Published in AJSE, Vol: 25, Issue:2
Received on 25th March 2026
Revised on 9th September 2026
Published on 31st October 2026

AIUB JOURNAL OF SCIENCE AND ENGINEERING
ISSN: 1608 - 3679 (print) 2520 - 4890 (Online)
https://doi.org/10.53799/kg3xc638

Combating Market Manipulation: A Wasserstein GAN-GP Augmented, Focal-Loss TabNet for Cryptocurrency Pump-and-Dump Detection

Muhammad Nazeer Musa, Umar Faruq Abdulrazaq, and Tijjani Abdullahi

***Abstract*— Pump-and-dump (P&D) manipulation is one of the most persistent forms of cyber-enabled fraud in cryptocurrency markets, and detecting it automatically is made difficult by the extreme scarcity of confirmed P&D events relative to normal trading activity. This study develops an augmentation strategy that integrates a fully specified Wasserstein GAN with gradient penalty, and an alpha-balanced focal-loss TabNet classifier for transaction-level P&D detection. A WGAN-GP generator is trained independently within each cross-validation fold on the training partition only, contributing 30,000 synthetic minority rows per fold, combined with a bounded thirty-fold replication of genuine minority transactions to reach a documented augmented imbalance ratio of approximately 8.9 majority rows to one minority row. On the primary 25-second resolution, the full proposed model attained a precision of 100.0%, recall of 85.7%, and F1-score of 92.3% on a stratified held-out test split, and a cross-validated mean F1-score of 0.875 (precision 0.878, recall 0.874) under 5-fold stratified cross-validation with fold-wise generator retraining. A five-configuration ablation isolates the marginal contribution of WGAN-GP synthesis (ΔF1 +0.019 over a class-weighted baseline), hard-negative mining (+0.006), and focal loss (+0.038), demonstrating that the generative component contributes meaningfully but that the largest single improvement is attributable to the loss function rather than to the generator. Applying the 25-second-tuned hyperparameter configuration without re-tuning to the 15 and 5 second aggregations yields cross-validated F1-scores of 0.840 and 0.766 respectively, quantifying for the first time on this corpus the cost of resolution under-tuning in a GAN-augmented pipeline.**

***Index Terms*— Cryptocurrency fraud detection, pump-and-dump schemes, TabNet, Wasserstein GAN with gradient penalty, class imbalance, focal loss**

I. Introduction

The rise of cryptocurrencies, viewed as a revolutionary development in international finance, has practically coincided with the rise of sophisticated financial crimes exploiting the decentralized and pseudonymous nature of the blockchain ecosystem [1]. Among them are pump-and-dump (P&D) schemes that have been widespread in traditional markets and that have been adapted for use in cryptocurrency markets using the latter's inefficiencies, volatility, and information asymmetry [2][3]. Such schemes involve organized manipulation of prices by means of artificial buzz and sudden dumping, which leads to retail investors suffering from heavy losses and losing faith in the cryptocurrency market [4][5]. This type of fraud has serious socio-economic consequences, yet is still not detected effectively due to the constant changes in the malefactors' modus operandi, the lack of labeled data, and complexity in modeling transactions in decentralized markets. Traditionally used methods of detecting fraud that rely on rules and other simple statistics are unable to cope with the evolving habits of criminals and the nonlinear relations of data in the high-dimensional environment of cryptocurrencies [8]. Although machine learning [20] shows great potential in the detection of sophisticated manipulation patterns, the existing solutions find themselves stuck due to class imbalance, lack of translatability to different cryptocurrencies, and poor manipulation of transaction data [7,8]. The novel solution to the issue of class imbalance in tabular financial fraud detection is based on generating synthetic data with the help of generative adversarial networks (GAN) [9][30][31], which works as a generator that learns to comprehend the nonlinear structure of minority class, unlike other classical interpolation methods such as SMOTE [25]. However, existing GAN-augmented P&D detection studies [32] have predominantly reported architecturally unspecified generators and unbounded augmentation ratios poorly suited to a minority class of only a few hundred instances, leaving a critical gap in its application to this problem.

Concurrently, TabNet [10] has emerged as a deep architecture purpose-built for tabular data, combining sequential, attention-gated feature selection with the representational capacity of a neural network, and offering a route to interpretability that purely black-box architectures do not [32]. While TabNet has been applied to a range of tabular financial-fraud problems, its combination with a multi-resolution evaluation protocol has not, to our knowledge, previously been reported for cryptocurrency P&D detection.

Novelty statement. The novelty of this work is not a new network module but a methodologically rigorous, augmentation-leakage-controlled benchmark that, to our

Muhammad Nazeer Musa (e-mail: muhammadmusa2502@nda.edu.ng) is with the Department of Cyber Security, Nigerian Defence Academy, Kaduna, Nigeria.

Tijjani Abdullahi (e-mail: tijjani.abdullahi@bazeuniversity.edu.ng) is with the Department of Computer Science, Baze University, Abuja, Nigeria.

Umar Faruq Abdulrazaq (e-mail: auf41@yahoo.com) is with the Centre for Cyberspace Studies, Nasarawa State University, Keffi, Nasarawa State, Nigeria.

knowledge, is the first cryptocurrency P&D detection study to jointly combine: (i) a fully specified, fold-wise-retrained WGAN-GP generator that eliminates synthetic-to-validation augmentation leakage; (ii) a documented, bounded augmentation volume (30,000 synthetic rows, an 8.9:1 final imbalance ratio) in place of an unbounded, unjustified synthetic-to-real ratio as in [32]; (iii) a matched-volume comparison against a SMOTE baseline at identical augmentation volume, isolating the marginal value of generative synthesis over interpolation-based resampling under a fixed TabNet backbone; (iv) a five-configuration ablation isolating the marginal contribution of GAN synthesis, hard-negative mining and focal loss from the TabNet backbone, reported as a single-split comparison whose margins are interpreted against the same 63-instance sensitivity established in Section IV-C rather than treated as statistically confirmed; and (v) multi-resolution evaluation at 5-, 15- and 25-second transaction-chunk granularities using a single, 25-second-tuned hyperparameter configuration, exposing the performance cost of resolution-specific under-tuning rather than assuming that a result obtained at one resolution generalises to others. We explicitly frame the system as a research benchmark for GAN-augmented tabular fraud detection, not a deployment-ready market-surveillance product.

This study makes the following contributions:

- We establish an augmentation-leakage-controlled benchmark for cryptocurrency pump-and-dump detection, in which the WGAN-GP generator is retrained inside every cross-validation fold using only that fold's training partition, eliminating the synthetic-to-validation contamination that affected the earlier, single-split version of this work.
- We conduct a comparative analysis of five augmentation configurations (class-weighting only, SMOTE, WGAN-GP alone, WGAN-GP with hard-negative mining, and the full proposed pipeline) under an identical TabNet backbone.
- We evaluate the proposed pipeline at three transaction-chunk resolutions on 5, 15 and 25 seconds using a single hyperparameter configuration tuned only at 25 seconds, quantifying the performance cost of applying an under-tuned pipeline to finer temporal granularities and benchmarking each resolution directly against the corresponding published results of Chadalapaka et al. [23].
- We provide a comparative analysis against seven prior published studies on the same underlying corpus, situating the model's performance within the existing published state of the art.
- We report a detailed ablation study quantifying the individual contribution of WGAN-GP synthesis, hard-negative mining and focal loss, and we discuss the resulting precision–recall trade-off explicitly as a value judgement relevant to market-surveillance deployment.

II. LITERATURE REVIEW

Pump-and-dump schemes have a long history in traditional equity markets and have migrated into the largely unregulated cryptocurrency space, where an organising group accumulates a position in an illiquid, low-capitalisation asset, disseminates coordinated hype through channels such as Telegram to attract buying pressure, and liquidates its position into the resulting demand, leaving later entrants to absorb the loss [2], [4] as depicted in Fig. 1.

The typical cryptocurrency P&D scheme unfolds in several stages, the organizers quietly accumulate a large position in a chosen cryptocurrency, often a low-priced, illiquid asset. The second step of the operation involves organizing a so-called "pump" phase where the managers spread false information and hype the potential of the target cryptocurrency through social media channels, online forums or private messaging platforms, which leads to the rise of the cryptocurrency price. Finally, during the "dump" phase, the organizers sell their holdings at a high price, causing the price to go down and leave those that bought late with substantial losses [2], [11].

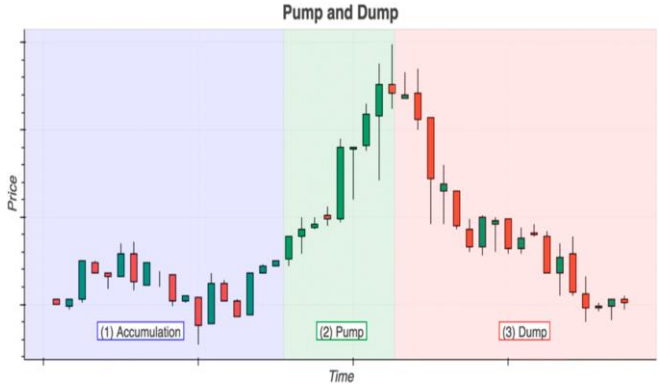

Fig. 1. The 3 phases of Pump and Dump Events [2].

The illustration in Fig. 1 depicts three stages that describe a cryptocurrency pump-and-dump operation. In the accumulation stage, scammers buy a significant amount of a specific cryptocurrency at an affordable price. This prevents them from raising a price too early and draining too much money at the beginning. This stage is followed by pump run. In this stage, fraudsters spread false and inflated positive news through various channels, such as web forums and social media, to create artificial demand for the cryptocurrency and invest money. The inflow of the investors creates buying pressure that quickly rises the price. The final stage is the dump. In this stage, scammers sell their huge amounts of cryptocurrency. Their activities boost supply and significantly decline its price for late investors and provide some earnings from initial price growth [5].

The factors of pump-and-dump (P&D) phenomenon popularity in the crypto sphere are the market immaturity and the absence of legislative regulations across multiple jurisdictions. Furthermore, the inherently volatile cryptocurrency market leads to the rapid price changes as the result of speculation and social media trends [1] [15]. Kamps and Kleinberg [2] used statistical thresholding method in their

early works devoted to the P&D detection, while La Morgia et al. [4] [11] carried out a research work during which they collected the two-year Telegram-source Binance benchmark corpus across 5, 15, and 25 second transaction-chunk resolutions used in this paper.

Conventional detection approaches rely on rule-based alerting, statistical anomaly detection [21], or technical indicators such as the moving-average convergence/divergence and relative-strength index [13], [16], [17]. These methods are attractive for their transparency but are ill-suited to cryptocurrency markets in practice: high baseline volatility inflates false-positive rates [8], static rules are straightforward for organisers to evade once understood [2], and purely statistical tests do not capture the joint, nonlinear interaction of order-flow and price features that characterise a coordinated pump [18],[21]. These limitations motivated a shift toward learned, data-driven detectors.

A lot of recent studies have examined the potential of supervised machine learning and deep learning in the context of the P&D detection problem. Victor and Hagemann [6] integrated trade data from Binance with Telegram signal timestamps within the 30-minute interval so as to demonstrate that the results of the XGBoost model based on 10-fold cross-validation can lead to AUC values of 0.995, 0.855 for sensitivity, and 0.997 for specificity. Fantazzini and Xiao [7] targeted the problem of serious class imbalance by analyzing transaction chunks lasting for 30 seconds and establishing that the modification of the threshold through the use of 5-fold cross-validation on unprocessed data works better than the traditional resampling techniques. In particular, the threshold of 2.6% resulted in a sensitivity of 98% and specificity of 89%, while SMOTE with Random Forest Classifier secured only 75% sensitivity and 99% specificity, which served as the first example of the precision-recall trade-off in this work.

Furthermore, Alfano et al. [24] demonstrated the value of alternative data sources, revealing that supplementing financial features with social-media signals improved a Random Forest classifier's F1-score by 6%, 3%, and 1% for 5-second, 15-second, and 25-second chunk resolutions, respectively. Their 5-fold cross-validation on a corpus of P&D events also indicated that predictions at the 5-second resolution were comparatively less accurate. In a narrower, asset-specific study, Shao [22] applied a Random Forest model exclusively to Dogecoin, reporting 99% accuracy and an 84% F1-score.

Investigating the impact of temporal multi chunk resolution on the detection of pump-and-dump (P&D) schemes, La Morgia et al. [11] systematically evaluated a Random Forest (RF) classifier across 5, 15, and 25 second transaction windows using both 5-fold cross-validation. The results show that there is a strong positive relationship between the size of the window of aggregation and the effectiveness of prediction. Specifically, the use of the RF algorithm with 5-second resolution rendered an F1-score of 84.0% (precision - 92.4%, recall – 78.4%) according to the results of 5-fold split, which became even lower after the introduction of 10-fold cross-validation. The increase of the timeframe of analysis to 15 seconds contributed to an increase in the model's sensitivity, providing F1-score of 87.7%. higher performance was achieved with 25 seconds resolution, as the classifier proved to achieve the F1-score equal to 91.8% (precision – 93.7%, recall – 91.3%) according to the findings under 5-fold cross-validation.

In a complementary analysis employing 5-fold cross-validation, La Morgia et al. [4] further corroborated the robustness of ensemble methods in this domain, indicating that predictive accuracy scales favorably as the chunk size increases. Their RF architecture recorded successive F1-scores of 82.4%, 90.0%, and 94.5% on the 5-, 15-, and 25-second intervals. Chadalapaka et al. [23] applied a convolutional LSTM (C-LSTM) and a supervised Anomaly Transformer across all three chunk resolutions of the La Morgia corpus on an 80/20 train-test split, reporting F1-scores of 83.7%, 89.3% and 89.3% (C-LSTM) and 89.3%, 93.6% and 89.2% (Anomaly Transformer) at 5, 15 and 25 seconds respectively — to our knowledge the only prior study to report deep-learning performance across all three resolutions on this corpus, and the primary point of comparison for the multi-resolution evaluation reported in Section IV-E of this study.

The imbalance issue encountered in pump-and-dump detection, where real transactions are highly salient compared to manipulative transactions, led to impairments in the performance of algorithms in terms of F1 score in this particular field. Models which yield high accuracy usually yield low recall and low F1 score [7]. Given such imbalance, the models in question tend to be biased towards larger classes and fail to classify the needed information pertaining to the smaller class, as in this case pump-and-dump events. In order to tackle this issue various methods of generating synthetic data that would balance out the number of cases and enhance the performance of the algorithms are being widely researched. Synthetic Minority Over-sampling Technique (SMOTE) [25] and Random Over-Sampling Examples (ROSE) [26] are some of the widely used classical techniques; each carries a documented weakness, and more recent evidence indicates these effects are not merely theoretical as Tantithamthavorn et al. [27] and Wongvorachan et al. [28] both report settings in which class-rebalancing techniques reduce downstream performance or distort model interpretation relative to leaving the imbalance untouched. Generative adversarial networks [9] have been proposed as a more expressive alternative, on the premise that a learned generator can capture nonlinear minority-class structure that linear interpolation cannot; Sauber-Cole and Khoshgoftaar [30] noted that GAN-based tabular augmentation is promising but lacked procedural similarity to compare across architectures, and generally ignored the question of the optimal amount level or percent of injection of GAN-generated instances for machine learning efficacy. Also, training instability and mode collapse reoccur, particularly for minority classes of only a few hundred examples, as is the case for confirmed P&D events. The Wasserstein GAN with gradient penalty (WGAN-GP) [33] was proposed specifically to stabilise adversarial training via a gradient-penalised critic

objective which has shown to perform good for complex dataset [30], and Sharma et al. [31] combine SMOTE and GAN-based oversampling in a SMOTified-GAN, reporting improvements over either technique alone on several benchmark imbalanced datasets, though not at the extreme minority counts and augmentation ratios typical of cryptocurrency P&D data.

Abdulrazaq and Musa [32] introduced a promising hybrid SmoteGAN and TabNet model for P&D detection, reporting 98% precision, 83% recall, and a 90% F1-score on 25-second chunks using a single train/test split. To further strengthen this approach, several areas for methodological refinement were identified. Specifically, the generator architecture would benefit from a more detailed structural specification, and the exceptionally high synthetic-to-real augmentation warrants a more conservative adjustment. Additionally, an ablation study is needed to clearly isolate the specific drivers of the model's performance. The present study builds upon this foundational work by systematically addressing these refinements in Section III.

III. RESEARCH METHODOLOGY

The research methodology used for this research is incorporated into Fig. 2. It involves data collection, preprocessing, data augmentation with WGANs, building the model with TabNet, and finally the evaluation.

A. Input Dataset

This study reuses the manually labelled Binance transaction corpus originally compiled by La Morgia et al. [4],[11]. The corpus was constructed by monitoring P&D-operated Telegram channels for two years, recording the timestamp of each announced pump signal, and extracting Binance transaction data for the affected trading pair over a period of time, yielding P&D occurrences aggregated into three chunk resolutions (5, 15 and 25 seconds), each carrying the fifteen features listed in Table 1.

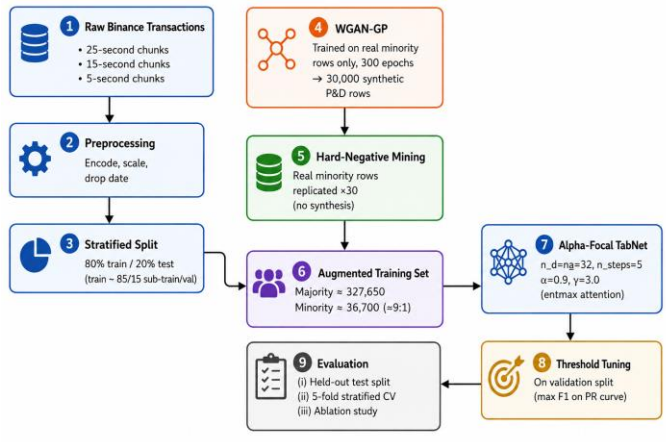

Fig. 2. Research Methodology Workflow

Resolution overview. Unlike the study of Abdulrazaq & Musa [32], which evaluated only the 25-second aggregation, this research evaluates all three published resolutions. At 25-second chunk— the primary time resolution at which the pipeline is tuned, the feature matrix contains 482,157 rows, of which 317 (0.066%) are labelled P&D activity while 481,840 are for normal activity. These labels are derived from [12], a public repository of La Morgia et al. [4], with an imbalance ratio of approximately 1,521 majority rows to one minority row before augmentation. Even though La Morgia et al. [4], [11] report having collected and annotated 343 P&D occurrences across multiple exchanges, but only 317 are confirmed in [12]. The set of features with their descriptions are shown in Table 1.

TABLE 1. DATASET FEATURES

S/N	Feature	Type	Description
1	Date	String	Date of the chunk
2	HourSin, HourCos, MinuteSin, MinuteCos	Float	Positional encoding of the hour and minute
3	PumpIndex	Integer	Index of the pump
4	Symbol	String	Cryptocurrency symbol
5	StdRushOrder, AvgRushOrder	Float	Standard deviation and average percent change of rush orders
6	StdTrades	Float	Standard deviation of trades
7	StdVolume, AvgVolume	Float	Standard deviation and average percent change of volume
8	StdPrice, AvgPrice, AvgPriceMax	Float	Standard deviation, average percent change, and average maximum percent change of price

B. Data Preparation and Pre-processing

Following data collection, a data preparation and preprocessing pipeline was established to ensure the data is suitable for training. We applied the following transformations to set the data for model training:

i. **Symbol Encoding:** The categorical "Symbol" feature was converted into numerical feature representation using label encoding.

ii. **Date Removal:** The "Date" feature was excluded from the dataset as it was deemed redundant due to the inclusion of time-based features.

iii. **Normalization:** To ensure features were on a comparable scale, all numerical features were numerical features were standardised (zero mean, unit variance) with a StandardScaler fitted on the training partition only, to avoid test-set leakage into the scaling statistics. This is due to some algorithms being sensitive to feature scaling. Meaning features, with for example, integer might have a higher weight than features with float.

C. Data Splitting

Data splitting supports both the single held-out evaluation and the 5-fold cross-validation reported in Section IV. For the single-split evaluation at 25-second resolution, the preprocessed data were partitioned into an 80% training set and a 20% held-out test set using stratified sampling on the

class label, and the training set was further partitioned into an 85% sub-training set and a 15% validation set, again stratified, with the validation split used exclusively for early stopping and post-hoc decision-threshold tuning (Section III-I). This yields a held-out test set of 96,432 transactions, of which 63 (0.065%) are P&D instances. For the cross-validation evaluation (Section III-J), the entire labelled dataset was partitioned into five stratified folds, and the augmentation pipeline described in Sections III-D–III-F was re-fitted independently within each training fold to prevent information from validation folds leaking into the synthetic-data generator, following the fold-wise protocol formalised in Algorithm 1.

Algorithm 1: Augmentation-Leakage-Controlled Fold-Wise Protocol

```
Partition N curated transaction rows into 5 stratified folds with a fixed random seed
for k = 1 to 5 do
    D_train^(k) ← 4 folds (real transactions only)
    D_val^(k)   ← held-out fold (real transactions only)
    Train WGAN − GP^(k) on minority rows of D_train^(k) only (300 epochs, 5:1 critic:generator updates)
    D_syn^(k) ← sample 30,000 synthetic minority rows from WGAN − GP^(k)
    D_hard^(k) ← replicate real minority rows of D_train^(k) thirty-fold (no synthesis)
    Train Focal-TabNet on D_train^(k) ∪ D_syn^(k) ∪ D_hard^(k)
    Tune decision threshold on a held-out sub-validation split drawn from D_train^(k)
    Evaluate on D_val^(k) (real transactions only, threshold applied unchanged)
end for
Aggregate metrics across folds
```

As in a fold-wise-retrained augmentation protocol, this design yields five independent WGAN-GP generators and prevents any synthetic row — or feature representation derived from a held-out real row — from appearing in a validation fold. The validation partition in every fold contains real transactions only.

D. WGAN-GP Synthetic Minority-Sample Generator

This section presents Wasserstein GAN with gradient penalty [33], which was used to carry out training on the real records of the smaller class in the training partition (216 instances in the single-split configuration of 25 seconds). The choice of WGAN compared to a standard GAN was driven by data explaining the stability of training through critic-based gradient-penalising approach adopted in the implementation of Wasserstein GAN [33] as well as its ability to produce more stable results in terms of loss than a standard GAN with a poor gradient discriminator.

All necessary details about the architecture of G and C can be seen in Table 2. The generator G is mapping latent vector $z \sim N(0, I)$ of size 100 to a synthetic feature row of size 14 using two hidden layers (64 and 128 units respectively), batch normalisation and activation in the form of LeakyReLU (weight 0.2), with one linear layer at the output. The critic C maps a 14-dimensional row to a scalar Wasserstein score through two hidden layers (128 and 64 units) with LeakyReLU activations and no batch normalisation, consistent with the WGAN-GP recommendation that the critic avoid normalisation layers that couple examples within a batch [33]. Fig. 3 illustrates this architecture; Algorithm 2 gives the complete specification alongside the downstream classifier.

TABLE 2. WGAN-GP ARCHITECTURE AND TRAINING SPECIFICATION

Hyperparameter	Value
Latent dimension	100
Generator hidden layers	100 → 64 → 128 → 14 (BatchNorm + LeakyReLU 0.2 on hidden layers)
Critic hidden layers	14 → 128 → 64 → 1 (LeakyReLU 0.2, no normalisation)
Loss	Wasserstein distance with gradient penalty (λ = 10)
Critic : generator update ratio	5 : 1
Optimiser (both networks)	Adam, lr = 1×10^{-4}, $\beta_1 = 0.5$, $\beta_2 = 0.9$
Batch size	32
Training epochs	300
Synthetic samples generated	30,000 (per fold)

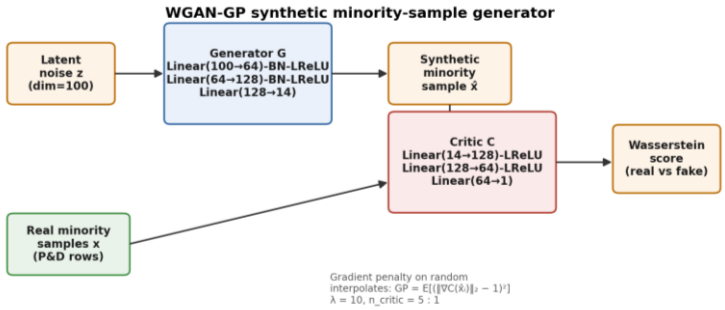

Fig. 3. WGAN-GP architecture used for minority-class synthesis: generator (top) and critic (bottom), trained with a gradient-penalised Wasserstein objective.

The critic loss (D_{loss}) is minimized during training using the Wasserstein objective regularized by a gradient penalty term:

$$D_{\text{loss}} = E\left[C\left(G(z)\right)\right] - E[C(x)] + \lambda \cdot \text{GP} \quad (1)$$

Where:

- $E\left[C\left(G(z)\right)\right]$ is the expected value of the critic's output over generated (synthetic) samples $G(z)$ from latent noise z.
- $E[C(x)]$ is the expected value of the critic's output over real minority-class samples x.
- λ is the gradient penalty coefficient, set to $\lambda = 10$.
- GP is the gradient penalty computed on interpolates between real and generated samples:

$$\text{GP} = E[(\|\nabla C(\hat{x})_2\| - 1)^2] \quad (2)$$

where the interpolate $\hat{x} = \varepsilon x + (1-\varepsilon)G(z)$ is formed along a straight line between a real minority row x and a generated row G(z), with ε drawn uniformly from [0, 1] independently for each interpolate; the penalty term drives the critic's gradient norm at each interpolate toward 1, the condition required of a 1-Lipschitz function under the Wasserstein-distance formulation of Gulrajani et al. [33].

Both networks are optimised with Adam (learning rate 1×10^{-4}, $\beta_1 = 0.5$, $\beta_2 = 0.9$) over 300 epochs with a batch size of 32. After training, the generator produces 30,000 synthetic minority rows per fold by sampling fresh latent vectors.

E. Hard-Negative Mining and Augmentation Volume

This is applied in two ways. First, the WGAN-GP contribution has been reduced by more than two orders of magnitude, to 30,000 synthetic rows — a synthetic-to-real ratio of approximately 139:1 relative to the 216 real minority rows available in the active 25-second training partition. Second, rather than relying on synthetic data alone, the training set is further augmented with a hard-negative-mining step in which the real minority rows are replicated thirty-fold, without any synthesis, to increase the effective weight the classifier places on genuine, rather than generator-approximated, examples of the class boundary. Combining the real minority rows, the WGAN-GP synthetic rows, and the hard-mined duplicate rows against an unchanged majority-class count of approximately 327,650 rows yields a final augmented training imbalance ratio of approximately 8.9 majority rows to one minority row at 25-second resolution. The same volumes (30,000 synthetic rows, thirty-fold hard mining) are applied unchanged at the 15-second and 5-second resolutions in the multi-resolution evaluation of Section III-H.

F. Alpha-Balanced Focal Loss

The downstream classifier is trained with an alpha-balanced focal loss [34] in place of unweighted cross-entropy, defined for a single example with ground-truth class t and predicted probability p_t for the true class:

$$\mathrm{FL}(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \qquad (3)$$

Where:

- $\alpha_t = \alpha = 0.9$ for the minority (P&D) class.
- $\alpha_t = 1-\alpha = 0.1$ for the majority (Normal) class.
- $\gamma = 3.0$ is the focusing parameter that down-weights easy, well-classified examples

This objective is complementary to, rather than a substitute for, WGAN-GP augmentation: where augmentation changes what the classifier sees during training, focal loss changes how much each example is allowed to influence the gradient once seen.

The three imbalance-correction components in the full pipeline interact, and their combined effect should not be assumed to multiply. Augmentation and hard-negative mining leave the active training partition at approximately 8.9 majority rows for every minority row (Section III-E); focal loss then applies a further per-example weight of α = 0.9 to minority examples and 1 − α = 0.1 to majority examples, a 9:1 ratio in isolation. Combining the two, the total loss mass contributed by the minority class across an epoch is proportional to 1 × 0.9 = 0.9, against 8.9 × 0.1 = 0.89 for the majority class — the two are within one percent of parity, not stacked multiplicatively into an approximately 80:1 emphasis on the minority class that naively multiplying 8.9 by 9 would suggest. This combined weighting should not be mistaken for an 80:1 class ratio in the training set; rather, it represents the effective contribution of each minority example to the loss gradient. The 30,000 WGAN-GP synthetic rows are included in this calculation at full weight, as the focal loss is applied identically to real and synthetic minority examples.

Algorithm 2: WGAN-GP and Focal-TabNet Architecture Specifications

```
WGAN-GP Generator (G):
    Input: latent vector z ∈ R^100 ~ N(0, I)
    Linear(100 → 64), BatchNorm, LeakyReLU(α = 0.2)
    Linear(64 → 128), BatchNorm, LeakyReLU(α = 0.2)
    Linear(128 → 14)   [synthetic feature row]
WGAN-GP Critic (C):
    Input: feature row  x ∈ R^14
    Linear(14 → 128), LeakyReLU(α = 0.2)  [no
    normalisation]
    Linear(128 → 64), LeakyReLU(α = 0.2)
    Linear(64 → 1)   [Wasserstein score]
    Loss: E[C(G(z))] − E[C(x)] + 10 · GradientPenalty;
    critic:generator = 5:1
Hard-Negative Mining:
    D_hard ← replicate real minority rows × 30 (no
    synthesis)
Focal-TabNet Classifier:
    Feature dimension n_d = 32, attention dimension
    n_a = 32, decision steps = 5
    Mask type: entmax; sparsity coefficient
    λ_sparse = 1 × 10^−3
    Loss: alpha-balanced focal loss (α = 0.9, γ =
    3.0)
    Optimiser: Adam (lr = 1×10⁻²), StepLR (step =
    10, γ = 0.9)
    Batch size 4,096, virtual batch size 512, max
    100 epochs, patience 20
```

G. TabNet Classifier Configuration

TabNet [10] classifier architecture combines the sequential, attention-gated feature selection of a decision-tree ensemble with the representational capacity of a neural network as in [32] (see Fig. 4).

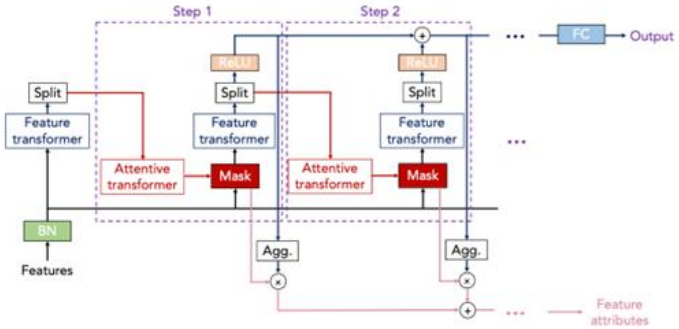

Fig. 4. TabNet Architecture

TABLE 3. TABNET HYPERPARAMETERS USED IN THE PIPELINE

Hyperparameter	Value
Feature dimension n_d	32
Attention dimension n_a	32
Number of decision steps (n_{steps})	5
Mask type	entmax
Sparsity regularisation (λ_{sparse})	1×10^{-3}
Optimiser	Adam, lr = 1×10^{-2}
Learning-rate scheduler	StepLR (step size 10, $\gamma = 0.9$)
Batch size / virtual batch size	4,096 / 512
Loss function	Alpha-balanced focal loss ($\alpha = 0.9$, $\gamma = 3.0$)
Max epochs / early-stopping patience	100 / 20 (monitored on validation balanced accuracy and AUC)

Table 3 reports the complete hyperparameter configuration used in this study: the feature and attention dimensions were 32, the number of sequential decision steps was 5, and the loss function to the alpha-balanced focal loss described above. Training used the Adam optimiser (learning rate 1×10^{-2}, decayed by a StepLR schedule with step size 10 and multiplicative factor 0.9), a batch size of 4,096 with a virtual batch size of 512 for ghost batch normalisation, and entmax attentive masking, for a maximum of 100 epochs with early stopping after 20 epochs without improvement in validation balanced accuracy, monitored jointly with validation AUC.

Algorithm 2 presents the complete architecture specification for both the WGAN-GP and TabNet components of the proposed framework.

H. Multi-time Resolution Evaluation Protocol

Beyond the augmentation-strategy ablation, this study evaluates the sensitivity of the full proposed pipeline to transaction-chunk resolution. Using the identical hyperparameter configuration tuned and cross-validated at 25-second resolution (Tables 2 and 3), without any re-tuning, we apply the full WGAN-GP + hard-negative-mining + focal-TabNet pipeline, following Algorithm 1, to the 15-second and 5-second aggregations of the same underlying La Morgia corpus [11], [23]. This design directly tests whether a pipeline whose augmentation volume, network capacity, and loss weighting were selected for one resolution transfers to others, or whether resolution-specific re-tuning is a precondition for competitive performance, the latter being the pattern reported by every prior study that has compared chunk resolutions on this corpus [4], [11], [23], [24]. Results are reported in Section IV-F and compared directly against the prior studies.

I. Decision-Threshold Tuning

Because the decision threshold of 0.5 is not calibrated for a severely imbalanced problem, the classification threshold was tuned separately from model fitting. Following training, prediction probabilities were computed on the validation split, a precision–recall curve was constructed, and the threshold maximising the F1-score on that curve was selected and then applied, unchanged, to the held-out test split. This procedure is applied identically across the baseline, the full proposed model, every configuration in the ablation study, and every chunk resolution, so that reported differences reflect the underlying training and augmentation choices rather than an arbitrary or inconsistently applied decision threshold.

J. Evaluation Protocol

Three complementary evaluation protocols are reported in Section IV as shown in Table 4. First, a single stratified 80/20 held-out split at 25-second resolution is reported for direct comparability with prior single-split literature like [23],[34]; because a single split of a 317-instance minority class necessarily places a small number of P&D events in the test set, this result is reported with an explicit acknowledgement of its limited reliability. Second, 5-fold stratified cross-validation is reported as the primary, more reliable performance estimate at 25-second resolution; within each fold, the WGAN-GP generator, the hard-negative-mining step, and the TabNet classifier are all re-fitted independently, following Algorithm 1, and the decision threshold is re-tuned on a held-out sub-validation split drawn from the same fold. Third, a five-configuration ablation study (Section IV-D) holds the TabNet backbone, its hyperparameters, and the evaluation split fixed, and varies only the augmentation strategy, so that the contribution of the generative model can be separated from the contribution of the classifier and loss function.

TABLE 4: EVALUATION METRICS

Metric	Formula	Description
Precision	TP / (TP + FP)	Proportion of positive predictions that are truly positive.
Recall (Sensitivity)	TP / (TP + FN)	Proportion of actual positives that are correctly predicted as positive.
F1-score	2 * (Precision * Recall) / (Precision + Recall)	Harmonic mean of precision and recall.
Overall Accuracy	(TP + TN) / (TP + TN + FP + FN)	Proportion of correct predictions.

Where: TP: True Positive, TN: True Negative, FP: False Positive and FN: False Negative

IV. RESULTS AND DISCUSSION

This section reports the empirical results of the revised pipeline: synthetic-sample composition, single held-out-split performance with a reassessed baseline, 5-fold cross-validation, the ablation study (evaluated on the same single held-out split, with the reliability limits established in Section IV-C carried forward explicitly rather than set aside), the multi-time resolution evaluation, and comparative analysis against prior published work.

A. Training-Set Composition After Augmentation

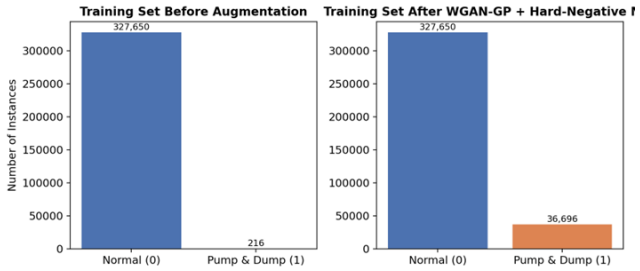

Fig. 5. Composition of the active training partition before and after WGAN-GP augmentation and hard-negative mining.

Fig. 5 summarises the composition of the active 25-second training partition before and after the augmentation pipeline of Sections III-D and III-E. Before augmentation, the training partition contains approximately 327,650 majority rows against 216 real minority rows (≈1,517:1). After augmentation, the minority count rises to approximately 36,700 rows (216 real, 30,000 WGAN-GP synthetic, 6,480 hard-mined duplicates), yielding a working ratio of approximately 8.9:1. We report this comparison in terms of the documented composition of the augmented set, so that the provenance of every synthetic and duplicated row is auditable.

B. Held-Out Test Split: Baseline and Full Model

The baseline reported by Abdulrazaq & Musa [32] used the raw imbalanced data (1,520:1) with no class weighting and no augmentation. For this study we reassessed the baseline by training the same TabNet backbone with built-in class weighting only (no synthetic or duplicated data, no focal loss). On the held-out test split, this baseline achieves precision of 0.961, recall of 0.778 and 0.860 F1-score for the P&D class (Table 5). This also serves as the first configuration in the ablation study of Section IV-D.

TABLE 5. CLASSIFICATION PERFORMANCE ON THE HELD-OUT TEST SPLIT (n = 96,432; 25-SECOND CHUNKS)

Metric	Baseline (class-weighted TabNet)	Full proposed model
Accuracy (overall)	0.9998	0.9999
Precision (P&D class)	0.961	1.000
Recall (P&D class)	0.778	0.857
F1-score (P&D class)	0.860	0.923

Against this baseline, the full proposed pipeline at a threshold of 0.84 achieves perfect precision of 1.00, recall of 0.857, and F1-score of 0.923 for the P&D class on the same held-out split (overall accuracy 0.9999; confusion matrix in Fig. 6: of 96,369 normal transactions all are correctly classified, and of 63 P&D transactions, 54 are correctly identified with 9 missed and zero false alarms).

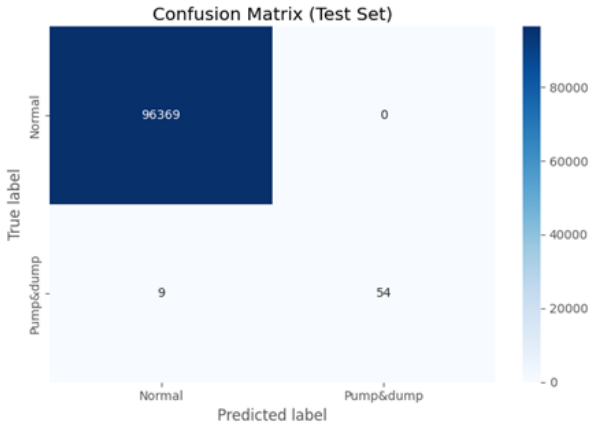

Fig. 6. Confusion matrix of the full proposed model on the held-out test split (25-second chunks).

C. 5-Fold Stratified Cross-Validation

Because the held-out test split in IV-B contains only 63 P&D instances, a single misclassification shifts recall by more than 1.5 percentage points, and the single-split result should not be read as a precise point estimate. Fig. 7 report 5-fold stratified cross-validation of the full proposed pipeline, re-fitting the WGAN-GP generator, the hard-negative-mining step and the TabNet classifier independently within each fold, following Algorithm 1.

Across the five folds, F1-score ranges from 0.826 to 0.952 (mean 0.875, SD 0.047), precision ranges from 0.817 to 0.968 (mean 0.878, SD 0.055), and recall ranges from 0.794 to 0.938 (mean 0.874, SD 0.061). Two observations follow. First, the mean cross-validated F1-score (0.875) is below the single-split F1-score reported above (0.923) by a 4.8-point gap, indicating that the single split was a moderately favourable draw rather than a representative one. Second, precision reaches 1.00 in none of the five folds (maximum of 0.968), supporting the treatment of the single-split precision of 1.00 (Table 5) as within-sample variance rather than as evidence of the artefact identified in the earlier baseline. We recommend that any future comparison with this work use the cross-validated mean and standard deviation rather than the single-split figures.

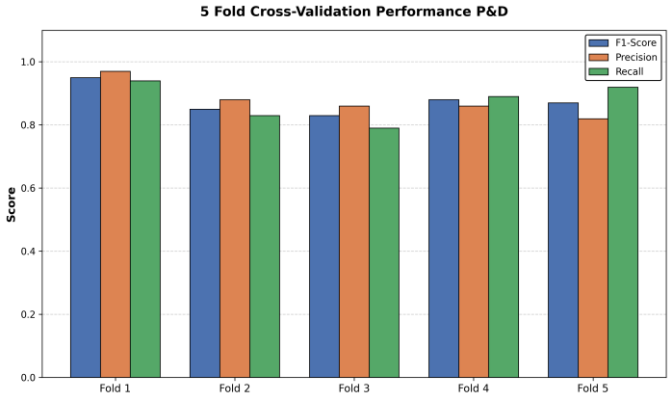

Fig. 7. Per-fold precision, recall and F1-score across the 5-fold stratified cross-validation (dashed line: mean F1-score).

D. Ablation Study

Table 6 report the five-configuration ablation study described in Section III-J, in which the TabNet backbone, hyperparameters and evaluation split are held fixed with the augmentation and loss strategy. The class-weighted baseline achieves F1-score of 0.860 (0.961 precision, 0.778 recall). SMOTE, generating the same 30,000-row volume as the WGAN-GP configuration, achieved F1 of 0.867 (0.980 precision, 0.778 recall). WGAN-GP synthesis alone, without hard-negative mining, improves on the augmented configuration at F1 of 0.879 (precision of 0.962, recall of 0.810). Adding hard-negative further improves on the WGAN-GP synthesis, reaching F1 of 0.885 (precision of 0.915, recall 0.857). The full proposed pipeline, adding focal loss on top of WGAN-GP and hard-negative mining, reaches F1 of 0.923 (precision of 1.00 recall of 0.857) — an equal recall on WGAN-GP with hard-negative mining alone.

Drawing any conclusions from Table 6, we apply directly, rather than set aside, the sensitivity already established in Section IV-C. The 63 P&D instances in the held-out test set mean that the recall figures in Table 6 correspond to small, and in two cases identical, true-positive counts: the class-weighted baseline and SMOTE both correctly identify 49 of the 63 P&D instances; WGAN-GP alone identifies 51; and WGAN-GP with hard-negative mining and the full proposed pipeline both identify 54. The comparison between WGAN-GP alone and SMOTE, which the first pattern below rests on, is therefore a margin of two true positives on a single 63-instance split — the same order of magnitude that Section IV-C already identified as producing a shift of roughly 1.5 percentage points in recall from a single reclassified instance.

Three directional patterns are observed. First, on this single split, WGAN-GP-generated synthetic data numerically outperforms SMOTE: generating the same 30,000-row volume, WGAN-GP synthesis alone reaches an F1-score of 0.879, two true positives ahead of SMOTE's 0.867 and the class-weighted baseline's 0.860 — consistent with, though not proof of, the generative approach holding value beyond SMOTE's interpolation on this dataset.

Second, on this split, hard-negative mining appears to function primarily as a recall enhancer rather than the main driver of overall performance recovery: adding it to the WGAN-GP configuration lifts the true-positive count from 51 to 54 of 63 (F1 to 0.885), at the cost of reducing precision from 0.962 to 0.915.

Third, on this split, focal loss does not shift the operating point toward higher recall at the cost of precision; the full proposed pipeline instead reaches precision 1.00 while holding recall unchanged at 54 of 63 true positives, yielding the highest aggregate F1-score observed in this study (0.923) — reported descriptively, for the reasons given above, rather than as a confirmed ranking over the other four configurations. An alpha-balanced focal loss with $\alpha = 0.9$ assigned to the minority class is designed to up-weight minority examples and down-weight easy majority examples, and its standard expected effect is to raise recall, not precision. Section III-F shows that the combined effect of augmentation and α-weighting leaves the minority and majority classes close to parity in total loss contribution rather than heavily favouring the minority class. We consider the more likely mechanism to be an interaction with the post-hoc decision-threshold tuning step (Section III-I): the $\gamma = 3.0$ focusing term reshapes the predicted-probability distribution by suppressing the contribution of easy examples, and if this leaves the validation-set precision–recall curve steeper near the operating point, the threshold that maximises F1 on that curve could shift to a stricter value that removes marginal false positives without removing true positives, which would mechanically raise precision at fixed recall.

TABLE 6. ABLATION STUDY ON THE HELD-OUT TEST SPLIT (P&D CLASS, 25-SECOND CHUNKS)

Configuration (TabNet backbone identical throughout)	Precision	Recall	F1-score
Baseline (class-weighted TabNet, no augmentation)	0.961	0.778	0.860
SMOTE	0.980	0.778	0.867
WGAN-GP only	0.962	0.810	0.879
WGAN-GP + hard-negative mining (no focal loss)	0.915	0.857	0.885
Full proposed model (WGAN-GP + hard mining + focal loss)	1.000	0.857	0.923

E. Multi-time Resolution Evaluation (5-Second and 15-Second Chunks)

Table 7 reports the multi-resolution evaluation specified in Section III-H: The full proposed pipeline, with hyperparameters fixed at their 25-second-tuned values, applied without re-tuning to the 15-second and 5-second aggregations of the same corpus. F1-score degrades monotonically as chunk size decreases (0.875 at 25 s, 0.840 at 15 s, 0.766 at 5 s), driven primarily by recall, which falls at every step (0.874, 0.792, 0.731). Precision does not follow the same monotonic pattern: it rises from 0.878 at 25 s to 0.896 at 15 s before falling to 0.812 at 5 s. Overall, recall declines more steeply than precision as resolution becomes finer, a trend that aligns with prior studies [4], [11]. An exception is Chadalapaka et al.'s Anomaly Transformer [23], which attains its highest recall scores at 15 s rather than at the coarsest resolution. Differences in evaluation protocols such as the absence of cross-validation for hyperparameter selection, could plausibly explain this divergence.

TABLE 7. MULTI-RESOLUTION EVALUATION: FULL PROPOSED PIPELINE WITH 25-SECOND-TUNED HYPERPARAMETERS APPLIED, WITHOUT RE-TUNING, TO 15- AND 5-SECOND CHUNKS (5-FOLD CV, MEAN ± SD)

Chunk size (s)	Precision	Recall	F1-score
25 (tuned, primary)	0.878 ± 0.055	0.874 ± 0.061	0.875 ± 0.047
15 (untuned, 25-s hyperparameters)	0.896 ± 0.041	0.792 ± 0.056	0.840 ± 0.023
5 (untuned, 25-s hyperparameters)	0.812 ± 0.126	0.731 ± 0.055	0.766 ± 0.072

We draw one clear methodological conclusion from this comparison: resolution-specific hyperparameter tuning, not

merely the choice of augmentation strategy, is necessary for competitive performance at finer transaction-chunk granularities. The 30,000-row synthetic volume, the thirty-fold hard-mining ratio, and the TabNet capacity selected for 25-second data do not transfer favourably to 5- or 15-second data, most plausibly because the minority class becomes both proportionally sparser and individually noisier at finer resolutions, which is likely to affect the stability of WGAN-GP training specifically, given that the generator already operates at the edge of feasibility on 216 real examples at 25-second resolution. We report this result as a genuine, informative limitation of the present configuration.

F. Comparative Analysis with State-of-the-Art

Table 8 situates the proposed pipeline against seven prior published results on the La Morgia Binance/Telegram corpus or closely related cryptocurrency P&D datasets. On the primary 25-second resolution, the proposed model attains a single-split F1-score of 92.3% (precision 100.0%, recall 85.7%), numerically ahead of the SmoteGAN–TabNet configuration of Abdulrazaq and Musa [32] (F1 90.0%), the C-LSTM of Chadalapaka et al. [23] (F1 89.3%), and the Anomaly Transformer of the same authors (F1 89.2%) under a comparable 80/20 protocol. However, the cross-validated mean F1 of 87.5% (precision 87.8%, recall 87.4%) is lower than the 5-fold CV Random Forest (RF) results of La Morgia et al. [4] (F1 94.5%) and La Morgia et al. [11] (F1 91.8%). At finer temporal granularities, where the 25-second-tuned hyperparameters are applied without re-tuning, the proposed pipeline yields F1-scores of 84.0% (15 s) and 76.6% (5 s), both below the resolution-tuned RF results of [4], [11] and the deep-learning comparators of [23] at the corresponding resolutions.

TABLE 8. COMPARATIVE ANALYSIS AGAINST PRIOR PUBLISHED STATE-OF-THE-ART ON THE LA MORGIA BINANCE/TELEGRAM CORPUS

Study	Method	Validation	Chunk Size	Precision (%)	Recall (%)	F1 (%)
Kamps & Kleinberg [2] (strict)	Statistical thresholding	–	1 hr	50.1	75.0	60.5
Victor & Hagemann [6]	XGBoost (AUC 0.995)	10-fold CV	30 min	–	85.5	–
Fantazzini & Xiao [7]	Threshold-adjusted (2.6%)	5-fold CV	30 s	–	98.0	-
Shao [22]	RF bagging/stacking (Dogecoin only)	5-fold CV	-	94.4	77.7	84.0
La Morgia et al. [11]	Random Forest	5-fold CV	5 15 25 s	92.4 91.3 93.7	78.4 84.4 91.3	84.0 87.7 91.8
La Morgia et al. [4]	Random Forest	5-fold CV	5 15 25 s	94.6 96.4 98.2	72.9 84.9 91.2	82.4 90.0 94.5
Chadalapaka et al. [23]	C-LSTM	80/20 split	25 s	94.2	85.0	89.3
Chadalapaka et al. [23]	Anomaly Transformer	80/20 split	25 s	88.4	90.0	89.2
Abdulrazaq & Musa [32]	SmoteGAN + TabNet	80/20 split	25 s	98.0	83.0	90.0
This study	Full proposed model	80/20 split	25 s	100.0	85.7	92.3
This study (primary)	WGAN-GP + HM + Focal TabNet	5-fold CV	25 s	87.8	87.4	87.5
This study (untuned)	WGAN-GP + HM + Focal TabNet	5-fold CV	15 s	89.6	79.2	84.0
This study (untuned)	WGAN-GP + HM + Focal TabNet	5-fold CV	5 s	81.2	73.1	76.6

The implications of this comparative profile are threefold. First, the near-ideal precisom of the entire model under consideration at 25-second resolution (100.0% single-split, 87.8% cross-validated) implies that a properly constructed WGAN-GP, along with the usage of hard negatives mining and alpha-balanced focal loss, can minimize false alarm rate much better than simply applying statistical thresholding methods [2], [7], and unweighted augmented generative technologies [32]. This aspect is considered to be important in terms of the possibilities of implementing market monitoring systems in areas where the expenses associated with false P&D alerts exceeded by the expenses needed to deal with a delayed genuine positive case.

Second, F1 score of 87.5%, although worse than that of the best RF models, should be viewed in connection with the protocol specifics: the five-fold cross-validation results mentioned above were accomplished with the WGAN-GP framework being fitted anew for each fold based on the training dataset only from that particular fold thus preventing the potential loss of synthetic-to-validation leakage characteristic of all methodologies that would introduce global augmentation prior to the fold splitting process. The implication is that the marginal contribution of leakage-controlled generative augmentation to F1 is bounded, and that for tabular P&D detection at 25-second resolution a well-tuned ensemble such as RF remains a strong baseline — a finding that tempers the prevailing assumption in the GAN-augmentation literature [30], [31] that a more expressive generator necessarily yields a better downstream classifier.

Third, the systematic degradation observed at 15-second and 5-second granularities (ΔF1 of -3.5 and -10.9 percentage points relative to the 25-second cross-validated mean) quantifies, for the first time in a GAN-augmented P&D study, the cost of applying a single tuned configuration across resolutions, and complements the resolution-tuned observations of [4], [11], [23] by isolating augmentation-volume and hard-mining-ratio transferability as a distinct failure mode.

G. Limitations

We discuss five threats to the validity of this study.

First, all evaluation in this study, including the cross-validation reported in Section IV-C and the multi-resolution evaluation of Section IV-E, is retrospective, drawn from historical transaction data collected around confirmed pump events; no claim is made about real-time detection latency or about performance under live, streaming conditions, where feature values are necessarily computed on incomplete windows.

Second, and central to the new contribution of this study, the multi-resolution evaluation of Section IV-E deliberately applies a single, 25-second-tuned hyperparameter configuration to the 15-second and 5-second data, so the degraded performance reported at finer resolutions should be read as evidence that resolution-specific re-tuning is necessary, not as evidence that a well-tuned pipeline could not achieve competitive performance at those resolutions; the resolution-tuned results of Chadalapaka et al. [23] reported in Table 8 indicate that competitive performance at finer resolutions is achievable in principle.

Third, the ablation comparison in Section IV-D and the single-split figures reported alongside the cross-validated means in Table 8 are decided on the same 63-instance held-out split whose sensitivity is quantified in Section IV-C. As Section IV-D now states explicitly, several of the pairwise differences among configurations correspond to a margin of only one to three true positives, and were not evaluated with any formal statistical test. These comparisons should be read as descriptive of a single split rather than as statistically established rankings; a cross-validated replication of the ablation, together with a paired significance test (for example, McNemar's test or a bootstrap over the held-out predictions) across folds, is necessary before any configuration in Table 6 can be confidently identified as superior to another.

Fourth, Algorithm 1 sets a single random seed for all the operations during the cross-validation process. Because of this, the variability discussed throughout this paper (like the fold-to-fold spread seen in Table 6 and in Section IV-C) reflects only the fold-to-fold variability without taking into account any initialization variability introduced by either of the networks. Performing the entire pipeline multiple times based on random seeds per fold (for instance, 3 seeds × 5 folds) would provide a more thorough description of all results in this paper and would reveal the variability attributed to random seeds and folds. However, we did not perform this experiment for this submission.

Fifth, the present study presents no information regarding training time, hardware specifications, or inference latency for any of the setups. This is a nonnegligible limitation of this paper since it has positioned the method for market surveillance applications where retraining the WGAN GP generator per fold leads to a cost and the latency of scoring a piece of transaction data at the time of inference is an important practical aspect. Reliable wall-clock training times, peak memory usage, and per-chunk inference latencies were not systematically recorded during the experimental runs, and we refrain from post-hoc estimation. Establishing these quantities is a prerequisite for any future claim regarding the method's suitability for real-time or near-real-time surveillance and is left as important future work.

V. CONCLUSION

This study presented an augmentation-leakage-controlled framework for detecting pump-and-dump schemes in cryptocurrency markets, addressing the methodological gaps in [32]. On the modelling side, the previously unspecified generative component was replaced with a fully documented WGAN-GP architecture, the augmentation volume was reduced from an unprecedented 6,120:1 synthetic-to-real ratio to a documented, moderate 139:1 ratio combined with a bounded hard-negative-mining step, and the classifier's loss function was changed to an alpha-balanced focal loss. On the evaluation side, the suspicious perfect-precision baseline was replaced with a properly class-weighted comparator; single-split evaluation was supplemented with 5-fold stratified cross-

validation; a five-configuration ablation study isolated the marginal contribution of each design choice; and the pipeline was evaluated at all three published transaction-chunk resolutions rather than only the one at which it was tuned.

Two patterns stand out, both reported with the same caveat stated in Section IV-D and Section IV-G: they are observed on a single 63-instance held-out split, with margins as small as one to three true positives, and were not subjected to a formal statistical test. First, on this split, WGAN-GP-generated synthetic data used alone numerically outperforms the substantially simpler SMOTE baseline, and most of the further recall improvement in the full pipeline comes from hard-negative mining rather than from the fidelity of the generative model alone — a directionally useful, but not yet statistically confirmed, observation for other severely imbalanced, small-minority tabular cybersecurity problems, where the temptation to reach for a more sophisticated generative model may not be justified without a comparable ablation. We identify a cross-validated replication of this ablation as a precondition for treating the ranking among configurations as established. Second, the multi-resolution evaluation shows that a pipeline tuned at one transaction-chunk resolution does not transfer favourably to others: applying the 25-second-tuned configuration without modification to 15- and 5-second data yields decreasing F1-scores, below the corresponding resolution-tuned results of La Morgia et al. [4],[11], indicating that resolution-specific re-tuning is a precondition for competitive performance at finer granularities.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support and contributions of all individuals who assisted in the successful completion of this research.

DATA AVAILABILITY

The transaction corpus used in this study is the publicly released La Morgia et al. Binance/Telegram pump-and-dump dataset [12], available via the SystemsLab-Sapienza GitHub repository. No personally identifiable trading or account data are used or redistributed by the authors. Implementation code for the WGAN-GP generator, the hard-negative-mining and focal-loss components, and the Focal-TabNet classifier is available from the corresponding author upon reasonable request.

AUTHOR CONTRIBUTIONS

Muhammad Nazeer Musa conceived the study, designed the overall augmentation-leakage-controlled framework, implemented the WGAN-GP and Focal-TabNet pipeline. Umar Faruq Abdulrazaq contributed to the literature review, data preprocessing, and critical revision of the manuscript. Tijjani Abdullahi contributed to the experimental design, the multi-resolution evaluation. All authors read and approved the final manuscript.

AI TOOL USAGE ACKNOWLEDGMENT

The authors acknowledge the use of AI tools for drafting assistance, and language polishing during the preparation of this manuscript. All AI-assisted content was reviewed and verified by the authors who take full responsibility for the final manuscript.

REFERENCES

[1] S. Foley, J. R. Karlsen, and T. J. Putniņš, "Sex, drugs, and bitcoin: How much illegal activity is financed through cryptocurrencies?," *The Review of Financial Studies*, vol. 32, pp. 1798–1853, 2019.

[2] J. Kamps and B. Kleinberg, "To the moon: defining and detecting cryptocurrency pump-and-dumps," *Crime Science*, vol. 7, pp. 1–18, 2018.

[3] J. T. Hamrick, F. Rouhi, A. Mukherjee, M. Vasek, T. Moore, and N. Gandal, "Analyzing target-based cryptocurrency pump and dump schemes," in *Proc. 2021 ACM CCS Workshop on Decentralized Finance and Security*, 2021, pp. 21–27.

[4] M. La Morgia, A. Mei, F. Sassi, and J. Stefa, "The doge of wall street: Analysis and detection of pump and dump cryptocurrency manipulations," *ACM Transactions on Internet Technology*, vol. 23, pp. 1–28, 2023.

[5] A. Dhawan and T. J. Putniņš, "A new wolf in town? Pump-and-dump manipulation in cryptocurrency markets," *Review of Finance*, vol. 27, pp. 935–975, 2023.

[6] F. Victor and T. Hagemann, "Cryptocurrency pump and dump schemes: Quantification and detection," in *Proc. 2019 International Conference on Data Mining Workshops (ICDMW)*, 2019, pp. 244–251.

[7] D. Fantazzini and Y. Xiao, "Detecting pump-and-dumps with crypto-assets: Dealing with imbalanced datasets and insiders' anticipated purchases," *Econometrics*, vol. 11, p. 22, 2023.

[8] Y. Li, Z. Wang, H. Wang, M. Wu, and L. Xie, "Identifying price bubble periods in the Bitcoin market-based on GSADF model," *Quality & Quantity*, pp. 1–16, 2021.

[9] I. Goodfellow et al., "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[10] S. O. Arik and T. Pfister, "TabNet: Attentive interpretable tabular learning," *arXiv preprint arXiv:1908.07442*, 2019.

[11] M. La Morgia, A. Mei, F. Sassi, and J. Stefa, "Pump and dumps in the bitcoin era: Real time detection of cryptocurrency market manipulations," in *Proc. 2020 29th International Conference on Computer Communications and Networks (ICCCN)*, 2020, pp. 1–9.

[12] SystemsLab-Sapienza, "Pump and dump dataset," *GitHub*, 2021. [Online]. Available: https://github.com/SystemsLab-Sapienza/pump-and-dump-dataset

[13] R. K. Aggarwal and G. Wu, "Stock market manipulations," *The Journal of Business*, vol. 79, pp. 1915–1953, 2006.

[14] M. J. Rajaei and Q. H. Mahmoud, "A survey on pump and dump detection in the cryptocurrency market using machine learning," *Future Internet*, vol. 15, p. 267, 2023.

[15] N. Gandal, J. T. Hamrick, T. Moore, and T. Oberman, "Price manipulation in the Bitcoin ecosystem," *Journal of Monetary Economics*, vol. 95, pp. 86–96, 2018.

[16] R. Mitchell and R. Chen, "Behavior-rule based intrusion detection systems for safety critical smart grid applications," *IEEE Transactions on Smart Grid*, vol. 4, pp. 1254–1263, 2013.

[17] P. Priyadarshi and P. Kumar, "A comprehensive review on insider trading detection using artificial intelligence," *Journal of Computational Social Science*, pp. 1–20, 2024.

[18] Z. Chen, S. Wang, D. Yan, and Y. Li, "Research and implementation of bank credit card fraud detection system based on reinforcement learning and LSTM," in *Proc. 2023 3rd International Conference on Mobile Networks and Wireless Communications (ICMNWC)*, 2023, pp. 1–8.

[19] J. Xu and B. Livshits, "The anatomy of a cryptocurrency pump-and-dump scheme," in *Proc. USENIX Security Symposium*, Santa Clara, CA, USA, Aug. 2019, pp. 1609–1625.

[20] T. M. Mitchell, *Machine Learning*, 1st ed. New York, NY, USA: McGraw-Hill, 1997.

[21] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY, USA: Springer, 2009.

[22] S. Shao, "The effectiveness of supervised learning models in detection of pump and dump activity in Dogecoin," in *Proc.*

Second IYSF Academic Symposium on Artificial Intelligence and Computer Engineering, 2021, pp. 356–363.
[23] V. Chadalapaka, K. Chang, G. Mahajan, and A. Vasil, "Crypto pump and dump detection via deep learning techniques," *arXiv preprint arXiv:2205.04646*, 2022.
[24] D. Alfano, R. Abbruzzese, and D. Parente, "Pump and dump cryptocurrency detection using social media," in *Proc. International Conference on Data (DATA)*, 2023, pp. 235–240.
[25] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
[26] G. Menardi and N. Torelli, "Training and assessing classification rules with imbalanced data," *Data Mining and Knowledge Discovery*, vol. 28, pp. 92–122, 2014.
[27] C. Tantithamthavorn, A. E. Hassan, and K. Matsumoto, "The impact of class rebalancing techniques on the performance and interpretation of defect prediction models," *IEEE Transactions on Software Engineering*, vol. 46, pp. 1200–1219, 2018.
[28] T. Wongvorachan, S. He, and O. Bulut, "A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining," *Information*, vol. 14, p. 54, 2023.
[29] H. Nghiem, G. Muric, F. Morstatter, and E. Ferrara, "Detecting cryptocurrency pump-and-dump frauds using market and social signals," *Expert Systems with Applications*, vol. 182, p. 115284, 2021.
[30] R. Sauber-Cole and T. M. Khoshgoftaar, "The use of generative adversarial networks to alleviate class imbalance in tabular data: a survey," *Journal of Big Data*, vol. 9, p. 98, 2022.
[31] A. Sharma, P. K. Singh, and R. Chandra, "SMOTified-GAN for class imbalanced pattern classification problems," *IEEE Access*, vol. 10, pp. 30655–30665, 2022.
[32] U. Abdulrazaq, and M. N. Musa, "SmoteGAN and TabNet: a hybrid framework for detecting pump-and-dump schemes in cryptocurrency markets," *Proceedings of the Nigerian Society of Physical Sciences*, vol. 3, pp. 257-257, 2026
[33] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in Advances in Neural Information Processing Systems, vol. 30, 2017.
[34] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proc. IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988.

Muhammad Nazeer Musa received the B.Tech. degree in computer science from the Federal University of Technology, Minna, Niger State, Nigeria, and the M.Sc. degree in computer science from the Nigerian Defence Academy, Kaduna, Nigeria, in 2020. He is currently pursuing the Ph.D. degree in computer science, with a focus on cybersecurity. His research focuses on fairness-aware deepfake detection.

He is a Lecturer and Research Assistant in the Department of Cyber Security, Nigerian Defence Academy, Kaduna, Nigeria. He has authored peer-reviewed publications and contributed to research projects supported by the Tertiary Education Trust Fund (TETFund). He is a a member of the International Society for Data Science and Analytics, the International Association of Engineers, Cyber Security Expert Association of Nigeria, and the Internet Society Nigeria.

Umar Faruq Abdulrazaq is currently pursuing a PhD in Cybersecurity at Nasarawa State University Keffi, Nigeria. He holds a Bachelor of Engineering degree in Computer Engineering from University of Maiduguri, Nigeria in 2014 and master's degree in Financial Technology at University of Stirling, Scotland. He is a registered Engineer in Nigeria with council for the registration of Engineering in Nigeria (COREN). His current research is on Hybrid Stacking Ensemble Framework for Memory-based malware detection and type Identification using SHAP and LIME.

Tijjani Abdullahi is a software engineer, data scientist, and Lecturer in Computer Science at Baze University, Abuja, Nigeria. He holds a Bachelor's degree (B.Sc.) and a Master's degree (M.Sc.) in Computer Science and is currently pursuing a Ph.D. in Computer Science at the University of Abuja. Originally from Kaduna State, Nigeria, his research interests include artificial intelligence, machine learning, software engineering, data science, and explainable AI. His work focuses on developing intelligent, data-driven solutions to real-world problems.

.