# 이더리움 블록체인에서의 자금 세탁 탐지 모델에 관한 연구

황윤서, 이정호, 최형기*
성균관대학교, 소프트웨어학과

qwdfgqw@skku.edu, dlwjdgh100@skku.edu, meosery@skku.edu

# A Study on the Money Laundering Detection Model on Ethereum Blockchain

Yoon-Seo Hwang, Jeongho Lee, Hyoung-Kee Choi*
Dept. of Software, Sungkyunkwan University.

요 약

본 논문은 이더리움 블록체인에서 발생하는 자금 세탁 행위를 탐지하기 위한 AI 모델을 개발하고 평가하였다. 최근 암호화폐 거래소의 해킹 사건이 증가함에 따라, 해킹된 자금의 흐름을 추적하고 세탁 패턴을 탐지하는 것이 중요한 과제로 떠올랐다. 본 연구에서는 업비트 해킹 사건의 데이터를 기반으로 GraphSAGE 와 GCN 을 결합한 앙상블 모델을 통해 바이비트 해킹 사건과 연관된 주소들을 예측하였으며, FBI 가 발표한 51 개 주소 중 90.19%를 탐지하였다. CryptopiaHack 데이터로 일반화 성능을 검증하며 한계와 개선 가능성을 탐구하였다.

## Ⅰ. 서 론

최근 암호화폐는 블록체인 기술의 발전과 가격의 급등으로 인해 전 세계적으로 큰 관심을 받았다. 이러한 관심은 암호화폐의 거래량 증가와 다양한 활용 사례로 이어졌으나, 동시에 보안에 대한 우려도 커졌다. 특히 암호화폐 거래소의 안전성에 대한 의문이 제기되었으며, 이를 보여주듯 최근 몇 년간 해킹 사건이 빈번히 발생하였다. 예를 들어, 2019 년 한국의 업비트 거래소에서 약 4,800 만 달러 상당의 이더리움이 도난당한 사건 [1]과 2025 년 바이비트 거래소에서 약 15 억 달러 규모의 암호화폐가 도난당한 해킹 사건 [2]이 대표적이다. 이들 사건에서 해커들은 훔친 자금을 무수히 많은 암호화폐 지갑을 통해 세탁하는 과정을 거쳤으며, 이는 자금 흐름의 추적을 매우 어렵게 만들었다.

본 논문에서는 해킹된 자금의 세탁 경로를 추적하기 위해 AI 모델을 개발하고, 자금 세탁 탐지의 효과를 입증하였다. 구체적으로, 업비트 해킹 사건의 데이터를 기반으로 학습한 모델을 활용하여 바이비트 해킹 사건의 자금 흐름을 예측하고 탐지하는 데 초점을 맞추었다. 나아가 CryptopiaHack 사건 데이터 [3]를 추가로 분석해 모델의 범용성을 평가하였다.

## Ⅱ. 본론

1. 개요

기존 자금 세탁 탐지 연구 중 하나인 DenseFlow 는 암호화폐 거래 네트워크에서 거래 밀도를 기반으로 비지도 학습 접근법을 취한다 [4]. DenseFlow 는 밀집 서브그래프를 탐지하고 최대 유량 기법을 적용해 자금 흐름을 추적하며, Precision 과 MCR (Money Coverage Ratio)을 평가 지표로 사용한다. 이 방식은 밀집된 자금 세탁 네트워크를 식별하는 데 효과적이지만, 희소 패턴이나 복잡한 다중 홉 거래를 포착하는 데 한계가 있다. 또한, 레이블 데이터를 활용하지 않아 학습된 패턴의 정확도가 제한적이다.

이러한 단점을 극복하고자 본 연구는 지도 학습 기반 그래프 신경망 (GNN)을 제안한다. 구체적으로, 업비트 해킹 사건 데이터를 기반으로 GraphSAGE 와 GCN (Graph Convolutional Network)을 결합한 앙상블 모델을 개발하여 자금 세탁 주소를 분류한다. 본 접근법은 레이블 데이터를 활용해 복잡한 자금 세탁 패턴을 학습하며, 새로운 거래 그래프에 일반화할 수 있다. 평가 지표로 F1 Score, Precision, Recall 을 사용하여 분류 성능을 정밀히 분석한다. 이를 통해 DenseFlow 대비 더 정교한 탐지와 실시간적용 가능성을 제공한다.

2. 용어 및 개념

본 연구를 이해하는 데 필요한 주요 용어와 개념은 다음과 같다:

이더리움 (Ethereum): 스마트 계약을 지원하는 블록체인 플랫폼으로, 본 연구의 주요 분석 대상 암호화폐이다.

주소 (Address): 이더리움 블록체인에서 자금 송수신을 위한 고유 식별자로, 16 진수 형태로 표현된다.

거래소 (Exchange): 암호화폐를 거래할 수 있는 플랫폼으로, 본 연구에서는 업비트와 바이비트 거래소가 주요 사례이다.

3. 결과 분석

a. 데이터셋 및 피처 구성

데이터셋은 구글 BigQuery 의 bigquery-public-data.goog_blockchain_ethereum_mainnet_us 에서 추출된 이더리움 메인넷 트랜잭션 기록을 기반으로 생성되었다. 각 트랜잭션에서 from_address, to_address, value, block_timestamp, gas, gas_price 데이터를 추출하였으며, 주소를 노드로, 송금 관계를 방향성 엣지로 정의하여 그래프 데이터셋 (노드 수 1,112,552, 엣지 수 4,983,233)을 구성하였다. 자금 세탁 트랜잭션은 업비트 해킹 사건과 관련된 주소에서 발생한 송금으로 정의하였으며, 정상 트랜잭션은 무작위 샘플링으로 선별하였다. 트랜잭션 클래스 분포는 [자금 세탁 트랜잭션: 765,831, 정상 트랜잭션: 4,217,402]였다. 이 트랜잭션 데이터를 주소 단위로 집계하여 노드 레이블을 설정하였으며, 노드 클래스 분포는 [정상 노드: 871,669, 자금 세탁 관련 노드: 240,883]이었다. 클래스 불균형을 보완하기 위해 클래스 가중치를 적용하였다.

b. 모델 성능

모델의 성능은 학습과 테스트 데이터로 분할하여 F1 Score, Precision, Recall 로 평가되었다. 앙상블 모델은 GCN 과 GraphSAGE 의 출력을 가중치 (GCN: 0.6, GraphSAGE: 0.4)로 결합하여 예측하였다. 결과는 표 1 에 제시되었다.

▪ 표 1: 모델별 성능 지표 비교

| 모델 종류 | F1 Score | Precision | Recall |
|---|---|---|---|
| GCN | 0.7660 | 0.7506 | 0.7897 |
| GraphSAGE | 0.7430 | 0.7257 | 0.7969 |
| Ensemble | 0.7702 | 0.7526 | 0.8005 |

앙상블 모델은 GraphSAGE 의 일반화 능력과 GCN 의 그래프 구조 분석 능력을 융합하여 F1 Score 0.7702, Recall 0.8005 를 달성하였다.

c. 특정 주소의 탐지 결과

바이비트 해킹 사건에서 Etherscan 에 Bybit Exploiter 1 로 태그된 주소 (0x47666...)는 본 모델에서 '자금 세탁 (Detection)'으로 예측되었다. FBI 의 발표 [5]에 기반한 바이비트 해킹 사건 관련 51 개 주소를 검증 데이터로 사용한 결과, 51 개 주소가 데이터셋에 존재하였으며, 이 중 46 개를 '자금 세탁 (Detection)'으로 탐지하여 자금 세탁 탐지 비율 0.9019 (46/51)를 기록하였다.

d. CryptopiaHack 사건을 통한 일반화 성능 검증

CryptopiaHack 사건의 데이터를 통해 모델의 일반화 성능을 검증하였다. CryptopiaHack 관련 6 개 주소에 대해 모델을 적용한 결과, 3 개가 자금 세탁으로 예측되어 자금 세탁 탐지 비율은 0.5000, 즉 3/6 로 계산되었다. 이 결과는 바이비트 해킹 사건의 탐지 비율인 0.9019 에 비해 낮은 성능을 보이며, 모델이 업비트 해킹 사건에 특화된 패턴을 학습했음을 시사한다. CryptopiaHack 의 자금 세탁 패턴이 업비트나 바이비트와 다를 가능성이 있으며, 이는 모델의 일반화 성능에 한계를 드러낸다. 향후 연구에서는 다양한 자금 세탁 사례를 포함한 데이터셋을 활용해 모델의 범용성을 강화할 필요가 있다.

## Ⅲ. 결론

본 연구에서는 업비트 해킹 사건 데이터를 기반으로 학습한 GraphSAGE 와 GCN 을 결합한 앙상블 모델을 통해 바이비트 해킹 사건의 자금 세탁을 예측하였다. 새로운 피처 엔지니어링과 앙상블 가중치 최적화를 통해 트랜잭션의 통계적 특성과 자금 흐름의 방향성을 효과적으로 반영하였으며, 이를 통해 F1 Score 0.7702, Recall 0.8005 를 달성하였다. 특히, FBI 가 발표한 51 개 해킹 관련 주소 중 46 개를 탐지하여 탐지 비율 0.9019 을 기록하였다. 업비트와 바이비트 해킹 사건이 유사한 자금 세탁 패턴을 보인 점이 높은 성능의 원인으로 보인다.

그러나 CryptopiaHack 사건을 통해 모델의 일반화 성능을 검증한 결과, 6 개 주소 중 3 개만 자금 세탁으로 탐지되어 탐지 비율이 0.5000 에 그쳤다. 이는 학습 데이터가 업비트 사건에 편향되어 다른 자금 세탁 패턴에 대한 적용력이 제한적임을 보여준다. 향후 연구에서는 Tornado Cash 등의 믹싱 패턴과 주소 분산 전략을 분석하기 위해 그래프 익명화 기법과 GNN 을 결합한 모델을 개발할 계획이다. 이를 통해 프라이버시 중심 트랜잭션에서도 자금 흐름을 효과적으로 추적할 수 있을 것으로 기대된다.

## ACKNOWLEDGMENT

이 논문은 2024 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구 결과 임 (No. RS-2024-00398745, 디지털 환경에서의 증거인멸행위 증명 및 대응 기술 개발)

## 참 고 문 헌

[1] Osborne, Charlie. " Upbit cryptocurrency exchange loses $48.5 million to hackers". ZDNet, Nov, 2019.

[2] Yaffe-Bellany, David. "Big Day for Crypto Goes South After Bybit Hack". The New York Times. Feb, 2025.

[3] O'Neal, Stephen. "Cryptopia Alleged Hack: Police Are on the Case While Community Tracks Down Stolen Funds." Yahoo Finance, 18 Jan. 2019, https://finance.yahoo.com/news/cryptopia-alleged-hack-police-case-200800523.html

[4] D. Lin, J. Wu, Y. Yu, Q. Fu, Z. Zheng, and C. Yang, “DenseFlow: Spotting cryptocurrency money laundering in ethereum transaction graphs.” in Proc. Int. Conf. World Wide Web, 2024, pp. 4429– 4438.

[5] Federal Bureau of Investigation. "North Korea Responsible for $1.5 Billion Bybit Hack." IC3, 21 Feb. 2025, https://www.ic3.gov/PSA/2025/PSA250226