Cross-Domain Fraud Detection in Ethereum Tokens Using On-Chain Time Series Based on Neural Networks and Reservoir Computing

Lívia Carrera[1], Raquel C.G. Pinto[1], Anderson Fernandes Pereira dos Santos[1,2]

[1] Departamento de Ciência da Computação
Instituto Militar de Engenharia (IME) – Rio de Janeiro, RJ – Brasil

[2]Venturus Innovation Center – Campinas, SP – Brasil

livia.emanuelle@ime.eb.br, raquel@ime.eb.br, anderson@ime.eb.br

Abstract. *Fraud in Ethereum token launches (ICOs, IDOs, and NFTs) produces observable on-chain behavioral patterns: anomalous asset concentration, artificial participant influx, fee manipulation, and abrupt activity collapse. This paper investigates whether four time series used for ICO fraud detection act as* transferable *behavioral indicators across launch modalities. The main contribution is the cross-domain evaluation protocol itself, applied uniformly to all model families: we empirically evaluate it on 758 labeled assets (ICO, IDO, NFT), comparing Keras baselines (MLP, CNN-MLP, LSTM-MLP), classical Echo State Networks (ESN), and Quantum Reservoir Computing (QRC) under a unified fixed decision threshold ($\tau = 0.5$). Performance is assessed with metrics robust to class imbalance (PR-AUC, ROC-AUC, F1, precision, recall) and a threshold-collapse diagnostic. IDO time series were fully extracted via Etherscan API v2; NFT reached 87.6% extraction coverage (219 of 250 labeled collections). Results show strong* intra-domain *predictivity across all three domains, but zero-shot transfer fails (ICO→IDO/NFT PR-AUC* ≈ 0.51*, below the* ≈ 0.60 *random baseline expected at observed prevalence): behavioral invariance was not observed across the domains and fraud definitions considered in this study. ESN-fusion-matched achieved the best ICO discriminative ranking (PR-AUC* $= 0.808$*), with no significant difference between classical and quantum reservoirs under the paired statistical protocol.*

1. Introduction

Ethereum smart contracts [Buterin 2014, Wood 2014] underpin several asset-launch mechanisms, Initial Coin Offerings (ICOs), Initial Decentralized-Exchange Offerings (IDOs), and Non-Fungible Token (NFT) collections, all of which produce publicly available, temporally ordered on-chain records. This transparency, however, does not preclude fraud: schemes exploit information asymmetry and artificial demand coordination, often amplified by social networks where narratives of rapid appreciation lead investors to mistake coordinated activity for organic adoption [Federal Trade Commission 2026].

Execution varies across modalities, yet the underlying mechanics converge. In ICOs, the U.S. Securities and Exchange Commission (SEC) charged Centra Tech's founders with fabricating partnerships [U.S. Securities and Exchange Commission 2018]. In NFTs, the U.S. Department

of Justice (DOJ) prosecuted the creators of *Frosties* for diverting $1.1 million in pre-sale funds and abandoning the project within hours [U.S. Department of Justice 2022]. In IDOs, rug pulls abruptly drain pool liquidity, stranding remaining participants [Xia et al. 2021, Cernera et al. 2023]. Across all three settings, recurring on-chain signatures emerge: artificial influxes of addresses, anomalous asset concentration, abrupt volume shifts, and atypical fee patterns.[da Mata Caffé et al. 2023] proposed a fraud-detection approach for ICOs based on normalized time series extracted from Ethereum transactions. Their Multilayer Perceptron (MLP), Convolutional–MLP (CNN-MLP), and Long Short-Term Memory–MLP (LSTM-MLP) architectures achieved up to 91 % Recall over 20-day windows, relying on four behavioral series, NEWHOLDER, NEWUSER, BIGBUYER, and GAS/GASLIMIT, as temporal representations of token activity. Two questions remain open: (i) are these series ICO-specific, or do they capture patterns transferable to other launch modalities? (ii) how do they behave across model families under a uniform evaluation protocol, particularly when the analysis extends to IDOs and NFTs?

This study re-evaluates the four series of [da Mata Caffé et al. 2023] as *candidate* cross-domain fraud indicators, without assuming their generalizability a priori. The contributions relative to [da Mata Caffé et al. 2023] are:

- **Cross-domain benchmark and protocol** (main contribution): we define a uniform labeling protocol spanning three launch modalities, with on-chain Etherscan extraction for IDO (250/250) and NFT (219/250), and a threshold-aware evaluation framework using a fixed threshold $\tau = 0.5$, Precision–Recall Area Under the Curve (PR-AUC) as primary metric, and a collapse diagnostic.
- **Behavioral reformulation**: we reinterpret each series as a candidate fraud indicator linked to an observable economic mechanism common to ICOs, IDOs, and NFTs.
- **Zero-shot transfer analysis**: we test whether models trained on ICO data transfer to IDO and NFT domains and characterize the failure modes encountered.
- **Controlled reservoir comparison**: Echo State Networks (ESNs) and Quantum Reservoir Computing (QRC) models are evaluated under matched feature dimensionality and identical readouts across all three domains, including behavior under realistic (5–10 %) fraud prevalence.

The remainder of this paper is organized as follows. Section 2 surveys related work on on-chain fraud detection and Reservoir Computing. Section 3 formalizes the problem, describes data collection and labeling, and details the feature-extraction pipeline. Section 4 presents results across the ICO, IDO, and NFT domains, including cross-domain transfer and prevalence analysis. Section 5 concludes and identifies future directions.

2. Related Works

2.1. On-Chain Fraud Detection in Ethereum

On-chain fraud detection in Ethereum has been investigated across different scopes, including Ponzi schemes in smart contracts, market manipulation, and, more recently, rug pulls in NFTs and DeFi. Taken together, this existing literature supports the central premise of the present work: fraudulent behaviors leave observable on-chain temporal

traces that are structurally similar across different launch mechanisms. Table 1 summarizes the most relevant related works (domains, methods, features, strengths, limitations); its final row contrasts them with the present cross-domain approach.

In the Ponzi domain, Chen et al. [Chen et al. 2018] detected schemes from transaction features and opcodes (200 verified contracts), later scaling to 3,780+ contracts with Random Forest (F1 $= 0.79$) [Chen et al. 2019], evidence that Ether-flow patterns discriminate even without source code. Xu and Livshits [Xu and Livshits 2019] showed that demand coordination, volume spikes, and asset concentration precede pump-and-dump collapses.

In the NFT domain, [Huang et al. 2023] presented the first systematic study of NFT rug pulls on Ethereum (173,000+ collections; 7,487 fraudulent), describing four recurring symptoms (mint-fee profit extraction, irreversible price collapse, abrupt on-chain decline, and social-media abandonment) and showing that a time-series model over on-chain events can alert up to 96 hours before a rug pull.

Large-scale scam-token studies. Xia et al. [Xia et al. 2021] conducted the first large-scale study of scam tokens on Uniswap, combining a guilt-by-association heuristic with a machine-learning classifier over all Uniswap V2 transactions to identify over 10,000 scam tokens (roughly half of all listings) with at least $16 million in losses across nearly 40,000 victims; Cernera et al. [Cernera et al. 2023] characterized rug pulls and sniper bots across DeFi, and Huang et al. [Huang et al. 2023] covered 173,000+ NFT collections. These efforts differ from ours in *scale* (10^4–10^5 tokens vs. 758 curated assets balanced over three modalities), *data* (a single venue or asset type vs. ICO/IDO/NFT corpora under one temporal-series representation), *labels* (venue-specific heuristics vs. verifiable per-asset labels with three distinct fraud definitions, Section 3.3), and *objective* (ecosystem measurement and single-domain early warning vs. a controlled test of cross-domain generalization). The smaller, balanced design is thus a deliberate choice dictated by the cross-domain question, not a coverage limitation.

2.2. Distribution Shift and Domain Adaptation for Time Series

Our central question, whether a detector trained in one domain transfers to another, is best read through the lens of *dataset shift*. Quiñonero-Candela et al. [Quiñonero-Candela et al. 2009] and Moreno-Torres et al. [Moreno-Torres et al. 2012] formalize its regimes: *covariate shift* (input distribution changes), *prior-probability shift* (class balance changes), and *concept shift* (the labeling function itself changes). Ben-David et al. [Ben-David et al. 2010] bound target error by source error plus a domain-divergence term, making explicit that zero-shot transfer can be arbitrarily poor under large divergence; practical remedies learn domain-invariant representations via domain-adversarial training [Ganin et al. 2016] and related unsupervised adaptation methods for time series [Wilson and Cook 2020]. Our three corpora differ simultaneously along all three axes — concept (distinct fraud definitions, Section 3.3), covariate (heterogeneous pipelines, Section 3.3.3), and prior-probability (different prevalences) — so standard theory predicts that naive zero-shot transfer should be hard; the adaptation techniques above are the natural recovery route, outlined as future work.

[da Mata Caffé et al. 2023] constitute the direct starting point of this work. The

Table 1. Summary of related work on on-chain fraud detection in Ethereum.

Work	Domain	Method	Features	Best	Limitation
[Chen et al. 2018]	Ponzi	ML	Opcodes, tx	High prec.	Small sample (n=200)
[Chen et al. 2019]	Ponzi	Rand. Forest	Bytecode, tx hist.	F1 = 0.79	No temporal modeling
[Xu and Livshits 2019]	P&D	Statistical	Volume, price	Qualitative	No classifier
[Huang et al. 2023]	NFT rug pull	TS + ML	Mint, swap, burn	Alert 96h	NFT-only
[Xia et al. 2021]	DeFi rug pull	Graph + ML	Token graph	High prec.	No time series
[da Mata Caffé et al. 2023]	ICO	MLP/CNN/LSTM	4 on-chain series	Rec. = 0.91	ICO-only
This work	**ICO,IDO,NFT**	**ESN, QRC**	**4 on-chain series**	**PR-AUC = 0.808**	**Zero-shot transfer not observed**

authors constructed normalized time series from Ethereum transaction tables in the ICO domain and evaluated MLP, CNN-MLP, and LSTM-MLP models, reporting Recall of up to 91% over 20-day windows. The contribution of the present work is not to replace these series, but to reinterpret the four proposed representations as candidate behavioral fraud indicators, examining whether they retain discriminative power in IDOs and NFTs under a single evaluation protocol applied uniformly to all model families. The two works contrast along every axis: *domains* (ICO only vs. ICO/IDO/NFT), *data* (258 pre-processed tokens vs. +500 newly labeled on-chain assets), *labels* (one fraud notion vs. three distinct definitions, Section 3.3), *models* (Keras baselines vs. those plus matched classical/quantum reservoirs), *protocol* (Recall-centric vs. PR-AUC-primary, fixed-threshold, collapse-diagnosed), and *analyses* (intra-domain only vs. transfer, pooling, and prevalence stress tests).

2.3. Classical and Quantum Reservoir Computing

Reservoir Computing (RC) drives a fixed high-dimensional dynamical system, the reservoir, with the input and trains only the output layer [Jaeger 2001, Lukoševičius and Jaeger 2009]. The recurrent reservoir captures sequential dependencies without explicit temporal feature engineering, and the fixed-dynamics/trained-readout split reduces cost and overfitting risk — constraints inherent to fraud detection in emerging token launches with scarce labels.

Echo State Networks (ESN), based on recurrent reservoirs with fixed random weights, are the most studied classical instantiation. By maintaining the same architectural principles across ESN and quantum implementations, we establish an isomorphic baseline that isolates the effect of the temporal representation independent of learning strategy.

Quantum Reservoir Computing (QRC) extends the paradigm to quantum systems: classical series are encoded into qubit rotations whose fixed unitary evolution expands the representation space with the number of qubits [Fujii and Nakajima 2017, Nakajima et al. 2019], and the readout remains classical and linear. All QRC circuits in this study are *classically simulated* at small scale (6–8 qubits); no quantum hardware is used and no quantum advantage is claimed. Because the reservoir is fixed and untrained, only the readout adapts per domain, the motivation for testing QRC in zero-shot transfer.

The QRC implementation utilizes the QRC-Lab framework [dos Santos 2026], a modular gate-based toolbox for Quantum Reservoir Computing. QRC-Lab provides,

by design: angle encoder, reservoir with randomly controlled rotation gates, temporal evolution simulator, observables module (local Pauli-Z and two-qubit correlations), and ridge regression readout.

3. Problem Formulation and Methodology

3.1. Overview

The method has six stages: (i) behavioral hypotheses; (ii) collection and labeling across three domains; (iii) normalized time series; (iv) feature extraction via classical/quantum reservoirs with linear readout; (v) threshold-aware evaluation robust to imbalance; and (vi) cross-domain transfer testing. Unlike [da Mata Caffé et al. 2023], every stage is defined uniformly for ICOs, IDOs, and NFTs, enabling the invariance question to be tested. Figure 1 shows the pipeline.

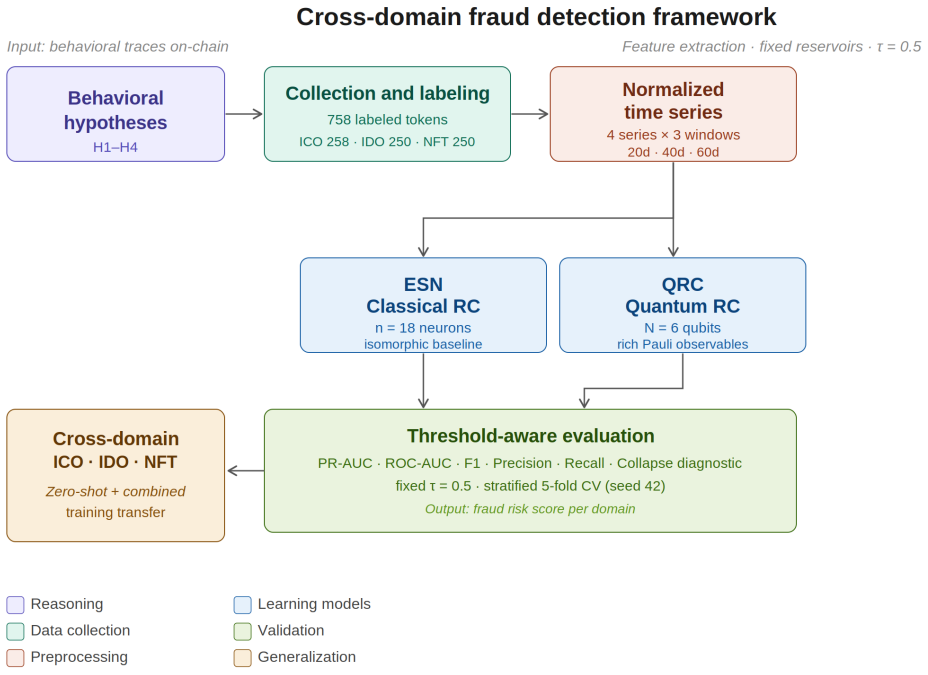

Figure 1. Complete framework pipeline for cross-domain fraud detection.

3.2. Research Hypotheses

The following four hypotheses are derived from the behavioral series originally proposed by [da Mata Caffé et al. 2023] for ICO fraud detection. We reinterpret each series as a candidate fraud invariant grounded in an observable economic mechanism, and investigate whether it retains discriminative power across ICO, IDO, and NFT domains.

H1 – NEWHOLDER [da Mata Caffé et al. 2023]. Fraud tends to show artificial, concentrated initial growth in unique holders followed by abrupt stagnation after liquidity extraction; legitimate projects grow organically. The series tracks cumulative unique-holder growth.

H2 – NEWUSER [da Mata Caffé et al. 2023]. Fraud frequently uses newly created accounts to simulate adoption; the series measures the daily share of addresses whose first on-chain record is an interaction with the token — persistently high values suggest artificial participants.

H3 – BIGBUYER [da Mata Caffé et al. 2023]. Supply concentration in one or few coordinated wallets characterizes pump-and-dump and rug pulls; the series tracks the daily fraction of supply held by the largest address.

H4 – GAS/GASLIMIT [da Mata Caffé et al. 2023]. Complex-contract calls, bots, and scripts exhibit gas-utilization ratios near one, while simple organic transfers do not; the series is the daily ratio of gas consumed to gas limit declared.

3.3. Data Collection and Labeling

Figure 2 summarizes the corpora: (a) 758 labeled tokens — ICO 258 (34%), IDO 250 (33%), NFT 250 (33%); (b) fraud prevalence ≈59/60/57% per domain; (c) extraction coverage of 245/258 ICO (95.0%), 250/250 IDO (100%), and 219/250 NFT (87.6%); (d) uniform ICO per-series coverage (245 tokens) at the 40-day window.

Heterogeneous fraud definitions across domains. The three corpora do not share one operational definition of "fraud": ICO labels derive from SEC-style enforcement and curated lists, IDO labels denote *rug pulls* (malicious liquidity removal), and NFT labels denote *cybersquatting* (impersonation) — distinct phenomena with different mechanics, victims, and on-chain footprints. Every cross-domain experiment thus tests generalization across launch modalities *and* across labeling functions: low transfer may stem from semantic (concept) shift of the labels rather than, or in addition to, absent shared behavioral signal. All cross-domain conclusions are relative to these specific definitions.

Corpus prevalence vs. real-world prevalence. All three corpora are approximately balanced (∼57–60% fraud) by construction, which is convenient for benchmarking but does not represent deployment conditions, where fraud prevalence in token markets is estimated at 5–10%. Results on the main benchmark therefore characterize discriminative signal under near-balanced conditions; operational claims are restricted to the dedicated prevalence analysis of Section 4.5.

3.3.1. ICO Domain

We use directly the dataset published in [da Mata Caffé et al. 2023], obtained from the public repository *ICOFraudDetection*. The dataset contains 258 token addresses with binary labels. Pre-computed pickles yield normalized series for 20/40/60-day windows without re-extraction. After filtering incomplete series, **245 tokens** enter experiments at the 40-day fusion window (95.0% of labeled ICOs), with fraud prevalence of **58.0%** (Panel (b)). The corpus is nearly balanced (∼58% fraud), unlike real markets (∼5–10%), motivating the prevalence-sensitivity analysis in Section 4.5.

3.3.2. IDO and NFT Domains

Labels are sourced from verifiable primary sources and versioned in the repository as `data/ido_labels.csv` and `data/nft_labels.csv`.

IDO fraudulent labels are rug-pull/malicious-liquidity-removal contracts from Dianxiang Sun's *rugpull_dataset* (Ethereum mainnet); IDO legitimate labels are high-persistent-liquidity Uniswap v2/v3 tokens from CoinGecko. NFT fraudulent labels come from the *Cybersquatting-NFT* dataset (valid `0x` contracts); NFT legitimate labels are active Ethereum collections from CoinGecko's NFT list.

Feature extraction. Daily series are built via Etherscan API v2 using `fetch_ido.py` (ERC-20 `tokentx`) and `fetch_nft.py` (ERC-721 `tokennfttx`), producing 60-day CSVs with four columns per token. **IDO:** all 250 labeled contracts were extracted and evaluated (100% coverage). **NFT:** 219 of 250 labeled collections (87.6%) were successfully extracted; 31 collections lack recoverable on-chain activity in the observation window and are excluded (Panel (c)). All extracted series are min–max normalized per token on load.

3.3.3. Pipeline Heterogeneity Between ICO and IDO/NFT

The ICO and IDO/NFT corpora come from *different* pipelines: ICO series are consumed exactly as published by [da Mata Caffé et al. 2023] (pre-computed, pre-normalized pickles whose low-level extraction choices are inherited), whereas IDO/NFT series were reconstructed here from raw Etherscan API v2 responses and min–max normalized *independently per token and per domain*. Three consequences follow: *(i) covariate shift* — pipeline differences can induce systematic feature differences unrelated to fraud; *(ii) normalization misalignment* (per-domain min–max scaling maps identical raw behaviors to different normalized shapes, penalizing zero-shot transfer; and *(iii) anchor mismatch*) t_0 differs across domains (Table 2), so day-index t is not semantically identical. These pipeline factors compound the label-semantics shift of Section 3.3 and are revisited when interpreting the zero-shot results (Section 4.4). A uniform re-extraction of all three domains through a single pipeline would isolate behavioral shift from pipeline shift, and is left as future work.

3.4. Time Series Construction

Each token is represented as a set of normalized daily time series. Let $D \in \{\text{ICO}, \text{IDO}, \text{NFT}\}$ be a domain, c a token or collection, and t_0 the launch event. For each window $w \in \{20, 40, 60\}$ days, we observe transactions in $[t_0, t_0 + w)$ and define Equations 1, 2, 3, and 4.

$$\text{NEWHOLDER}(t) = \frac{|\text{unique holders up to } t|}{|\text{total unique holders in window}|} \tag{1}$$

$$\text{NEWUSER}(t) = \frac{|\text{new network addresses on day } t|}{|\text{transactions on day } t|} \tag{2}$$

$$\text{BIGBUYER}(t) = \frac{\max_a \text{balance}(a, t)}{\text{total supply on day } t} \tag{3}$$

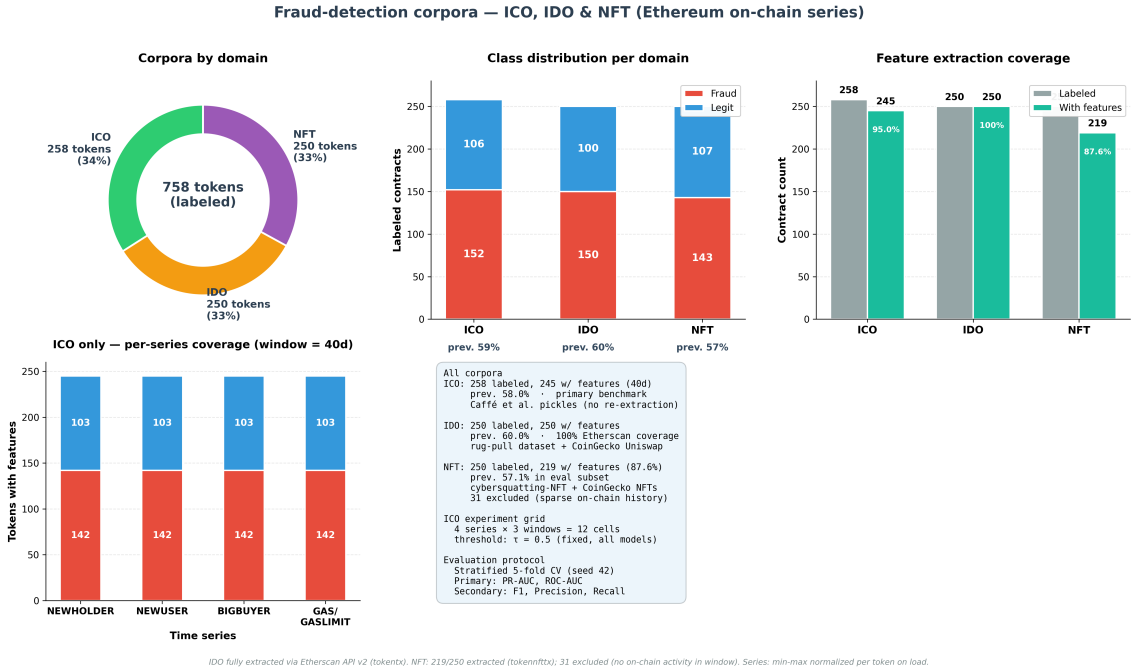

Figure 2. Fraud-detection corpora across ICO, IDO, and NFT domains on Ethereum, 758 labeled assets total. (a) Domain distribution by token count. (b) Class distribution (fraud vs. legitimate) per domain, with fraud prevalence annotated below each bar. (c) Feature extraction coverage: labeled contracts vs. contracts with recovered on-chain series. (d) ICO per-series token coverage at the 40-day window across the four behavioral time series (NEWHOLDER, NEWUSER, BIGBUYER, GAS/GASLIMIT).

$$\text{GAS/GASLIMIT}(t) = \frac{\sum_{\tau \in \mathcal{T}_t} \text{gasUsed}(\tau)}{\sum_{\tau \in \mathcal{T}_t} \text{gasLimit}(\tau)} \tag{4}$$

Table 2 defines domain-specific anchors (t_0) and data sources for each series.

Table 2. Cross-domain time series adaptation.

Series	ICO	IDO	NFT
Day 0	First transaction or token launch	Liquidity pair creation on DEX	First mint transaction
NEWHOLDER	Cumulative unique holders	Holders after pool swaps	Unique wallets with collection NFTs
NEWUSER	New addresses among buyers	New addresses among swappers	New addresses among minters/buyers
BIGBUYER	Max balance / total supply	Max balance / reserve	Max NFTs per wallet / minted supply
GAS/GASLIMIT	Daily ratio in contract txs	Daily ratio in pool swaps	Daily ratio in mint / marketplace txs

3.5. Feature Extraction Pipeline

For each token c and window w, the pipeline follows four stages.

Stage 1: Normalized time series. The series $\mathbf{u}^{(c)} = (u_1, \ldots, u_w) \in [0,1]^w$ is computed according to Equations 1–4. For the *fusion* variant, all four series are processed

simultaneously as a multichannel signal $\mathbf{U}^{(c)} \in [0,1]^{w\times 4}$.

Stage 2: Reservoir drive. The series drives the reservoir in continuous temporal mode, the state at each step accumulating memory of the dynamics.

Stage 3: Feature vector. The final $k = 5$ state vectors are concatenated. For QRC with $N = 6$ qubits and *rich* observables $(\langle Z_i\rangle, \langle X_i\rangle, \langle Y_i\rangle)$, dimension is $k \times 3N = 90$. For ESN-matched with $n = 18$ neurons, dimension is $k \times n = 90$, making both models isomorphic: 91 trainable parameters each (90 weights + 1 bias).

Stage 4: Linear readout. A logistic regression classifier with L_2 regularization ($C = 0.01$) is trained on the feature vectors. **All models use a fixed decision threshold of** $\tau = 0.5$ on validation scores (`tune_objective=fixed_05`), matching Keras baselines under a single, uniform decision rule.

4. Results and Analysis

4.1. Experimental Protocol

4.1.1. Models

We compare: **MLP, CNN-MLP, LSTM-MLP** (Keras, ICO only); **ESN-matched** ($n = 18$) and **ESN-large** ($n = 500$); **ESN-fusion** variants (multichannel, input_dim $= 4$); **QRC** ($N = 6$ qubits) and **QRC-fusion**. All reservoir models use a classical linear readout; the comparison targets the temporal representation, not the readout family.

4.1.2. Metrics and threshold policy

All ESN and QRC runs use a **fixed decision threshold** $\tau = 0.5$, identical to the Keras baseline policy. **Primary** (threshold-independent): PR-AUC and ROC-AUC. **Secondary**: Recall, Precision, F1, and Balanced Accuracy. **Collapse diagnostic**: a fold is flagged when simultaneously (i) Recall ≥ 0.95, (ii) Precision $\leq$ prevalence $+0.05$, and (iii) predicted positive rate ≥ 0.85. Collapse score $= 5\times$ mean collapse fraction across folds. Because $\tau = 0.5$ is itself a protocol choice, the per-fold score files released in the artifact allow Precision–Recall curves and threshold sweeps ($\tau \in [0.3, 0.7]$) to be reproduced for any model; threshold sensitivity is revisited in Section 4.5.

4.1.3. Data and cross-validation

Stratified 5-fold CV (seed 42). Table 3 reports corpus coverage.

4.1.4. Statistical testing

Model families are compared with a paired two-sided **Wilcoxon signed-rank test** on per-fold PR-AUC differences over the same 5 CV partitions (H_0: median paired difference is zero; H_1: non-zero; $\alpha = 0.05$), with **Holm–Bonferroni** correction over the family of pair/window tests and matched-pairs **rank-biserial correlation** as effect size. With

only 5 folds the power is limited, so non-significance is reported as "no significant difference", never as equivalence or superiority; significance is claimed only after correction. The test is computed by `compute_dispersion.py` from the per-fold metrics (`results/metrics_per_fold.csv`). Applied to this study, none of the six comparisons reaches significance after Holm correction. For the fusion pair (ESN-fusion-matched vs. QRC-fusion; n=5 paired folds): W=2.0/4.0/1.0, p_{Holm}=0.75/0.88/0.63 at 20/40/60d, with rank-biserial $+0.73/+0.47/+0.87$. For the per-series pair (ESN-matched vs. QRC; n=20 series×fold pairs): W=69.0/91.0/54.0, p_{Holm}=0.75/0.88/0.35, rank-biserial $+0.34/+0.13/+0.49$. We therefore report **no significant difference** between classical and quantum reservoirs under this protocol.

Table 3. Corpus coverage and experimental subsets (fusion 40d evaluation unit). Prevalence = fraud fraction among tokens *with features*.

Domain	Labeled	With features	Coverage	Prevalence	Role
ICO	258	245	95.0%	58.0%	Primary benchmark
IDO	250	250	100.0%	60.0%	Cross-domain benchmark
NFT	250	219	87.6%	57.1%	Cross-domain (partial)

4.2. ICO Domain: Overall Performance

Table 4 reports all models across windows (means over 5 folds, $\tau = 0.5$). With a unified threshold policy, the collapse phenomenon largely disappears: QRC-fusion collapse scores fall to 0.0 at all windows, and Recall values align with Precision across all model families.

The strongest model on primary metrics is **ESN-fusion-matched**: PR-AUC $=$ 0.808 at 40d, ROC-AUC $= 0.768$, F1 $= 0.751$, collapse score $= 0.0$. ESN-fusion-large achieves near-identical performance (0.806). QRC-fusion reaches PR-AUC $= 0.790$ at 40d — competitive with the classical multichannel variant, with the difference between the two assessed statistically in Section 4.1.4 rather than asserted from the point estimates alone.

QRC does not show a significant advantage over ESN on primary metrics. QRC mean PR-AUC across series and windows ($\approx$ 0.69–0.72) lies below both ESN-fusion variants (0.75–0.83). The paired Wilcoxon test of Section 4.1.4 finds no significant difference at any window (all Holm-corrected $p \geq 0.35$), so the quantum reservoir shows no significant ranking advantage in this setting — nor do we claim superiority of ESN or equivalence of the two families: with 5 folds the test has limited power, and the observed gap is reported as a descriptive difference.

Fusion helps classical reservoirs. ESN-fusion-matched gains $\approx +0.07$ PR-AUC over per-series ESN-matched at 40d; QRC-fusion does not replicate this (PR-AUC 0.58–0.79 at 40–60d), suggesting no benefit from the multichannel signal in this regime.

Table 4. ICO results: mean ± standard deviation over 5 folds, $\tau = 0.5$. Per-series rows aggregate the four behavioral series per fold; fusion rows are multichannel runs. Collapse score $= 5\times$ collapse fraction. Keras baseline rows report fold means; their per-fold dispersion is provided in the artifact.

Model	PR-AUC			Recall			F1			Coll.
	20d	40d	60d	20d	40d	60d	20d	40d	60d	40d
MLP	0.74	0.76	0.75	0.73	0.71	0.72	0.69	0.68	0.69	0.0
CNN-MLP	0.76	0.77	0.75	0.73	0.73	0.71	0.70	0.70	0.69	0.0
LSTM-MLP	0.72	0.70	0.76	0.66	0.63	0.70	0.64	0.63	0.68	0.0
ESN-matched	.71±.03	.73±.03	.74±.04	.87±.02	.80±.03	.83±.02	.72±.02	.69±.02	.72±.02	0.5
ESN-large	.70±.05	.75±.03	.75±.06	.71±.03	.74±.03	.74±.03	.66±.02	.70±.02	.71±.03	0.0
ESN-fusion-matched	.75±.05	**.81±.06**	.81±.08	.77±.10	.76±.10	.73±.11	.71±.05	**.75±.05**	.71±.07	0.0
ESN-fusion-large	.76±.08	.81±.05	**.83±.06**	.67±.11	.71±.08	.72±.10	.68±.06	.73±.04	**.74±.07**	0.0
QRC	.73±.05	.72±.03	.71±.04	.72±.03	.71±.02	.72±.05	.68±.03	.69±.02	.69±.02	0.0
QRC-fusion	.81±.03	.79±.07	.75±.08	.73±.12	.75±.08	.73±.07	.72±.08	.74±.04	.70±.06	0.0

4.3. Hypothesis Verdicts

Table 5 reports the best PR-AUC per hypothesis at 40d. All four series are **supported intra-domain** (PR-AUC substantially above prevalence in all three corpora). Cross-domain invariance, however, **was not observed** across the domains and fraud definitions considered in this study, as the zero-shot experiments of Section 4.4 show.

A note of caution applies to IDO and NFT results: PR-AUC values near 0.94–0.99 may partly reflect *trivial separability* between the fraudulent and legitimate label sources — simple covariates such as collection age, activity level, series length, or transaction counts could separate the two sources by construction — rather than the behavioral mechanisms of H1–H4. We therefore treat these figures as a *hypothesized* corpus upper bound rather than evidence of deployable discriminability; a control analysis with a simple-covariate baseline (number of transactions, active days, contract age) and activity-matched samples is the required validation, and is left as an explicit open task in the artifact.

Table 5. Hypothesis verdicts (window = 40d). Lift ≈ PR-AUC − domain prevalence. "Transfer?" refers to invariance across the domains and fraud definitions considered in this study. PR-AUC cells report mean ± standard deviation over 5 folds.

Hyp.	Mechanism	ICO		IDO	NFT	Transfer?
		PR-AUC	Model	PR-AUC	PR-AUC	
H1	NEWHOLDER	.75±.08	QRC	.95±.02	.98±.01	**Not obs.**
H2	NEWUSER	.78±.03	ESN-l.	.94±.01	.99±.01	**Not obs.**
H3	BIGBUYER	.76±.05	ESN-m.	.94±.02	.99±.00	**Not obs.**
H4	GAS/GASLIMIT	.78±.06	ESN-l.	.95±.03	.99±.01	**Not obs.**

All hypotheses supported intra-domain. Zero-shot transfer not observed (see Section 4.4).

4.4. Cross-Domain Transfer

4.4.1. Zero-shot (train ICO → test IDO/NFT, ESN-fusion 40d)

- **ICO → IDO**: PR-AUC $= 0.513$, Recall $= 0.140$.

- **ICO → NFT**: PR-AUC $= 0.515$, Recall $= 0.040$.

Since the expected PR-AUC of a random scorer is approximately the positive-class prevalence, the observed ≈ 0.51 lies **below** the ≈ 0.60 random baseline, confirming zero-shot failure rather than near-random performance. Consistent with Section 2.2, three factors plausibly drive it: (i) *prior-probability and covariate shift* from the differing temporal anchor t_0; (ii) *concept shift* — SEC enforcement (ICO), rug pulls (IDO), and cybersquatting (NFT) are distinct phenomena (Section 3.3), so the model must recognize a different concept at test time; and (iii) heterogeneous pipelines with per-domain normalization (Section 3.3.3). The design cannot disentangle these factors.

4.4.2. Combined training (pool ICO+IDO+NFT)

Per-domain test PR-AUC: ICO 0.740, IDO 0.892, NFT 0.981. Pooling helps IDO/NFT but *degrades* ICO (-0.068 vs. ICO-only). This asymmetry confirms domain shift: the pooled model learns predominantly from IDO/NFT patterns, which are more separable, at the cost of ICO discriminability.

4.5. Realistic Prevalence Analysis

In real-world deployment, fraud prevalence in token markets is estimated at 5–10%, far below the $\sim 59\%$ of the benchmark datasets. Table 6 reports PR-AUC and Recall under stratified subsampling to 5%, 7.5%, and 10% fraud prevalence with $\tau = 0.5$. The resulting samples are small ($n \approx 77$–114 per configuration) and the reported figures come from a single stratified draw; they should be read as illustrative of the qualitative behavior under rare-event conditions, not as stable estimates — multi-seed replication of this analysis is left to the artifact. We adopt subsampling, rather than class weights or probability calibration, because it changes the *test-time* prevalence — the quantity that differs between benchmark and deployment — while leaving the training procedure of every model family untouched, preserving the controlled comparison; class weighting and calibration alter the fitted models themselves and are listed as future work.

Table 6. PR-AUC / Recall under realistic fraud prevalence ($\tau = 0.5$, fusion 40d). Single stratified draw; small n (≈ 77–114).

Domain	Model	5%	7.5%	10%
ICO	QRC-fusion	.57 / .00	.17 / .00	.27 / .00
	ESN-fusion-m.	.20 / .00	.35 / .00	.27 / .00
IDO	QRC-fusion	.39 / .00	.56 / .10	.54 / .67
	ESN-fusion-m.	.44 / .00	.46 / .00	.44 / .00
NFT	QRC-fusion	.80 / .40	.77 / .30	.51 / .30
	ESN-fusion-m.	.67 / .00	.90 / .40	.90 / .80

Cells: PR-AUC / Recall.

Under realistic prevalence, Recall collapses to zero for most ICO configurations, indicating that scores are not calibrated for rare-event detection. PR-AUC remains informative (ICO 0.20–0.57, NFT 0.51–0.90), but operational deployment requires threshold recalibration or asymmetric cost weighting.

4.6. Discussion

Four findings emerge from the empirical analysis. Throughout, we distinguish *negative results* (what was tested and not observed), *experimental limitations* (what the setup cannot decide), and *general conclusions* (what the evidence supports):

1. **Intra-domain signal is real.** All four behavioral series achieve PR-AUC well above the random baseline within their respective domains, indicating that on-chain temporal traces carry fraud information across ICO, IDO, and NFT mechanisms.
2. **Cross-domain transfer was not observed.** Zero-shot generalization from ICO to IDO/NFT fails, with PR-AUC *below* the random baseline. Within the domains and fraud definitions considered in this study, the series are predictive intra-domain but not invariant; whether this reflects genuinely domain-specific behavior or differences between the fraud definitions and pipelines cannot be decided by this design.
3. **Classical fusion reservoirs are competitive.** ESN-fusion-matched is the best-performing model on the primary ICO benchmark (PR-AUC 0.808), with no significant advantage for QRC under the paired test of Section 4.1.4. In this setting the multichannel architecture appears to combine complementary fraud signals more effectively than per-series or quantum alternatives, though with 5 folds we report this as a descriptive ranking rather than a proven separation.
4. **A unified threshold policy enables fair cross-family comparison.** Fixing $\tau = 0.5$ for all families with PR-AUC as the primary criterion makes ranking differences reflect representations, not operating points. We do not attribute the high Recall of prior work to threshold choice: at $\sim$58% prevalence a fixed $\tau = 0.5$ is not expected to inflate Recall, which is consistent with genuine class separation; the collapse diagnostic instead flags degenerate always-positive folds in our own runs (e.g. ESN-matched at 20d). Recall/Precision differences vs. prior work may also stem from architecture: LSTMs adapt their internal representation, reservoirs train only the readout, and this adaptive-capacity gap shifts trade-offs independently of any threshold.

5. Conclusion

This paper investigated whether four on-chain behavioral time series — NEWHOLDER, NEWUSER, BIGBUYER, and GAS/GASLIMIT, proposed for ICO fraud detection by [da Mata Caffé et al. 2023] — act as transferable fraud indicators across ICOs, IDOs, and NFT collections, under a uniform collection protocol, matched classical/quantum reservoir models, and a threshold-aware framework. The main contribution is the cross-domain evaluation itself — benchmark, protocol, and transfer analysis — rather than a new fraud detector. Intra-domain, the four series discriminate above baseline: under fixed $\tau = 0.5$, ESN-fusion-matched reaches PR-AUC $= 0.808$ (ROC-AUC $= 0.768$) on ICO without collapse, and all hypotheses H1–H4 are supported in each domain with PR-AUC well above prevalence.

Notwithstanding these intra-domain results, behavioral invariance was not observed across the domains and fraud definitions considered in this study. Zero-shot transfer from ICO to IDO and NFT yielded PR-AUC values *below* the ≈ 0.60 random baseline (0.51–0.52), and combined training partially recovers IDO and NFT performance while degrading ICO discriminability, a pattern consistent with domain shift. This negative result should be read as absence of transfer under the studied conditions — not as

proof that no shared behavioral signal exists across launch modalities. Under realistic fraud prevalence (5–10%), Recall collapses to zero for most configurations with $\tau = 0.5$, underscoring that threshold recalibration is a prerequisite for any operational screening deployment.

Three methodological contributions extend beyond the empirical findings: the reformulation of the series as hypotheses grounded in observable economic mechanics; the reproducible cross-domain protocol (Etherscan extraction at IDO 100%/NFT 87.6%, versioned labels, standardized anchoring); and the threshold-aware framework with a collapse diagnostic that flags degenerate always-positive folds in our own runs — not a reinterpretation of prior results.

On the model comparison, ESN-fusion reached a slightly higher point estimate than QRC-fusion on ICO (PR-AUC 0.808 vs. 0.790) under matched dimensionality and fixed threshold, with no significant difference under the paired test (Section 4.1.4). This neither refutes nor endorses QRC as a paradigm: for short, normalized on-chain series with a linear readout, the simulated configurations provided no measurable benefit. QRC ran as exact classical simulation of small circuits (6–8 qubits); no hardware or speedup was claimed. The multichannel ESN ingests the four series natively, whereas QRC-fusion maps channels to qubits by fixed assignment, which may explain its weaker fusion gains.

Limitations. The three corpora embody distinct fraud definitions — SEC-style enforcement (ICO), rug pulls (IDO), and cybersquatting (NFT) — so cross-domain results conflate generalization across launch modalities with generalization across labeling functions, and the observed lack of transfer may partly reflect this semantic shift of the labels (Section 3.3). NFT coverage is 87.6%, with 31 excluded collections likely biased toward sparse on-chain activity. IDO/NFT labels (rug-pull, cybersquatting) may also be structurally more separable than SEC-style ICO fraud, inflating cross-domain corpus PR-AUC as an upper bound. ICO and IDO/NFT feature pipelines differ (paper pickles vs. Etherscan CSVs with per-token normalization), which may compound zero-shot failure. QRC did not demonstrate a systematic advantage over ESN under the fair fixed-threshold protocol.

Future work: (i) threshold recalibration (Platt scaling, isotonic regression) and class weighting at 5–10% prevalence; (ii) domain adaptation — feature alignment and domain-adversarial training [Ganin et al. 2016, Wilson and Cook 2020] — to test whether transfer can be recovered once covariate and concept shift are controlled; (iii) deeper study of the absence of QRC gains, whose candidate explanations include limited circuit expressivity at N=6–8 qubits, the linear readout, the 91 trainable parameters shared by both matched models, $n \leq 250$ per domain, and the near-Markovian nature of short normalized series — with hardware-based QRC and larger qubit counts as next steps; and (iv) extension to emerging modalities such as liquid staking and real-world asset tokenization.

Acknowledgments

The authors acknowledge the financial support of the National Institute of Science and Technology for Applied Quantum Computing through CNPq process No. 408884/2024-0. We acknowledge the use of Claude and ChatGPT as auxiliary tools during the writing, textual revision, translation, and code organization stages. All technical content, method-

ological decisions, result analyses, and the final version of the article were reviewed, validated, and remain the responsibility of the authors.

References

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). A theory of learning from different domains. *Machine Learning*, 79(1–2):151–175.

Buterin, V. (2014). Ethereum: A next-generation smart contract and decentralized application platform. White Paper, Ethereum Foundation.

Cernera, F., La Morgia, M., Mei, A., and Sassi, F. (2023). Ready, aim, snipe! Analysis of sniper bots and their impact on the DeFi ecosystem. In *Proceedings of the ACM Web Conference (WWW)*.

Chen, W., Zheng, Z., Cui, J., Ngai, E., Zheng, P., and Zhou, Y. (2018). Detecting Ponzi schemes on Ethereum: Towards healthier blockchain technology. In *Proceedings of the 27th International Conference on World Wide Web (WWW)*, pages 1409–1418.

Chen, W., Zheng, Z., Ngai, E. C.-H., Zheng, P., and Zhou, Y. (2019). Exploiting blockchain data to detect smart Ponzi schemes on Ethereum. *IEEE Access*, 7:37575–37586.

da Mata Caffé, L. A. Z., Braga, R. Z., Júnior, L. A. P., de Azevedo Castro Cesar, C., and Marcondes, C. A. C. (2023). Detecção de fraudes em criptomoedas utilizando métodos de classificação de séries temporais baseados em redes neurais. In *Anais do XXIII Simpósio Brasileiro de Cibersegurança (SBSeg)*, pages 484–497, Juiz de Fora/MG. SBC.

dos Santos, A. F. P. (2026). QRC-Lab: An educational toolbox for quantum reservoir computing. *arXiv preprint arXiv:2602.03522*.

Federal Trade Commission (2026). Consumer sentinel network data book: Social media fraud losses. Technical report, FTC Bureau of Consumer Protection.

Fujii, K. and Nakajima, K. (2017). Harnessing disordered-ensemble quantum dynamics for machine learning. *Physical Review Applied*, 8(2):024030.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35.

Huang, J., He, N., Ma, K., Xiao, J., and Wang, H. (2023). A deep dive into NFT rug pulls. In *Proceedings of the ACM Internet Measurement Conference (IMC)*.

Jaeger, H. (2001). The "echo state" approach to analysing and training recurrent neural networks. *GMD Technical Report*, 148.

Lukoševičius, M. and Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149.

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., and Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530.

Nakajima, K., Fujii, K., Negoro, M., Mitarai, K., and Kitagawa, M. (2019). Boosting computational power through spatial multiplexing in quantum reservoir computing. *Physical Review Applied*, 11(3):034021.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D., editors (2009). *Dataset Shift in Machine Learning*. MIT Press, Cambridge, MA.

U.S. Department of Justice (2022). Two defendants charged in non-fungible token (NFT) fraud and money laundering scheme. Press release 22-087, U.S. Attorney's Office, Southern District of New York.

U.S. Securities and Exchange Commission (2018). SEC charges founders of centra tech inc. with fraudulent initial coin offering. Press release 2018-53, SEC Enforcement Division, Cyber Unit.

Wilson, G. and Cook, D. J. (2020). A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology*, 11(5):1–46.

Wood, G. (2014). Ethereum: A secure decentralised generalised transaction ledger. Technical report, Ethereum Project. Yellow Paper.

Xia, P., Wang, H., Gao, B., Su, W., Yu, Z., Luo, X., Zhang, C., Xiao, X., and Xu, G. (2021). Trade or trick? Detecting and characterizing scam tokens on Uniswap decentralized exchange. *Proceedings of the ACM on Measurement and Analysis of Computing Systems (POMACS)*, 5(3):1–26.

Xu, J. and Livshits, B. (2019). The anatomy of a cryptocurrency pump-and-dump scheme. In *Proceedings of the 28th USENIX Security Symposium*, pages 1609–1625.

Appendix: Reproducibility

All code, versioned label files (`data/ido_labels.csv`, `data/nft_labels.csv`), extraction scripts, per-fold metric outputs, and step-by-step execution instructions are provided in the public repository `https://github.com/Lixipluv/FraudDetectionICOIDONFT`. All runs use stratified 5-fold cross-validation with a fixed seed (42) and the fixed decision threshold $\tau = 0.5$; the ESN configurations are $n = 18$ (matched) and $n = 500$ (large) with spectral radius 0.9, and QRC uses $N = 6$ qubits with an Ising topology and rich Pauli observables, all feeding an L_2 logistic readout ($C = 0.01$). The complete repository layout, the full command lines to reproduce every experiment, and the exhaustive hyperparameter listing are provided as supplementary material in the artifact, so that the paper retains only what is needed to understand and re-run the study.