로벌 모델의 성능을 저하시키는 악성 클라이언트가 존재함에도 해당 클라이언트의 가중치를 업데이트에 포함시켜 모델 성능이 크게 감소하는 결과를 보였다. FedAVG 방법은 악성 클라이언트의 비율이 증가할수록 최대 약 5% 이상 모델의 성능이 하락하여 악성 클라이언트에 대해 매우 취약한 점을 확인할 수 있다.

그에 반해 제안하는 TSFL 방법은 악성 클라이언트의 비율이 증가하더라도 약 3% 이내로 악성 클라이언트가 없는 상황에 근접하는 성능을 보인다. BlockFLow의 방법과 비교하였을 때, 악성 클라이언트가 존재에 대해 조금 더 성능이 개선되었으며, 각 클라이언트 사이에 모델을 공유하지 않음으로 악성 클라이언트가 정상 클라이언트를 복제 또는 참고할 수 있는 문제를 방지할 수 있다.

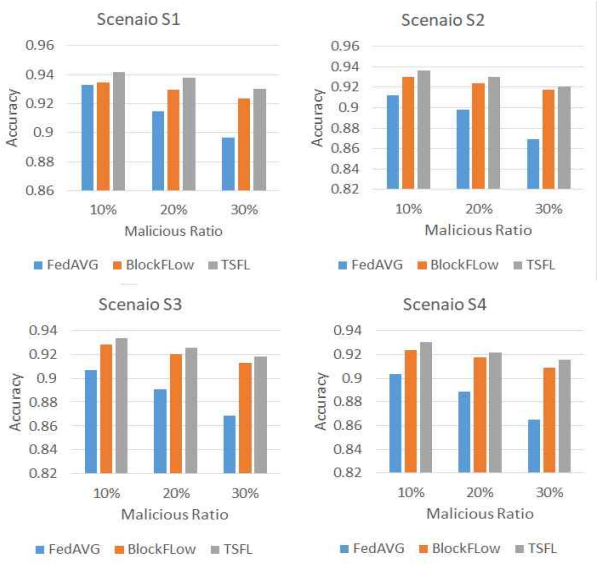


그림 7. 시나리오별 악성 클라이언트 비율에 따른 성능 비교
Fig. 7. Performance comparison according to malicious client ratio by scenario

2) 글로벌 모델 취합 소요 시간 비교

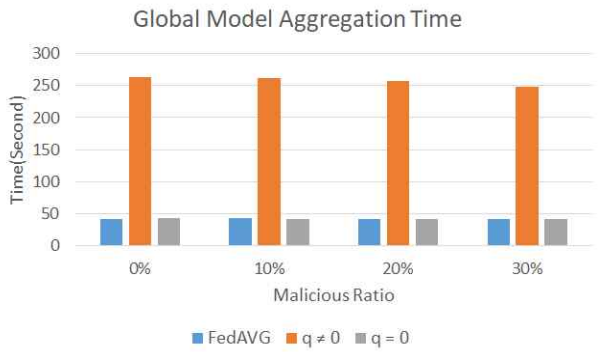


그림 8. FedAVG 와 TSFL 글로벌 모델 집계 시간 비교
Fig. 8. FedAVG vs TSFL global model aggregation time comparison

그림 8은 TSFL와 FedAVG 방법을 적용하였을 때, 글로벌 모델을 생성 시 소요되는 시간을 비교한 그래프이다. 제안하는 TSFL 방법은 q값의 반영 여부에 따라서 $q = 0$인 경우 임시 모델을 생성하지 않기 때문에 최종 글로벌 모델 업데이트하기까지 소요되는 시간이 FedAVG의 글로벌 모델 취합 소요 시간과 비교하였을 때 비슷하였으며, 악성 노드의 비율이 늘어날수록 학습에 참여할 수 있는 클라이언트 수가 적어져,이에 따라 소요되는 시간이 감소하는 결과를 확인할 수 있다. 반면, $q \neq 0$인 경우 신뢰도 점수 계산 단계에서 임시 모델을 생성하고 결과를 비교하는 과정이 추가됨으로 FedAVG와 비교하였을 때, 약 5배 이상의 시간이 소요되는 점을 확인할 수 있다. 그러나 이는 모델에 성능을 방어하기 위한 과정으로 학습 소요 시간과 모델 성능을 교환하였기 때문에 부정적으로만 보기는 어렵다.

3) p, q, r 비율에 따른 TSFL 성능 비교

	p	q	r
Case 1	1	0	0
Case 2	0	1	0
Case 3	1/3	1/3	1/3
Case 4	1/4	1/4	1/2
Case 5	1/2	0	1/2

	Case 1	Case 2	Case 3	Case 4	Case 5
S1	0.9291	0.9302	0.9392	0.9378	0.9328
S2	0.9223	0.9241	0.9286	0.9297	0.9247
S3	0.9154	0.9165	0.9239	0.9253	0.9201
S4	0.9076	0.9092	0.9195	0.9213	0.9176

그림 9. 사례별 p, q, r 비율 및 시나리오 성과에 따른 사례 분류
Fig. 9. Case classification according to p, q, r ratio and scenario performance per case

그림 10은 전체 노드의 20%가 악의적인 클라이언트인 상황에서 그림 9의 Case1과 Case2는 식 (6.a) Case3, Case4, Case5는 (6.b)의 (p, q, r) 비율에 따라 신뢰도 점수를 계산하였을 때의 모델 성능을 나타낸다. 각 요소를 개별로 반영한 경우에 반해 각 요소를 같이 반영하여 모델을 취합했을 경우에 더 좋은 성능을 나타내는 것을 확인할 수 있다. 식 (3.a), (4.a), (5.a)의 ϵ_p, ϵ_q, ϵ_r값은 각각 0.03, 0.015, 0.05를 적용하여 실험을 진행하였다. 또한, Fig 10에서 각 시나리오 별 (p, q, r)의 비율에 대한 Case를 적용하였을 때, 4번 Case의 비율이 1번 시나리오를 제외한 상황에서 가장 좋은 성능이 나타남을 확인하여, 식 (6.a)와 (6.b)의 (p, q), (p, q, r) 값은 각각 1/2과 Case4에 대한 비율을 적용하였다.

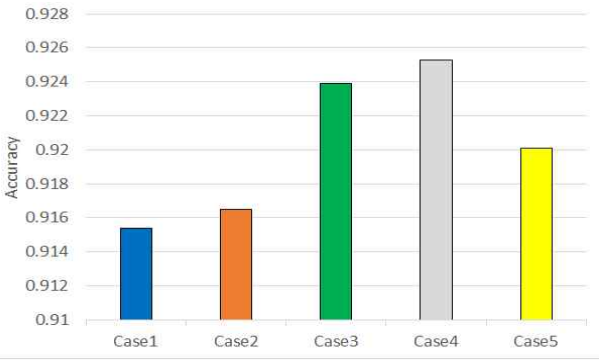


그림 10. p, q, r 비율의 적용에 따른 성능
Fig. 10. Performance as a function of p, q, r ratio

4) 학습 참여 기회 임계치에 따른 성능 비교

정상적인 클라이언트임에도 학습을 진행하는 과정에서 모델의 정확도가 감소하는 상황이 발생하여, 클라이언트의 신뢰도 점수가 임계치보다 낮은 점수를 획득할 수 있다. 이러한 경우에 대한 평가를 진행하기 전, 학습 클라이언트가 학습에 참여할 수 있는 기회에 대한 수치 c값을 설정하기 위한 실험을 진행하였다. 학습 참여 기회를 부여하면 글로벌 모델 업데이트 진행 시 해당 클라이언트의 가중치가 집계에서만 제외되기 때문에 모델 성능 개선에 기여를 하지 못하더라도 업데이트된 글로벌 모델을 획득할 수 있다. 20%의 악성 클라이언트가 존재하는 상황에서 그림 6의 시나리오 S3을 적용한 실험 결과는 그림 11에 보여진다. 참여 클라이언트에 대해 2번의 학습 참여 기회를 부여하였을 때 92.53%의 가장 나은 성능을 보였으며, 이를 기반으로 성능 평가를 진행할 때 $c = 2$로 설정하고 실험을 진행하였다.

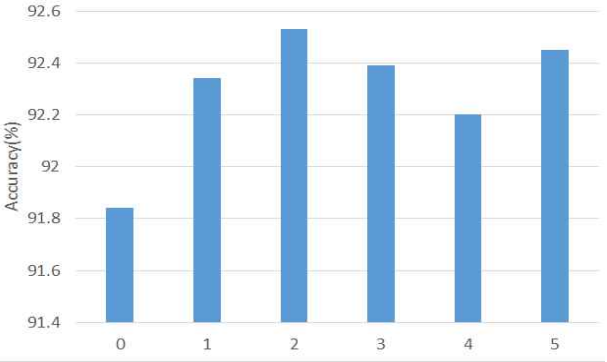


그림 11. 클라이언트의 학습 참여 기회에 따른 성과
Fig. 11. Performance according to the client's opportunity to participate in learning threshold

V. 결 론

본 논문에서는 연합학습에서 악성 클라이언트가 존재하는 경우 클라이언트의 신뢰도 점수에 따라 글로벌 모델의 성능에 부정적인 영향을 미치는 악성 클라이언트를 판단하고 글로벌 모델의 성능을 향상시킬 수 있는 신뢰도 점수 기반의 연합학습 방법을 제안하였다.

신뢰도 점수의 임계치 미만으로 인해 글로벌 모델 성능 개선에 기여 하지 못했더라도 학습에 참여할 수 있는 기회를 부여하여 학습에서 배제될 수 있는 상황을 개선하였다. MNIST 데이터와 Fashion-MNIST 데이터를 사용하여 FedAVG 방법, BlockFLow의 연구방법과 제안하는 TSFL 방법에 대해 성능 평가를 진행하여 신뢰도 점수 기반의 연합학습 방법이 일반적인 연합학습에서 보편적으로 사용되는 FedAVG 방법에 비해 전체 학습 참여 클라이언트 중 악성 클라이언트의 비율이 늘어나는 경우에 최대 약 7% 이상의 개선된 성능을 달성하는 결과를 확인하였다. 또한, 중앙 서버가 없는 블록체인 기반 연합학습 환경에서 클라이언트 간 모델 공유로 인해 발생할 수 있는 문제점을 블록체인을 활용하지 않음으로써 사전에 방지하는 해결책을 제시하였고 블록체인을 활용한 연합학습 프레임워크보다 더욱 개선된 성능을 확인하였다[14].

감사의 글

본 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임. (IITP-2022-0-01203) 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 지역지능화혁신인재양성사업의 연구결과로 수행되었음 (IITP-2023-RS-2022-00156287)

참고문헌

[1] B. Custers, A. M. Sears, F. Dechesne, I. Georgieva, T. Tani, and S. van der Hof, *EU Personal Data Protection in Policy and Practice*, Hague, Netherlands: TMC Asser Press, 2019.

[2] B. M. Gaff, H. E. Sussman, and J. Geeter, “Privacy and Big Data,” *Computer*, Vol. 47, No. 6, pp. 7-9, June 2014. https://doi.org/10.1109/MC.2014.161

[3] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient Learning of Deep Networks from Decentralized Data,” arXiv:1602.05629, February 2016. https://doi.org/10.48550/arXiv.1602.05629

[4] Q. Yang, Y. Liu, T. Chen, and Y. Tong, “Federated Machine Learning: Concept and Applications,” arXiv: 1902.04885, February 2019. https://doi.org/10.48550/arXiv.1902.04885

[5] L. Huang, A. D. Joseph, B. Nelson, B. I. P. Rubinstein, and J. D. Tygar, “Adversarial Machine Learning,” in *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence (AISec ’11)*, New York: NY, pp. 43-58, October 2011. https://doi.org/10.1145/2046684.2046692

[6] V. Shejwalkar and A. Houmansadr, “Manipulating the Byzantine: Optimizing Model Poisoning Attacks and Defenses for Federated Learning,” in *Proceedings of the Network and Distributed System Security (NDSS) Symposium 2021*, San Diego: CA, pp. 21-24, February 2021. https://doi.org/10.14722/ndss.2021.24498

[7] V. Tolpegin, S. Truex, M. E. Gursoy, and L. Liu, “Data Poisoning Attacks against Federated Learning Systems,” in *Proceedings of the 25th European Symposium on Research in Computer Security (ESORICS) 2020*, Guildford, UK, pp. 480-501, September 2020. https://doi.org/10.1007/978-3-030-58951-6_24

[8] C. Fung, C. J. M. Yoon, and I. Beschastnikh, “Mitigating Sybils in Federated Learning Poisoning,” arXiv: 1808.04866, July 2020. https://doi.org/10.48550/arXiv.1808.04866

[9] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, “Analyzing Federated Learning through an Adversarial Lens,” in *Proceedings of the 36th International Conference on Machine Learning*, Long Beach: CA, pp. 634-643, June 2019.

[10] Y. Liu, J. Peng, J. Kang, A. M. Iliyasu, D. Niyato, and A. A. Abd El-Latif, “A Secure Federated Learning Framework for 5G Networks,” *IEEE Wireless Communications*, Vol. 27, No. 4, pp. 24-31, August 2020. https://doi.org/10.1109/MWC.01.1900525

[11] Y. Mugunthan, R. Rahman, and L. Kagal, “BlockFLow: An Accountable and Privacy-preserving Solution for Federated Learning,” arXiv:2007.03856, July 2020. https://doi.org/10.48550/arXiv.2007.03856

[12] S. K. Lo, Y. Liu, Q. Lu, C. Wang, X. Xu, H.-Y. Paik, and L. Zhu, “Towards Trustworthy AI: Blockchain-based Architecture Design for Accountability and Fairness of Federated Learning Systems,” *IEEE Internet of Things Journal*, Vol. 10, No. 4, pp. 3276-3284, January 2022. https://doi.org/10.1109/JIOT.2022.3144450

[13] Y. Li, C. Chen, N. Liu, H. Huang, Z. Zheng, and Q. Yan, “A Blockchain-based Decentralized Federated Learning Framework with Committee Consensus,” *IEEE Network*, Vol. 35, No. 1, pp. 234-241, December 2020. https://doi.org/10.1109/MNET.011.2000263

[14] S. H. Oh, Study of Reliable Federated Learning for Depending Data Poisoning Attack, Master’s Thesis, Chonnam National University, Gwangju, February 2023.

오석환(Sukhuan Ou)

2021년 : 전남대학교 (학사)
2023년 : 전남대학교 대학원 (석사)

2015년~2021년: 전남대학교 컴퓨터정보통신공학과
2021년~2023년: 전남대학교 정보보안협동과정 석사
※관심분야 : 정보보호(Personal Information), 데이터관리 등

정송헌(Songheon Jeong)

2022년 : 광주대학교 (학사)
2023년 : 전남대학교 대학원 (석사과정)

2016년~2022년: 광주대학교 사이버보안경찰학과
2022년~현 재: 전남대학교 정보보안협동과정 석사과정
※관심분야 : 정보보호(Personal Information), 클라우드, 블록체인 등

김경백(Kyungbeak Kim)

1999년 : 한국과학기술원 전기공학 및 컴퓨터공학(학사)
2001년 : 한국과학기술원 전기공학 및 컴퓨터공학(석사)
2007년 : 한국과학기술원 전기공학 및 컴퓨터공학(박사)

2007년~2008년: Network and Distributed Systems Group, Computer Science, in University of California Irvine
2008년~2012년: Information Systems Group, Computer Science, University of California Irvine
2012년~2016년: 전남대학교 전자컴퓨터공학과 조교수
2016년~2021년: 전남대학교 전자컴퓨터공학과 부교수
2021년~현 재: 전남대학교 소프트웨어공학과 인공지능학과 교수
※관심분야 : 지능형 분산시스템, SDN/NFV, 빅데이터 플랫폼, 인공지능, 블록체인, 소셜네트워크 등