Journal of KIIT. Vol. 24, No. 8, pp. 147-157, Aug. 31, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.8.147

중앙화 믹서 기반 전이학습을 활용한 탈중앙화 믹서 탐지 시스템

김성빈*, 정현준**

Detection System of Decentralized Mixer using Transfer Learning from Centralized Mixers

Seongbin Kim*, Hyunjun Jung**

요 약

이 연구는 중앙화 믹서(BitcoinFog, ChipMixer)의 행동 패턴이 탈중앙화 CoinJoin 믹서(Wasabi Wallet) 탐지로 전이될 수 있는지를 검증한다. 비트코인 트랜잭션 그래프에서 10가지 행동 피처를 추출하고, GATv2를 사전학습한 후 소량의 Wasabi 라벨로 미세조정하는 전이학습 프레임워크를 설계하였다. 복합 소스 모델(Macro F1 0.914)의 zero-shot 전이는 Wasabi F1 0.013으로 큰 도메인 간극을 보였고 10% 라벨에서는 부정적 전이가 관찰된 반면, 25% 라벨에서 사전학습 모델(Wasabi F1 0.984±0.003)이 랜덤 초기화 모델(Wasabi F1 0.972±0.004) 및 100% 랜덤 초기화 모델(Wasabi F1 0.978±0.001)을 상회하여 행동 패턴 전이의 임계 데이터량 존재와 라벨 절감 효과를 함께 입증하였다.

Abstract

This study investigates whether behavioral patterns learned from centralized mixers (BitcoinFog, ChipMixer) can be transferred to detect decentralized CoinJoin mixers (Wasabi Wallet). We design a transfer learning framework that extracts 10 behavioral features from Bitcoin transaction graphs, pre-trains a GATv2 model, and fine-tunes it with a small amount of Wasabi labels. While the combined source model (Macro F1 0.914) yielded a zero-shot Wasabi-class F1 of only 0.013 and negative transfer was observed at 10% labels, at 25% labels the pre-trained model (Wasabi F1 0.984±0.003) surpassed both the from-scratch model (Wasabi F1 0.972±0.004) and the 100% from-scratch model (Wasabi F1 0.978±0.001), demonstrating effective behavioral pattern transfer above a critical data threshold with reduced label requirements.

Keywords

bitcoin, mixer detection, graph neural network, GATv2, transfer learning, wasabi mixer

* 국립군산대학교 소프트웨어학과 학사과정
- ORCID: https://orcid.org/0009-0006-6504-4153
** 국립군산대학교 소프트웨어학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-6717-1395

· Received: Apr. 09, 2026, Revised: Jul. 30, 2026, Accepted: Aug. 02, 2026

· Corresponding Author: Hyunjun Jung
Dept. of Software at Kunsan National University, 558, Daehak-ro, Kunsan-si, Jeollabuk-do, Republic of Korea
Tel.: +82-63-469-8917, Email: junghj85@kunsan.ac.kr

Ⅰ. 서 론

비트코인(Bitcoin)은 2009년 사토시 나카모토에 의해 제안된 최초의 탈중앙화 디지털 통화로, 모든 트랜잭션이 공개 블록체인에 기록되는 의사익명 시스템이다[1]. 블록체인은 주소 클러스터링과 트랜잭션 패턴 분석을 통해 사용자의 신원이 노출될 수 있다[2]. 이러한 의사익명성의 한계를 보완하기 위해 여러 사용자의 자금을 혼합하여 입출금 주소 간 연결을 끊는 믹서(Mixer) 서비스가 등장하였다[1]. 믹서는 정당한 프라이버시 보호 목적으로 사용될 수 있으나, 자금세탁 등 불법 활동에 악용되는 사례가 지속적으로 보고되고 있다[3]. 미국 FinCEN의 AML 규정 강화[4]를 시작으로, OFAC은 2022년 5월 Blender.io, 2022년 8월 Tornado Cash, 2023년 11월 Sinbad.io를 연이어 제재 지정하였으며[5], 유럽연합 또한 2024년 발효된 MiCA 규정을 통해 익명성 강화 가상자산 서비스에 대한 감독을 강화하였다. 산업계 보고에 따르면 2022년 비트코인 믹서로 유입된 자금 중 약 23%가 제재 대상 또는 불법 출처에서 비롯되었으며(2021년 12% 대비 증가), 북한 Lazarus Group 등 국가 후원 행위자의 자금 세탁 통로로 활용된 사례가 다수 보고되었다[6]. 이처럼 믹서 규모의 양적 증가와 규제 압박이 병행되며, 믹서 주소의 자동 탐지 기술은 블록체인 포렌식의 핵심 과제로 부상하고 있다.

비트코인 믹서는 운영 방식에 따라 중앙화 믹서와 탈중앙화 믹서로 구분된다. BitcoinFog[7], ChipMixer[8]등 중앙화 믹서는 서비스 운영자가 자금을 수탁하여 혼합을 수행하므로, 서버 압수나 운영자 추적을 통해 자금 흐름을 복원할 수 있는 단일 실패점이 존재하며, 실제로 2023년 BitcoinFog 운영자의 유죄 판결과 ChipMixer 인프라 압수가 이루어졌다. 반면, Wasabi Wallet이 채택한 CoinJoin 프로토콜[9]은 비트코인 트랜잭션의 다중 입출력(UTXO) 구조를 활용한 탈중앙화 방식이다. 여러 사용자가 각자의 입력과 출력 주소를 하나의 트랜잭션에 결합하면, 외부 관찰자는 어떤 입력이 어떤 출력에 대응되는지 판별하기 어렵다. Wasabi Wallet[10]은 ZeroLink 프로토콜을 기반으로 이를 구현하며, 코디네이터 서버가 트랜잭션 구성만 조율할 뿐 자금을 수탁하지 않는다. 또한 모든 참여자가 동일한 금액의 출력을 생성하도록 강제하여 금액 기반 연결 분석을 차단한다. 이처럼 CoinJoin은 서비스 운영자를 통한 추적이 불가능하고 동일 금액 출력 구조로 인해, 중앙화 믹서보다 탐지가 본질적으로 어렵다. 실제로 중앙화 믹서에 대한 법 집행 강화 이후, 사용자들이 CoinJoin 기반 탈중앙화 믹서로 이동하는 추세가 관찰되고 있다[4]. 그러나 기존 믹서 탐지 연구는 대부분 중앙화 믹서에 초점을 맞추어 왔으며[11], 구조적으로 근본적인 차이를 갖는 탈중앙화 믹서에 이러한 기법을 직접 적용하기 어렵다.

비트코인 트랜잭션 데이터는 주소를 노드로, 자금 흐름을 엣지로 표현하는 그래프 구조로 모델링되며, 그래프 신경망(GNN, Graph Neural Network)[12]-[14] 중 GATv2[15]의 동적 어텐션 메커니즘은 다양한 거래 관계가 혼재하는 비트코인 그래프에서 이웃 노드의 중요도를 적응적으로 평가할 수 있다. 선행 연구[16]에서 GATv2의 3-hop 구성이 1-hop 대비 F1 스코어 약 30%p 향상을 보인 바 있어, 이 연구는 GATv2와 3-hop BFS 확장 전략을 채택하였다. 그러나 지도학습 기반 접근법은 충분한 라벨 데이터를 전제로 하며, 탈중앙화 믹서는 프로토콜 특성상 개별 주소의 라벨링에 높은 전문성과 수작업이 요구되어 라벨 확보가 근본적으로 제한된다. 전이학습은 라벨이 풍부한 소스 도메인의 지식을 라벨이 부족한 타겟 도메인으로 전이하여 이 문제를 해결할 수 있다[17][18]. 이 연구의 핵심 가설은 중앙화 믹서와 탈중앙화 믹서가 "자금 출처 은닉"이라는 공통 목적을 공유하므로, 거래 규모의 비정상성, 자금 흐름의 방향성, 시간적 행동 패턴 등 믹서 공통의 행동적 특성이 도메인 간 전이될 수 있다는 것이다. 구체적으로, BitcoinFog와 ChipMixer의 트랜잭션 그래프에서 GATv2 모델을 사전학습한 후, 소량의 Wasabi 라벨로 미세조정하여 전이 효과를 검증한다.

이 연구의 주요 기여는 다음과 같다. 첫째, 복수의 중앙화 믹서(BitcoinFog, ChipMixer)를 결합한 복합 소스 모델이 단일 소스 대비 높은 일반화 성능(Macro F1 0.914 vs 0.897)을 달성하며, 학습된 행동 패턴이 탈중앙화 믹서로 전이될 수 있음을 실증한다. 둘째, 전체 라벨의 25%만으로 Wasabi F1 0.984

을 달성하여 100% 랜덤 초기화 모델(F1 0.978)보다 높은 성능이며, 라벨 요구량을 절감할 수 있음을 입증한다. 셋째, 10% 라벨에서의 부정적 전이(PT F1 0.900 < SC F1 0.942)와 25% 이상에서의 긍정적 전이를 통해, 도메인 간극 극복을 위한 임계 데이터 요구량의 존재를 밝힌다. 이 논문은 2장에서 배경 및 관련 연구를, 3장에서 제안 프레임워크를, 4장에서 실험 및 결과를, 5장에서 결론을 제시한다.

II. 관련 연구

2.1 비트코인 믹서 서비스

비트코인 믹서는 서론에서 기술한 바와 같이 중앙화 믹서와 탈중앙화 믹서로 분류된다. 중앙화 믹서인 BitcoinFog[7]와 ChipMixer[8]는 운영자가 자금을 수탁하여 혼합하는 방식으로, 단일 실패점이 존재한다. 반면 Wasabi Wallet[10]의 CoinJoin[9]은 ZeroLink 프로토콜 기반의 탈중앙화 방식으로, 서비스 운영자를 통한 추적이 불가능하다. Möser et al.[11]은 비트코인 믹서의 운영 구조와 약점을 체계적으로 분석하고 중앙화 믹서의 트랜잭션에 기반한 탐지 가능성을 제시하였다. Meiklejohn et al.[2]은 주소 클러스터링 휴리스틱을 통해 비트코인 네트워크의 익명성 한계를 분석하며, 믹서 서비스를 포함한 주요 서비스의 자금 흐름을 추적할 수 있음을 보였다. 그러나 이러한 연구들은 중앙화 믹서의 특성에 초점을 맞추고 있으며, CoinJoin 기반 탈중앙화 믹서에 대한 탐지 연구가 필요하다.

2.2 그래프 신경망 기반 암호화폐 주소 분류

암호화폐 트랜잭션 데이터는 주소를 노드로, 자금흐름을 엣지로 모델링하는 그래프 구조를 형성하며, 이를 활용한 주소 분류 연구가 진행되고 있다.

Weber et al.[13]은 Elliptic 데이터셋 기반으로 GCN을 활용해 약 20만 건의 트랜잭션 노드에 대한 합법/불법 이진 분류를 수행하였으며, GCN이 수작업 피처 기반 분류기와 경쟁력 있는 성능을 달성함을 보였다. Hu et al.[14]은 시간적 정보를 결합한 T-GCN을 제안하여, 트랜잭션 발생 시점과 금액 변화의 시간적 추이를 이용해 분류 성능 향상을 확인하였다. Wu et al.[12]은 이더리움 네트워크에서 트랜잭션 그래프의 구조적 정보를 노드 임베딩으로 변환하여 피싱 주소를 식별하였으며, 그래프 기반 표현 학습이 암호화폐 네트워크의 악의적 행위 탐지에 효과적임을 보였다.

GATv2[15]는 기존 GAT[19]의 정적 어텐션 한계를 극복한 모델로, 이웃 노드를 먼저 연결한 후 비선형 변환을 적용하여 동적 어텐션 할당을 가능하게 한다. 이러한 특성은 다양한 유형의 거래 관계가 혼재하는 비트코인 트랜잭션 그래프에서 이웃 노드의 중요도를 반영하는 데 유리하다.

GAT를 활용한 다중 홉 트랜잭션 그래프 기반 비트코인 믹서 주소 탐지 성능 비교 연구[16]에서는 비트코인 트랜잭션 그래프에 GAT 및 GATv2를 1-hop과 3-hop 구조로 적용하여 믹서 주소 탐지 성능을 비교하였다. 1-hop 구성에서는 F1 스코어가 0.503–0.613에 그쳤으나 3-hop 구성에서는 GATv2가 F1 0.908을 달성하여 다중 홉 이웃 정보의 활용이 믹서 탐지 성능 향상에 효과적임을 보였다.

2.3 그래프 도메인 전이학습

전이학습은 소스 도메인의 지식을 상이한 타겟 도메인으로 전이하는 기법이다. 컴퓨터 비전과 자연어 처리에서 사전학습-미세조정 패러다임이 성과를 보였으나 그래프 도메인은 비유클리드적 특성과 도메인 간 구조적 차이가 존재한다.

Hu et al.[17]은 GNN의 사전학습 전략을 체계적으로 분석하고, 노드 수준과 그래프 수준 사전학습의 결합이 효과적임을 보였다. Qiu et al.[18]은 GCC(Graph Contrastive Coding)를 통해 구조적 유사성 기반의 범용 그래프 사전학습 프레임워크를 제안하였다. Sun et al.[20]은 G-Tuning을 통해 사전학습 그래프와 다운스트림 그래프의 구조적 차이가 전이 성능에 미치는 영향을 분석하였다.

Chen et al.[21]은 미세조정 시 발생하는 파국적 망각(Catastrophic forgetting)과 부정적 전이의 상호작용을 관찰자 관점에서 분석하고, 소스 지식의 전이 불가능한 성분을 억제하는 BSS(Batch Spectral Shrinkage) 기법을 제안하였다. 이 연구는 미세조정

데이터가 증가할수록 소스의 지식이 훼손될 수 있음을 보여, 소량 라벨에서의 전이학습과 대량 라벨에서의 효과 감소 현상에 대한 근거를 제공한다.

이 연구는 다음 두 측면에서 선행 연구와 차별화된다. 첫째, 데이터셋 측면에서 기존 비트코인 믹서 탐지 연구[22]는 2017년 이전 트랜잭션 데이터를 기반으로 중앙화 믹서만을 대상으로 하였으며, GCN 기반 일반 불법 거래 분류 연구[13]는 Elliptic 데이터셋(2013-2017)의 시간적 범위로 제한된다. 이 연구는 2023년 BitcoinFog 압수 및 ChipMixer 인프라 압수 사건을 통해 공개된 최신 라벨(시드 15,165개)과 활동 중인 Wasabi Wallet 라벨(시드 8,054개)을 결합하여 단일 믹서 소스 대비 51.6% 더 많은 믹서 주소를 활용한다. 둘째, 방법론 측면에서 기존 연구는 단일 데이터셋 내 분류[12]-[14][22] 또는 동일 도메인 학습에 초점을 맞춘 반면, 본 연구는 중앙화 믹서에서 학습된 행동 패턴을 운영 방식이 근본적으로 다른 탈중앙화 CoinJoin 믹서로 전이하는 도메인 간 전이학습 프레임워크를 제시한다. 이는 신종 믹서 등장 시 라벨 확보 부담을 완화하는 실용적 함의를 가진다.

비트코인 분석에서의 전이학습 연구는 초기 단계에 있으며, 특히 복수의 중앙화 믹서에서 학습된 행동 패턴을 운영 방식이 다른 탈중앙화 CoinJoin 믹서로 전이하는 연구가 필요하다.

III. 중앙화 비트코인 믹서 사전학습 기반 전이학습을 활용한 탈중앙화 믹서 탐지 시스템

3.1 전체 프레임워크

그림 1은 이 논문에서 제안하는 프레임워크다. 이 프레임워크는 알려진 믹서에서 행동 패턴을 사전학습하고, 이를 새로운 믹서 유형으로 전이하는 두 단계로 구성된다.

소스 도메인 사전학습에서는 중앙화 믹서인 BitcoinFog와 ChipMixer의 라벨 데이터를 결합하여 복합 소스 도메인을 구성한다. 시드 주소로부터 3-hop BFS 확장을 통해 트랜잭션 그래프를 생성하고, 10가지 행동 기반 피처를 추출한 후, GNN 모델을 통해 믹서의 행동 패턴을 학습한다.

타겟 도메인 전이 및 미세조정에서는 탈중앙화 믹서인 Wasabi Mixer 주소를 시드로 하여 동일한 파이프라인(3-hop BFS 확장, 피처 추출)을 적용하고, 일반(normal) 비트코인 주소와 함께 타겟 그래프를 구성한다. 소스 모델의 사전학습 가중치를 초기값으로 사용하고, 소량의 타겟 라벨 데이터로 미세조정하여 탈중앙화 믹서에 적응시킨다.

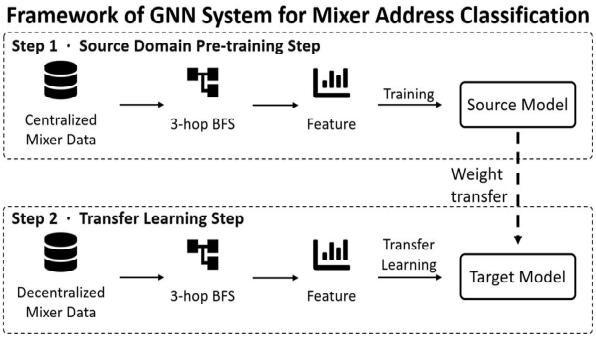

그림 1. 믹서 주소 분류를 위한 GNN 시스템 프레임워크
Fig. 1. Architecture of GNN system for mixer address framework

3.2 트랜잭션 그래프 구성

비트코인 트랜잭션 데이터를 주소 수준의 방향 그래프로 변환한다. 노드 집합은 비트코인 주소를, 엣지 집합은 트랜잭션을 통한 자금 흐름을 나타낸다. 트랜잭션에서 주소 a가 입력으로, 주소 b가 출력으로 참여한 경우, 방향 엣지를 생성한다.

시드 주소(알려진 믹서 또는 일반 주소)로부터 3-hop BFS 확장을 수행하여 주변 이웃 노드를 포함한다. 이를 통해 시드 주소뿐 아니라 거래 상대방, 간접 연결된 주소 등의 맥락 정보가 그래프에 반영된다. 확장된 그래프에서 시드 주소에만 라벨(믹서 또는 일반)이 부여되며, BFS 확장 과정에서 추가된 주소는 미라벨(Unlabeled) 상태로 유지된다.

3.3 노드 피처 설계

각 노드(주소)에 대해 트랜잭션 이력 기반의 10가지 피처를 추출한다. 이 피처들은 믹서 주소의 행동 패턴을 수치화하도록 설계되었으며, 전이학습의

기반이 되는 행동적 특성을 표현한다. 표 1은 실험에 사용된 노드 피처의 설명이다. 피처 1 - 6은 주소의 거래 규모와 방향성을, 피처 7은 트랜잭션의 구조적 특성(다수 출력)을, 피처 8 - 10은 시간적 행동 패턴을 반영한다. 이 피처들은 분산 원장 시스템의 트랜잭션 구조적 특성에 기반하여 설계되었으며 [23], 시간 및 흐름 기반의 믹서 탐지 연구[22]와 믹서 운영 구조 분석 연구[11]에서 믹서와 일반 주소를 구분하는 데 유효성이 검증된 행동 지표들이다.

표 1. 노드 피처 설명
Table 1. Node feature description

Number	Feature name	Explain
1	tx_in_count	Number of transactions as input
2	total_in	Total received amount
3	avg_in_value	Average received amount
4	tx_out_count	Number of transactions as output
5	total_out	Total sent amount
6	in_out_ratio	Input/output amount ratio
7	fanout_avg	Average number of outputs per transaction
8	tx_gap_mean	Mean time interval between transactions
9	tx_burst_ratio	Ratio of transactions with less than 60-second intervals
10	lifetime	Duration from first to last transaction

3.4 전이학습 전략

소스 도메인에서 학습된 GNN 모델을 타겟 도메인으로 전이하는 두 가지 전략을 사용한다.

Zero-shot 전이에서는 소스 모델의 가중치를 그대로 타겟 도메인에 적용한다. 소스 도메인의 피처 정규화 파라미터(평균, 표준편차)를 타겟 피처에 적용하여 입력 분포를 정렬한 후, 미세조정 없이 타겟 테스트 세트에서 평가한다.

Few-shot 미세조정에서는 소스 모델의 사전학습 가중치를 초기값으로 사용하고, 타겟 학습 데이터를 사용하여 전체 모델을 미세조정한다. 미세조정 시 학습률과 조기 종료 설정은 소스 학습과 동일하게 유지한다. 비교를 위해, 동일한 타겟 학습 데이터로 랜덤 초기화(from-scratch) 모델도 학습한다.

Ⅳ. 실험 설계 및 결과

4.1 실험 환경 및 데이터셋

실험환경은 표 2와 같다. 실험의 일관성을 확보하기 위해 동일한 환경에서 5회 반복 수행했으며, 평균 성능을 보고한다. 데이터셋은 7:3의 비율로 훈련 데이터와 테스트 데이터로 분할해 사용했다.

표 2. 실험 환경 구성
Table 2. Experimental environment configuration

Operating system	Ubuntu 22.04.1
Framework	Python 3.10.12
Library	PyTorch 2.1.0 (CUDA 11.8) PyTorch Geometric 2.6.1 scikit-learn 1.6.1
CPU	Xeon(R) Silver 4210R
GPU	NVIDIA RTX A6000×2
Memory	128 GB

표 3은 소스 도메인 데이터셋 구성 결과이다. 소스 도메인은 두 종류의 중앙화 믹서 데이터를 결합하여 구성하였다. WalletExplorer[24]의 클러스터 라벨링을 기준으로 활동이 관찰된 BitcoinFog 및 ChipMixer 주소를 시드로 사용하였으며, 일반(normal) 주소는 동일 시간 범위 내 비믹서 카테고리(거래소, 결제 서비스, 일반 지갑)의 클러스터에서 카테고리별 균등 분포를 유지하도록 30,000개 후보군에서 무작위 추출하였다. 클러스터 간 동일 주소가 중복 라벨링된 경우 우선순위(믹서 > 일반)에 따라 단일 라벨을 부여하였으며, 시드 집합 내·간 중복 주소는 그래프 구성 전에 단일 노드로 병합하였다. 시드 주소에 대해 3-hop BFS 확장을 수행하여 약 169만 개 노드로 구성된 그래프를 생성하였다. 두 믹서를 결합한 복합 소스 도메인은 단일 믹서 소스 대비 51.6% 더 많은 믹서 주소를 포함하여, 보다 일반화된 믹서 표현을 학습할 수 있다.

표 3. 소스 도메인 데이터셋 구성
Table 3. Source domain dataset configuration

Category	BitcoinFog	ChipMixer	Total
Mixer seeds	10,000	5,165	15,165
Training seeds (70%)	8,000	3,615	11,615
Test seeds (30%)	2,000	1,550	3,550

표 4는 타겟 도메인 데이터셋 구성 결과이다. 타겟 도메인은 Wasabi Wallet의 CoinJoin 주소로 구성하였다. WalletExplorer에서 동일 기간 라벨링된 Wasabi 시드 주소를 수집하고, 3-hop BFS 확장을 통해 그래프를 구성하여 피처를 계산하였다.

표 4. 타겟 도메인 데이터셋 구성
Table 4. Target domain dataset configuration

Item	Value
Wasabi seed addresses	8,054
Training seeds (70%)	5,637
Test seeds (30%)	2,417
Normal addresses	3,213
Unlabeled nodes	37,256
Total nodes	48,523
Total edges	103,759

4.2 GATv2 소스 모델

소스 모델은 GATv2[15] 아키텍처를 기반으로 하며, 기존 GAT[19]의 정적 어텐션 한계를 극복한 동적 어텐션 메커니즘을 활용한다. GATv2는 이웃 노드의 피처를 먼저 연결한 후 비선형 변환을 적용하여, 이웃 노드에 대한 동적 어텐션 스코어 할당이 가능하다.

모델은 2개의 GATv2 층으로 구성된다. 첫 번째 층은 10차원 입력 피처를 64차원 은닉 표현으로 변환하고, 두 번째 층은 64차원 은닉 표현을 유지한다. 각 층은 2개의 어텐션 헤드를 사용하며, 각 헤드가 32차원의 표현을 학습한 후 연결하여 64차원 출력을 생성한다. 각 층 사이에 Dropout(p=0.2)을 적용하고, 최종 출력층은 64차원 표현을 두 클래스로 분류하는 Linear 층과 Softmax로 구성된다. 총 학습 파라미터 수는 약 13K이다.

손실 함수로는 클래스 불균형에 대응하기 위해 Focal Loss(α=0.75, γ=2)[25]를 사용한다. 최적화에는 학습률 0.001을 사용하고, 대규모 그래프에서의 미니배치 학습을 위해 NeighborLoader(이웃 샘플 [10, 10, 10], 배치 크기 512)를 적용하여 각 층에서 최대 6개의 이웃을 샘플링한다. 조기 종료는 Patience 10, 최대 1,000 에폭으로 설정하였다.

4.3 소스 모델 학습 결과

복합 소스 도메인의 효과를 검증하기 위해 세 가지 소스 모델을 학습하고 비교하였다. 그림 2는 소스 모델 성능 비교 표이다. 그림 2를 살펴보면 BitcoinFog 모델은 Mixer F1이 높으나 Normal 분류가 약하고, ChipMixer 모델은 반대로 Normal F1은 높으나 Mixer F1이 낮아 각각 편향된 성능을 보였다. BitcoinFog + ChipMixer 복합 모델은 두 믹서의 행동 패턴을 모두 반영하여 Mixer F1 0.934, Macro F1 0.914로 가장 균형 잡힌 성능을 달성하였으며, 단일 소스(BitcoinFog) 대비 +1.7% 향상되었다. 이후 전이학습 실험에서는 BitcoinFog + ChipMixer 모델을 소스 모델로 사용한다.

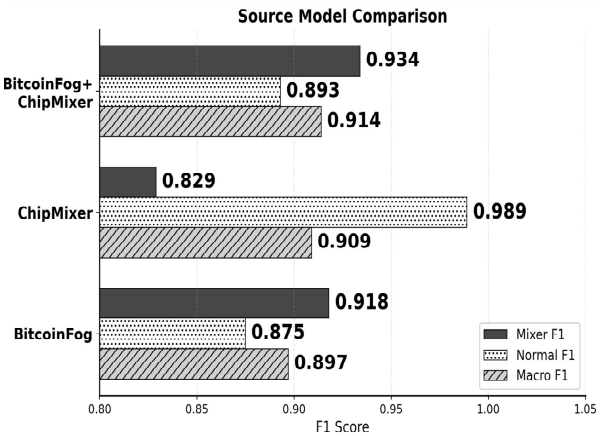

그림 2. 소스 모델 성능 비교
Fig. 2. Source model comparison

4.4 Zero-shot 전이 및 도메인 간극 분석

표 5는 비라벨 실험 결과이다. 비라벨 실험 결과를 살펴보면 Zero-shot 전이의 결과는 Wasabi F1 0.013으로 유의미한 분류 성능을 보이지 못하였다. 모델은 대부분의 테스트 주소를 Normal로 예측하여

Wasabi recall이 0.007에 그쳤다. 이는 중앙화 믹서와 탈중앙화 믹서 간 도메인 간극이 존재함을 보여준다.

표 5. 비라벨 실험 결과
Table 5. Zero-shot transfer results

Class	Precision	Recall	F1 Score
Normal	0.045	0.970	0.087
Wasabi	0.821	0.007	0.013
Macro avg	0.433	0.488	0.050

표 6의 혼동행렬은 이러한 오분류 패턴을 시각적으로 확인한다. 모델은 Wasabi 클래스 주소의 99.3%를 Normal로 오분류하였으며, 이는 사전학습된 결정 경계가 타겟 도메인에서 사실상 무효화되었음을 보여준다.

표 6. 비라벨 실험 혼동행렬
Table 6. Zero-shot transfer confusion matrix

True \ Predict	Pred normal	Pred wasabi
True normal	97.0	3.0
True wasabi	99.3	0.7

이러한 도메인 간극은 두 믹서 유형의 구조적 차이에 기인한다. 트랜잭션 구조 측면에서 중앙화 믹서는 단일 중개자를 통한 순차적 입출금 패턴을 보이는 반면, Wasabi Wallet은 다수 참여자의 동시 입출력이 하나의 트랜잭션에 결합된다. 금액 분포 측면에서 중앙화 믹서는 다양한 금액의 트랜잭션을 처리하지만, Wasabi Wallet은 고정 금액 단위로 운영된다. 시간적 패턴 측면에서 중앙화 믹서는 임의의 지연 시간을 삽입하지만, CoinJoin은 코디네이터의 라운드 주기에 따른 규칙적 패턴을 보인다. 이러한 행동 패턴의 차이가 소스 모델의 결정 경계를 타겟 도메인에서 무효화시킨 것으로 분석된다.

이러한 도메인 간극이 두 믹서 유형의 행동 패턴 차이가 아니라 데이터셋의 표면적 차이에서 비롯될 가능성을 검토할 필요가 있다. 이 연구는 소스(BitcoinFog, ChipMixer)와 타겟(Wasabi Wallet)을 모두 비트코인 UTXO 모델에서 추출하였으며, 이는 임의의 선택이 아니라 본 연구의 문제 정의상 필연적이다. 첫째, 중앙화 믹서와 탈중앙화 CoinJoin 믹서는 모두 비트코인 프로토콜 위에서 동작하므로, 동일한 트랜잭션 구조(UTXO, 입출력 주소, 수수료 모델 등)를 공유한다. 둘째, 소스를 다른 블록체인(예: 이더리움의 Tornado Cash)에서 추출할 경우 트랜잭션 모델 자체가 계정 기반과 스마트 컨트랙트로 변경되어, 검증 대상이 "믹서 행동 패턴 전이"가 아닌 "이종 블록체인 간 표현 정렬"로 바뀐다. 셋째, 일반 주소 또한 동일 비트코인 네트워크에서 샘플링되어야 트랜잭션 그래프 구조의 일관성이 보장된다. 따라서 본 실험에서 관찰된 도메인 간극은 동일 비트코인 풀 내에서 운영 방식(중앙화 중개 vs 분산 CoinJoin)의 차이로부터만 발생하며, 이는 본 연구가 검증하고자 하는 행동 패턴 전이의 본질적 도전 과제에 해당한다.

4.5 Few-shot 미세조정 결과 및 분석

그림 3인 소량 라벨 미세조정 결과를 살펴보면 10% 라벨 조건(563개 믹서 시드)에서 사전학습 모델(Wasabi F1 0.900±0.025)이 랜덤 초기화 모델(F1 0.942±0.015)보다 −4.1%p 낮은 성능을 보였다. 소스 가중치가 중앙화 믹서의 행동 패턴에 최적화되어 있어, Wasabi Wallet의 상이한 패턴을 학습하기에 563개의 시드가 불충분한 것이다. 사전학습 모델의 높은 표준편차(0.025 vs SC의 0.015)는 소스 가중치의 초기 편향이 미세조정의 수렴 경로를 불안정하게 만든다는 것을 의미한다.

25% 라벨부터 사전학습 모델의 이점이 나타난다. 표 7의 혼동행렬을 살펴보면 Wasabi F1 기준 +1.2%p, Macro F1 기준 +1.8%p의 향상이 관찰되었으며, 이는 전이학습 효과가 가장 큰 구간이다. 50%에서 +0.4%p, 100%에서 +0.1%p로 라벨 비율이 증가할수록 사전학습 이점이 감소하는 패턴은, 타겟 데이터가 충분해지면 미세조정 과정에서 소스 표현이 점진적으로 대체되기 때문이다[21]. 표 8의 랜덤 초기화 모델의 혼동행렬과 함께 비교해보면 25% 사전학습 모델은 Wasabi precision(0.987 vs 0.973)과 Normal recall(0.976 vs 0.948) 모두에서 랜덤 초기화 모델을 상회하여, 더 균형 잡힌 분류를 수행함을 확인할 수 있다.

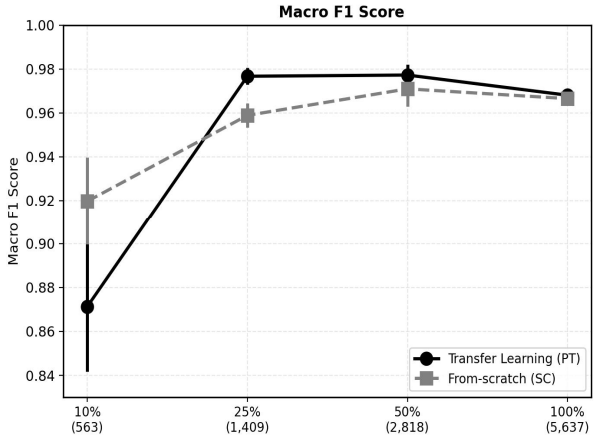

그림 3. 소량 라벨 미세조정 결과
Fig. 3. Few-shot fine-tuning results

표 7. 25% 라벨 미세조정 사전학습 모델 혼동행렬
Table 7. Pre-trained model confusion matrix (25% labels)

True \ Predict	Pred normal	Pred wasabi
True normal	97.6	2.4
True wasabi	1.9	98.1

표 8. 랜덤 초기화 모델 혼동행렬
Table 8. Confusion matrix of the scratch-trained model

True \ Predict	Pred normal	Pred wasabi
True normal	97.0	3.0
True wasabi	99.3	0.7

전이학습의 실용적 이점은 25–50% 라벨 구간에서 가장 두드러진다. 이 구간에서 사전학습 모델은 절대 성능이 F1 0.984–0.985로 이미 포화 상태에 도달하면서도, 랜덤 초기화 모델 대비 +1.2–0.4%p의 유의미한 성능 우위를 유지한다. 표 7과 표 8의 혼동행렬을 비교하면, 25% 조건에서 사전학습 모델은 Normal 클래스의 거짓 양성을 5.2%에서 2.4%로 절반 이상 감소시켜 양 클래스에 대한 결정 경계가 더 안정적임을 확인할 수 있다.

25% 사전학습 모델(Wasabi F1 0.984±0.003)이 100% 랜덤 초기화 모델(F1 0.978±0.001)을 상회하여, 전이학습이 라벨 데이터의 절감 효과를 가져올 수 있음을 보였다. 한편 10%에서의 부정적 전이와 25%에서의 긍정적 전이 간 전환은 소스-타겟 도메인 간극이 큰 전이학습에서 임계 미세조정 데이터량이 존재함을 시사한다. 본 실험에서 이 임계량은 약 563개–1,409개 믹서 시드 사이에 해당하며, 이는 신종 믹서 탐지 시스템 구축에 필요한 최소 라벨 규모의 하한을 제시한다.

Ⅴ. 결론 및 향후 과제

이 연구는 알려진 중앙화 비트코인 믹서(BitcoinFog, ChipMixer)에서 학습된 행동 패턴이 운영 방식이 근본적으로 상이한 탈중앙화 CoinJoin 믹서의 탐지에 전이될 수 있는지를 검증하였다.

실험 결과, 복합 소스 모델(Macro F1 0.914)의 zero-shot 전이는 F1 0.013으로 두 믹서 유형 간 상당한 도메인 간극을 확인하였다. Few-shot 미세조정에서는 10% 라벨에서 부정적 전이가, 25% 라벨부터 긍정적 전이가 관찰되어 임계 데이터 요구량의 존재를 밝혔다. 25% 사전학습 모델(F1 0.984)이 100% 랜덤 초기화 모델(F1 0.978)을 상회하여 라벨 데이터 절감 효과를 입증하였다.

타겟 그래프의 일반 주소가 소스 도메인과 동일한 비트코인 네트워크에서 추출되었으므로, 차수(Degree) 분포 등에서 Wasabi 주변 일반 주소와 통계적 차이가 존재할 수 있다. 그러나 이 연구가 검증하는 가설은 ‘사전학습이 행동 패턴 전이를 통해 미세조정 효율을 개선한다’는 것이며, 사전학습 모델과 랜덤 초기화 모델이 완전히 동일한 그래프, 동일한 라벨 분포, 동일한 학습 설정에서 평가되었다는 점에서 두 모델 간의 상대적 성능 차이는 그래프 구조나 데이터셋 표면적 특성에 의해 편향되지 않는다. 이 실험에서 관찰된 성능 우위는 사전학습 가중치에서만 유래한 차이이다.

향후 연구 방향은 다음과 같다. 첫째, 전이학습 메커니즘에 대한 정량적·통제 분석이 필요하다. 본 연구는 사전학습/랜덤 초기화 비교를 통해 사전학습의 전체 효과를 검증하였으나, 관찰된 성능 향상이 중앙화 믹서의 행동 패턴 학습에서 기인하는지 동일 도메인 사전학습의 일반적 초기화 효과인지를 구분하지 못하였고, zero-shot 비교에서 분석한 도메인 간극의 구체적 크기를 정량화하지 못하였다. 향후 라벨 셔플링·자기지도 사전학습 등의 통제 실험을 통해 사전학습 효과의 원인을 분리하고, 소스-타겟 도메인의 임베딩 공간 거리 측정과 피처별 분포

차이 정량화를 통해 도메인 간극의 주요 요인을 식별할 수 있다. 둘째, 도메인 적응(Domain adaptation) 기법의 적용을 통해 본 연구의 두 가지 한계를 완화할 수 있다. 본 연구의 zero-shot 전이는 Wasabi F1 0.013으로 사실상 실패하였으며, 이는 사전학습된 소스 표현이 타겟 분포로 정렬되지 않았기 때문이다. 사전학습 또는 미세조정 과정에 MMD(Maximum Mean Discrepancy)[26], CORAL (Correlation Alignment)[27] 등 분포 정렬 손실을 보조 목적함수로 통합하면, 소스-타겟 임베딩 공간을 명시적으로 정렬하여 zero-shot 성능을 개선할 수 있다. 또한 10% 라벨 조건(563개 시드)에서 관찰된 부정적 전이(−4.1%p)는 사전학습 가중치가 타겟 분포로 미세조정되기에 라벨 수가 부족할 때 발생한 현상이며, 분포 정렬 손실을 미세조정 보조 목적함수로 추가하면 본 연구가 제시한 임계 라벨 수(약 1,400개)를 더 낮출 가능성이 있다. 셋째, 피처의 행동 패턴 반영 정도에 대한 정량적 검증이 필요하다. 본 연구의 10가지 피처는 시간·흐름 기반 믹서 탐지 연구[22]와 믹서 운영 구조 분석 연구[11]에 기반하여 설계되었으나, 각 피처가 실제 믹서 행동 패턴을 얼마나 반영하는지에 대한 직접적 검증은 수행하지 않았다. 향후 SHAP, permutation importance 등을 활용한 피처 중요도 분석과 믹서/비믹서 클래스별 피처 분포 비교를 통해 피처 설계의 타당성을 정량적으로 검증할 수 있다. 넷째, 피처 설계의 확장이 가능하다. 현재 10가지 피처는 주소 수준의 통계치에 기반하며, CoinJoin 트랜잭션의 구조적 특성을 명시적으로 반영하는 피처를 추가함으로써 탐지 성능을 향상시킬 수 있다.

이 연구는 신종 믹서가 등장하여 라벨 확보가 제한된 초기 단계에서, 기존 믹서의 행동 패턴을 전이하여 신속하게 탐지 모델을 구축할 수 있음을 보여준다. 다만 전이학습의 효과가 발현되기 위해서는 최소한의 타겟 라벨이 전제되어야 하며, 이는 실무에서 신종 믹서 대응 시 라벨 확보 전략 수립에 참고가 될 수 있다.

References

[1] M. Möser, R. Böhme, and D. Breuker, "An inquiry into money laundering tools in the Bitcoin ecosystem", Proc. APWG eCrime Researchers Summit, San Francisco, CA, USA, pp. 1-14, Sep. 2013. https://doi.org/10.1109/eCRS.2013.6805780.

[2] D. Ron and A. Shamir, "Quantitative analysis of the full Bitcoin transaction graph", Proc. Financial Cryptography and Data Security (FC), Okinawa, Japan, pp. 6-24, Apr. 2013. https://doi.org/10.1007/978-3-642-39884-1_2.

[3] A. L. Reid and D. A. Orr, "Bitcoin investigations: Evolving methodologies and case studies", Journal of Forensic Research, Vol. 9, No. 3, pp. 1-4, May 2018. https://doi.org/10.4172/2157-7145.1000420.

[4] FinCEN, "Application of FinCEN's Regulations to Certain Business Models Involving Convertible Virtual Currencies", FIN-2013-G001, U.S. Department of the Treasury, May 2019.

[5] Chainalysis, "OFAC Sanctions Popular Ethereum Mixer Tornado Cash for Laundering Crypto Stolen by North Korea's Lazarus Group", https://www.chainalysis.com/blog/tornado-cash-ofac-designation-sanctions/. [accessed: May 27, 2026]

[6] Chainalysis, "2023 Crypto Crime Trends: Illicit Cryptocurrency Volumes Reach All-Time Highs Amid Surge in Sanctions Designations and Hacking", https://www.chainalysis.com/blog/2023-crypto-crime-report-introduction/. [accessed: May 27, 2026]

[7] U.S. Department of Justice, "Bitcoin Fog operator convicted of money laundering conspiracy", https://www.justice.gov/opa/pr/bitcoin-fog-operator-convicted-money-laundering-conspiracy. [accessed: May 27, 2026]

[8] Y. J. Fanusie and T. Robinson, "Bitcoin laundering: An analysis of illicit flows into digital currency services", https://www.blockchainwg.eu/wp-content/uploads/2018/01/Bitcoin_Laundering.pdf. [accessed: May 27, 2026]

[9] G. Maxwell, "CoinJoin: Bitcoin privacy for the

real world", https://bitcointalk.org/index.php?topic=279249.0. [accessed: May 27, 2026]

[10] nopara73 and TDevD, "ZeroLink: The Bitcoin fungibility framework", https://github.com/nopara73/ZeroLink. [accessed: May 27, 2026]

[11] J. Pakki, Y. Shoshitaishvili, R. Wang, T. Bao, and A. Doupé, "Everything you ever wanted to know about Bitcoin mixers (but were afraid to ask)", Proc. Financial Cryptography and Data Security (FC), Online, pp. 117-146, Mar. 2021. https://doi.org/10.1007/978-3-662-64322-8_6.

[12] J. Wu, Q. Yuan, D. Lin, W. You, W. Chen, C. Chen, and Z. Zheng, "Who are the phishers? Phishing scam detection on Ethereum via network embedding", IEEE Transactions on Systems, Man, and Cybernetics: Systems, Vol. 52, No. 2, pp. 1156-1166, Feb. 2022. https://doi.org/10.1109/TSMC.2020.3016821.

[13] M. Weber, G. Domeniconi, J. Chen, D. K. I. Weidele, C. Bellei, T. Robinson, and C. E. Leiserson, "Anti-money laundering in Bitcoin: Experimenting with graph convolutional networks for financial forensics", Proc. KDD Workshop on Anomaly Detection in Finance (ADF-KDD), Anchorage, AK, USA, pp. 1-9, Aug. 2019. https://doi.org/10.48550/arXiv.1908.02591.

[14] Y. Xie, J. Jin, J. Zhang, S. Yu, and Q. Xuan, "Temporal-amount snapshot MultiGraph for Ethereum transaction tracking", Proc. International Conf. on Blockchain and Trustworthy Systems (BlockSys), Guangzhou, China, pp. 133-146, Aug. 2021. https://doi.org/10.1007/978-981-16-7993-3_10.

[15] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?", Proc. International Conf. on Learning Representations (ICLR), Online, pp. 1-10, Apr. 2022. https://doi.org/10.48550/arXiv.2105.14491

[16] S. Kim and H. Jung, "A performance comparison of Bitcoin mixer address detection based on multi-hop transaction graphs using GAT", Proc. KIIT Conference, Jeju, Republic of Korea, pp. 536-540, Jun. 2025.

[17] W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V. Pande, and J. Leskovec, "Strategies for pre-training graph neural networks", Proc. International Conf. on Learning Representations (ICLR), Addis Ababa, Ethiopia, Apr. 2020. https://doi.org/10.48550/arXiv.1905.12265

[18] J. Qiu, Q. Chen, Y. Dong, J. Zhang, H. Yang, M. Ding, K. Wang, and J. Tang, "GCC: Graph contrastive coding for graph neural network pre-training", Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD), Online, pp. 1150-1160, Aug. 2020. https://doi.org/10.1145/3394486.3403168.

[19] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks", Proc. International Conf. on Learning Representations (ICLR), Vancouver, Canada, pp. 1-8, Apr. 2018. https://doi.org/10.48550/arXiv.1710.10903.

[20] Y. Sun, Q. Zhu, Y. Yang, C. Wang, T. Fan, J. Zhu, and L. Chen, "Fine-tuning graph neural networks by preserving graph generative patterns", Proc. AAAI Conf. on Artificial Intelligence, Vancouver, Canada, Vol. 38, No. 8, pp. 9053-9061, Feb. 2024. https://doi.org/10.1609/aaai.v38i8.28755.

[21] X. Chen, S. Wang, B. Fu, M. Long, and J. Wang, "Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning", Proc. Advances in Neural Information Processing Systems (NeurIPS), Vancouver, Canada, Vol. 32, pp. 1908-1918, Dec. 2019. https://dl.acm.org/doi/10.5555/3454287.3454458.

[22] J. Wu, J. Liu, W. Chen, H. Huang, Z. Zheng, and Y. Zhang, "Detecting mixing services via mining Bitcoin transaction network with hybrid motifs", IEEE Transactions on Systems, Man, and Cybernetics: Systems, Vol. 52, No. 4, pp.

2237-2249, Apr. 2022. https://doi.org/10.1109/TSMC.2021.3049278.

[23] A. Zamyatin, M. Al-Bassam, D. Zindros, E. Kokoris-Kogias, P. Moreno-Sanchez, A. Kiayias, and W. J. Knottenbelt, "SoK: Communication across distributed ledgers", Proc. Financial Cryptography and Data Security (FC), Online, pp. 3-36, Mar. 2021. https://doi.org/10.1007/978-3-662-64331-0_1.

[24] WalletExplorer: https://www.walletexplorer.com/. [accessed: Apr. 23, 2026]

[25] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection", Proc. IEEE Int. Conf. on Computer Vision (ICCV), Venice, Italy, pp. 2980-2988, Oct. 2017. https://doi.org/10.1109/ICCV.2017.324

[26] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test", Journal of Machine Learning Research, Vol. 13, pp. 723-773, Mar. 2012. https://jmlr.csail.mit.edu/papers/v13/gretton12a.html.

[27] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation", Proc. AAAI Conf. on Artificial Intelligence, Phoenix, AZ, USA, pp. 2058-2065, Feb. 2016. https://doi.org/10.1609/AAAI.V30I1.10306.

저자소개

김 성 빈 (Seongbin Kim)

2021년 3월 ~ 현재 : 국립군산대학교 소프트웨어학과 학사과정

관심분야 : 보안, 블록체인

정 현 준 (Hyunjun Jung)

2008년 3월 : 삼육대학교 컴퓨터과학과(공학사)

2010년 3월 : 숭실대학교 컴퓨터학과(공학석사)

2017년 9월 : 고려대학교 컴퓨터·전파통신공학과(공학박사)

2017년 8월 ~ 2020년 8월 : 광주과학기술원 블록체인인터넷경제연구센터 연구원

2021년 3월 ~ 현재 : 국립군산대학교 소프트웨어학과 교수

관심분야 : 블록체인, 데이터 사이언스, 센서 네트워크, 사물인터넷, 머신러닝