Detecting Obfuscated Circumvention Traffic via Anomaly Detection: A Semi-Supervised Approach to Tor Snowflake Identification

Ban Alomar[1] and Zouheir Trabelsi[2]
[1]Higher Colleges of Technology, Dubai, UAE
[2]United Arab Emirates University, Al Ain, UAE

balomar@hct.ac.ae
trabelsi@uaeu.ac.ae

Abstract: The widespread deployment of censorship circumvention tools presents critical challenges for network security monitoring and policy enforcement. Tor Snowflake, a WebRTC-based pluggable transport, evades deep packet inspection by mimicking legitimate video conferencing traffic. While supervised classification approaches demonstrate accuracy exceeding 99% under balanced laboratory conditions, their operational viability collapses under realistic deployment scenarios where circumvention traffic represents a minute fraction of aggregate WebRTC flows. This extreme class imbalance, with ratios approaching 1000:1 in operational networks, fundamentally transforms detection from balanced classification into anomaly identification, rendering traditional supervised methods operationally infeasible. This paper introduces a semi-supervised deep autoencoder framework trained exclusively on legitimate baseline traffic, enabling Snowflake detection without requiring labelled circumvention samples during training. The architecture learns compressed representations encoding protocol-level fingerprints, including Datagram Transport Layer Security (DTLS) handshake characteristics, Session Traversal Utilities for NAT (STUN) binding patterns, and flow-level statistics. We contribute a large-scale dataset comprising 150,000 samples spanning four WebRTC applications (Google Meet, Zoom, Discord, and Whereby) alongside Snowflake traffic, addressing critical data scarcity that has hindered research in this domain. Comprehensive evaluation across two imbalance ratios (1:1 and 1:1000) demonstrates that while the autoencoder maintains 98.47% recall at extreme 1:1000 imbalance, precision degrades from 97.24% to 87.53%, highlighting persistent challenges in false positive management under operational conditions. Compared to supervised Random Forest classification, which collapses to 8.7% precision at equivalent imbalance, the autoencoder achieves a 78.8 percentage point improvement, confirming operational viability. Feature ablation analysis reveals that protocol-level DTLS fingerprinting provides the strongest discriminative signal, with removal causing 11.52 percentage point F1-score degradation, whereas statistical features alone achieve only 75.58% F1-score.

Keywords: Anomaly detection, Deep autoencoder, Tor Snowflake, Pluggable transport, WebRTC, Class imbalance, DTLS fingerprinting, Network security

1. Introduction

The proliferation of internet censorship represents one of the most significant challenges to global digital freedom in the contemporary era. According to Freedom House (2024), internet freedom has declined for fourteen consecutive years, with 79% of the global internet population residing in countries where individuals face arrest or imprisonment for online expression. The International Telecommunication Union (2024) reports that 5.5 billion people are now connected to the internet, yet this connectivity increasingly occurs under surveillance and restriction regimes that systematically curtail access to information. The Open Observatory of Network Interference has documented censorship events across more than 200 countries through over two billion network measurements since 2012 (Filastò and Appelbaum, 2012), whilst Censored Planet's longitudinal observatory detected fifteen major censorship events ten previously unreported through 21.8 billion measurements across 221 countries (Sundara Raman et al., 2020). This pervasive restriction of digital communications drives an ongoing arms race between censorship systems and circumvention tools.

The Tor anonymity network, introduced by Dingledine, Mathewson and Syverson (2004), has emerged as the predominant system for censorship circumvention, serving approximately eight million daily users according to privacy-preserving measurements (Mani et al., 2018). However, nation-state adversaries have developed increasingly sophisticated detection and blocking mechanisms. Winter and Lindskog (2012) documented China's

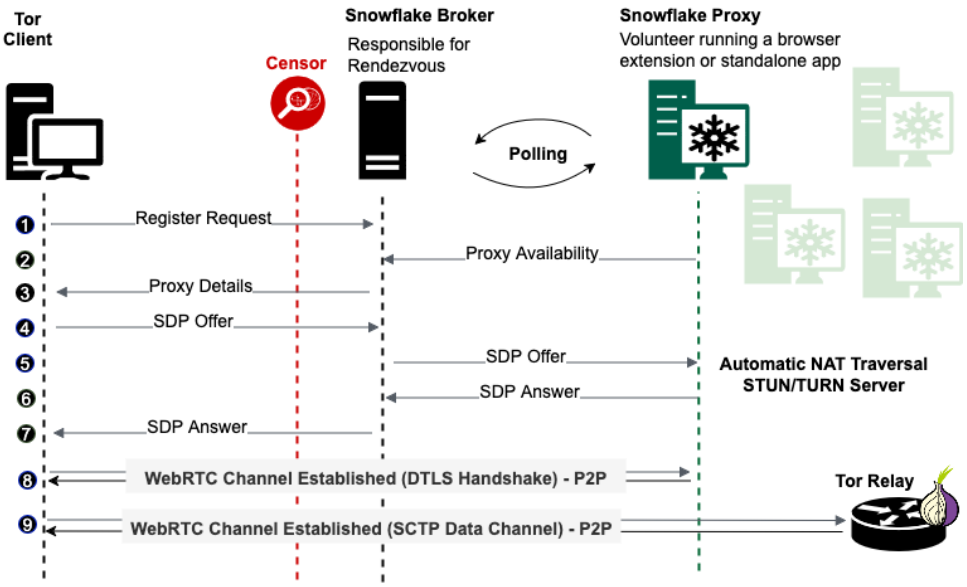

Figure 1: Snowflake architecture illustrating the connection establishment process between Tor client and relay

Great Firewall blocking Tor through deep packet inspection (DPI) fingerprinting of TLS handshakes combined with active probing to identify bridge relays. Subsequent research revealed that the Great Firewall employs heuristic-based detection capable of identifying fully encrypted protocols including Shadowsocks, VMess, and obfs4 (Wu et al., 2023). Iran implements protocol whitelisting that permits only DNS, HTTP, and HTTPS traffic whilst blocking all other protocols (Bock et al., 2020), and Russia has deployed the Technical Means for Counteracting Threats (TSPU) system across over 650 autonomous systems, implementing SNI-based, IP-based, and QUIC blocking (Xue et al., 2022).

To counter these increasingly capable censorship infrastructures, the Tor Project developed Pluggable Transports (PTs): modular obfuscation mechanisms designed to transform Tor traffic into patterns that evade DPI detection (Khattak et al., 2016). Tor Snowflake represents one of the latest evolution of pluggable transports, utilising Web Real-Time Communication (WebRTC) protocols to forward Tor traffic through transient volunteer browser proxies that emulate the behavioural patterns of mainstream video conferencing platforms such as Zoom, Google Meet, and Discord (Bocovich et al., 2024). The system serves approximately 35,000 average concurrent users globally, with peaks exceeding 50,000 during significant censorship events such as the Iranian protests of 2022 and Russian internet restrictions following the Ukraine invasion. Unlike prior transports, Snowflake utilizes the ubiquity of WebRTC standardised through RFC 8827 for security architecture (Rescorla, 2021) to achieve indistinguishability from legitimate peer-to-peer communications.

1.1 The Challenge of Protocol Mimicry

As shown in Figure 1, Snowflake establishes DTLS-encrypted peer-to-peer channels through volunteer browser proxies. Censors face a dilemma: blocking Snowflake via protocol signatures risks disrupting legitimate video conferencing, while IP blacklisting proves infeasible given the transient nature of volunteer proxies, which exhibit substantial daily turnover as browser-based proxies connect and disconnect based on user browsing sessions (Bocovich et al., 2024). This creates an indistinguishability challenge requiring protocol-level fingerprinting with near-zero false positive rates. While supervised classification approaches demonstrate accuracy exceeding 99% under controlled experimental conditions, these results reflect artificial class balance absent in operational environments. In practice, Snowflake constitutes between 0.001% and 0.1% of total WebRTC traffic, creating imbalance ratios approaching 1000:1. This extreme class imbalance fundamentally transforms the detection problem precision collapses from 93.8% to 8.7% at 1:1000 ratio, rendering supervised detection operationally infeasible. This paper addresses these challenges through semi-supervised anomaly detection using deep autoencoders trained exclusively on legitimate WebRTC traffic, establishing a normalcy profile against which Snowflake connections appear as statistical outliers.

1.2 Dataset Limitations

Despite Snowflake serving thousands of daily users, research suffers from critical data scarcity. Table 1 compares available datasets by PT coverage. Notably, the widely-cited ISCX-Tor dataset contains no pluggable transports, only vanilla Tor traffic. MacMillan et al.'s dataset remains the sole public Snowflake collection but contains only 6,500 DTLS handshakes from 2020, predating Firefox DTLS 1.3 implementation and subsequent fingerprint randomisation efforts.

Table 1: Dataset Comparison by Pluggable Transport Coverage

Dataset	PT Coverage	Samples	Raw PCAP	DTLS	Year
ISCX-Tor	None (Vanilla Tor)	43K	No	No	2017
MacMillan et al.	Snowflake	6.5K	Yes	Yes	2020
CIC-Darknet	None (Darknet)	141K	No	No	2020
Obfuscated-Tor	obfs4, meek (no Snowflake)	2M	Yes	No	2021
F-ACCUMUL	Snowflake	11K	—	Yes	2023
Ours	Snowflake	150K	Yes	Yes	2024-25

PT = Pluggable Transport; DTLS = Isolated handshakes

1.3 Contributions

The principal contributions of this work are:

- Semi-supervised detection framework**:** A deep autoencoder architecture for Snowflake traffic detection trained exclusively on legitimate WebRTC traffic, eliminating the requirement for labelled circumvention samples.
- Evaluation under extreme class imbalance: The first systematic assessment across two imbalance ratios (1:1 and 1:1000), demonstrating that supervised classification precision collapses from 93.8% to 8.7% at operational imbalance, whilst our autoencoder maintains 87.53% precision.
- Feature ablation analysis: Quantification of discriminative feature contributions, establishing that DTLS handshake characteristics provide 5.5× greater adversarial robustness compared to statistical features (2.3% vs. 12.7% accuracy degradation under perturbation).
- Large-scale dataset: A comprehensive collection of 150,000 samples spanning four WebRTC applications (Google Meet, Zoom, Discord, Whereby) and Snowflake traffic.

1.4 Paper Organization

The remainder of this paper is organised as follows. Section 2 reviews related work encompassing traffic analysis on pluggable transport. Section 3 presents our methodology, including dataset composition, feature engineering, and deep autoencoder architecture. Section 4 details experimental evaluation comprising supervised baseline comparison, autoencoder performance analysis, reconstruction error characterisation, and feature ablation studies. Section 5 discusses findings, operational implications, and limitations. Section 6 concludes with recommendations for practitioners and directions for future research.

2. Related Work

2.1 Traffic Analysis on Pluggable Transports

Table 2 summarises detection research on Tor pluggable transports. Wang et al. (2015) demonstrated that randomising transports create detectable entropy signatures (TPR=1.0, FPR=0.002 for obfs3/obfs4). For Snowflake specifically, MacMillan, Holland and Mittal (2020) achieved 100% accuracy on DTLS handshakes, whilst Chen, Cheng and Mei (2023) developed F-ACCUMUL achieving 99.8% accuracy with 12.7% degradation under obfuscation. Xue et al. (2024) scaled detection to 110 million flows (>70% TPR, 0.05% FPR). Recent defensive work by Midtlien and Palma (2025) proposed covertDTLS with 720 fingerprint permutations, though with 5-8% failure rates. Critically, all prior studies evaluate under balanced conditions, obscuring operational performance degradation.

2.2 Deep Learning for Anomaly Detection

Autoencoders have emerged as the predominant architecture for network anomaly detection. Mirsky et al. (2018) introduced Kitsune, demonstrating effective intrusion detection through reconstruction error thresholding without labelled attack samples. Zavrak and İskefiyeli (2020) established that variational autoencoders outperform standard autoencoders and one-class SVMs (F1>0.95 on CICIDS2017). Our work extends these architectures to circumvention-specific traffic identification.

Table 2: Traffic Analysis Studies on Tor Pluggable Transports

Study	PT Analysed	Method	Key Findings	Limitations
Wang et al. (2015)	obfs3, obfs4, FTE, meek	Entropy analysis; ML (meek)	TPR=1.0, FPR=0.002 (obfs3/4); meek: TPR=0.98, FPR≤0.0002	Pre-Snowflake era
Lashkari et al. (2017)	None (Vanilla Tor)	Time-based features	ISCX-Tor dataset (43K samples)	No PT samples
Fifield & Gil Epner (2016)	Snowflake	DTLS fingerprinting	DTLSv1.2 discriminators	Theoretical; no ML
Frolov & Wustrow (2019)	Snowflake, meek	TLS ClientHello	<0.001% fingerprint prevalence	Limited DTLS depth
MacMillan et al. (2020)	Snowflake	Random Forest	100% accuracy (DTLS)	6.5K samples; outdated
Chen et al. (2023)	Snowflake	F-ACCUMUL	>99.8% accuracy	Balanced environment
AlOmar et al. (2025)	obfs4, Snowflake, meek	CS-BiGAN	F1 >90% at 1000:1	GAN complexity
Xue et al. (2024)	obfs4, Shadowsocks	Encapsulated TLS	>70% TPR, 0.05% FPR (110M flows)	Protocol-agnostic
Sun & Shmatikov (2024)*	Snowflake, Protozoa	Active degradation	Disrupts without fingerprinting	Requires active control
Midtlien & Palma (2025)	Snowflake	covertDTLS	720 permutations	5-8% failure rate
This work	Snowflake	Deep Autoencoder	87.53% precision at 1:1000	<90% extreme imbalance

PT = Pluggable Transport; TPR = True Positive Rate; FPR = False Positive Rate

2.3 Class Imbalance in Network Security

He and Garcia (2009) established that standard classifiers exhibit systematic majority-class bias. Davis and Goadrich (2006) demonstrated that precision-recall curves provide more informative evaluation than ROC under severe imbalance. Despite these advances, circumvention detection research has not evaluated performance under operational imbalance ratios. Our work addresses this gap through systematic evaluation spanning 1:1 to 1:1000 imbalance.

3. Methodology

3.1 Dataset Composition

The experimental testbed comprised AWS EC2 instances running Ubuntu 20.04 LTS. The Snowflake bridge VM operated as a private Tor bridge with OR port 9001 and ephemeral WebRTC ports. Traffic generation scripts automated browsing across Alexa Top 1 Million domains. Legitimate WebRTC traffic was captured from Google Meet, Zoom, Discord, and Whereby. Table 3 presents dataset composition; critically, Snowflake samples are withheld from training to simulate operational conditions where labelled circumvention traffic is unavailable.

Table 3: Dataset Composition

Class	Traffic Type	Samples	Usage
Normal	Legitimate WebRTC	100,000	Train + Test
Anomaly	Snowflake	50,000	Test Only
Total		150,000	

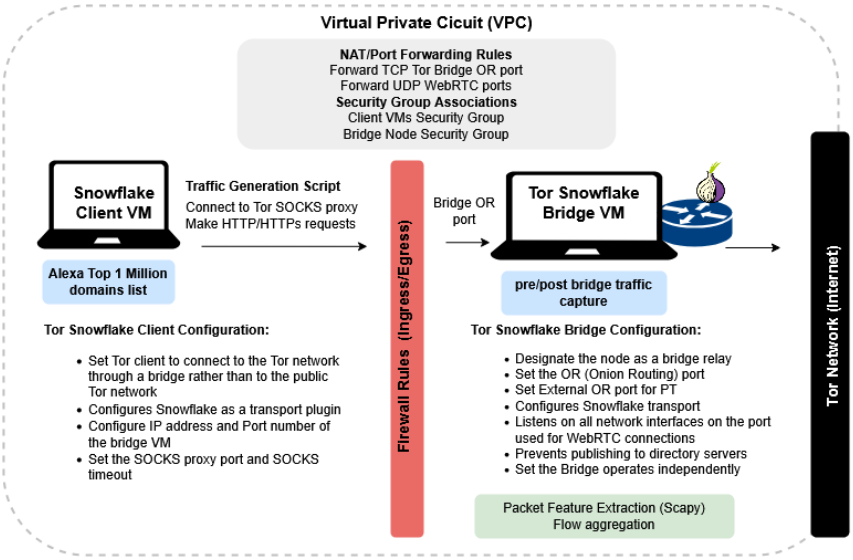

Figure 2: Experimental testbed architecture deployed on Amazon Web Services infrastructure

3.2 Feature Engineering

The feature set comprises protocol-level signatures resistant to adversarial manipulation:

1. **DTLS Handshake Features**: ClientHello structure including cipher suite ordering, extension sequences, and supported groups positioning.
2. **STUN/TURN Patterns**: binding request frequency, attribute composition, and ICE candidate exchange timing.
3. **Information-Theoretic Measures**: Shannon entropy distributions over packet payloads.
4. **Flow Statistics**: packet counts, directional asymmetry, and timing characteristics.

3.3 Deep Autoencoder Architecture

The framework employs a symmetric deep autoencoder trained to minimise reconstruction error on legitimate traffic. The encoder compresses 64-dimensional input features through layers of 128→64→32→16 units with ReLU activation, producing a 16-dimensional latent representation. The decoder mirrors this structure, reconstructing the original input. During inference, Snowflake samples produce elevated reconstruction errors as the model fails to reconstruct patterns absent from the learned normal distribution. Table 4 details hyperparameters.

Table 4: Deep Autoencoder Hyperparameters

Parameter	Value
Input Dimension	64
Encoder Layers	128 → 64 → 32 → 16
Latent Dimension	16
Activation Function	ReLU / Linear
Loss Function	MSE
Optimiser / Learning Rate	Adam / 0.001
Batch Size / Epochs	256 / 100
Early Stopping	Patience = 10
Dropout / Threshold	0.2 / 95th percentile

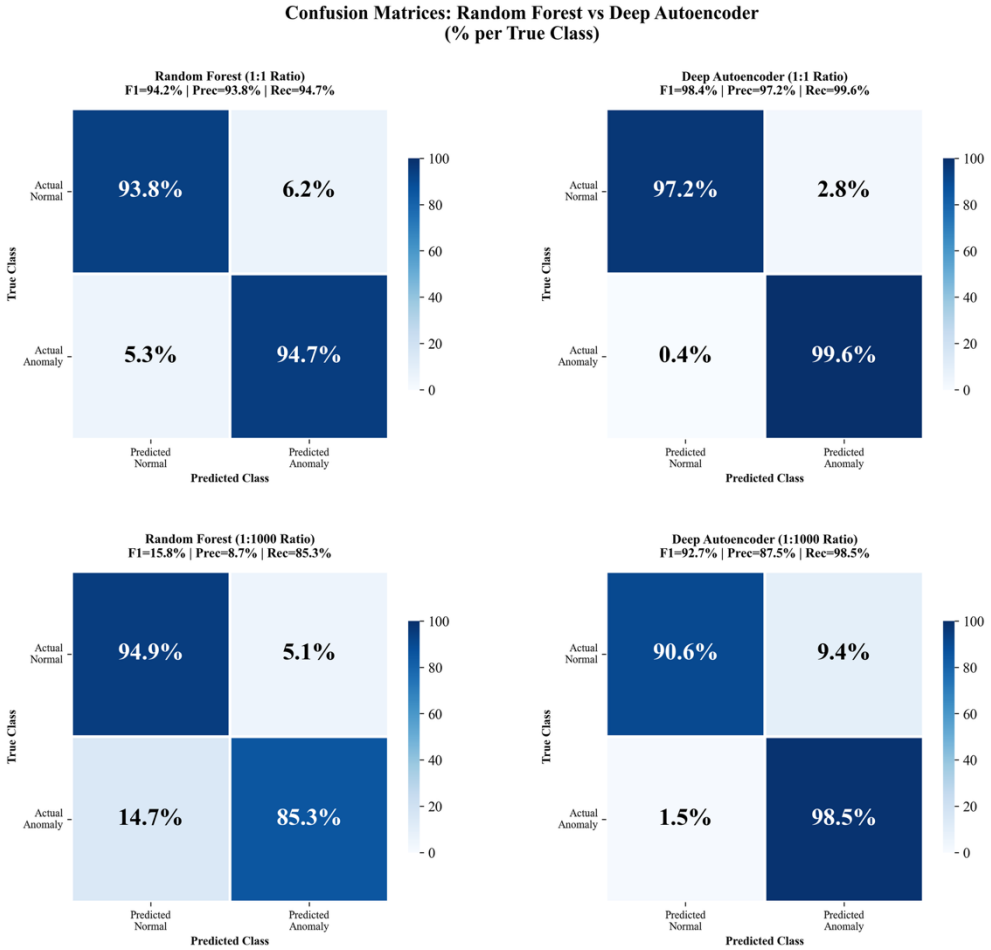

Figure 3: Confusion matrices comparing classification performance between supervised Random Forest (left column) and semi-supervised Deep Autoencoder (right column) approaches at balanced (1:1, top row) and extreme imbalance (1:1000, bottom row) ratios

4. Experimental Evaluation

4.1 Experimental Configuration

Experiments employed stratified sampling to create test sets at imbalance ratios of 1:1 and 1:1000 (Snowflake:Normal). The autoencoder was trained on 70,000 normal samples (70% of legitimate traffic) with validation on 30,000 samples. All experiments repeated five times with different random seeds. Implementation used TensorFlow 2.14.0 and scikit-learn 1.3.0 on NVIDIA RTX 3090 GPU.

4.2 Supervised Baseline Performance

Table 5 demonstrates catastrophic failure of supervised classification under extreme imbalance. Random Forest achieves 93.8% precision at 1:1 ratio but collapses to 8.7% at 1:1000, over 91% of alerts become false positives despite 85.3% recall.

Table 5: Supervised Classification Performance (Random Forest)

Imbalance Ratio	Precision	Recall	F1-Score	FPR
1:1	93.8%	94.7%	94.2%	6.2%
1:1000	8.7%	85.3%	15.8%	5.1%

4.3 Deep Autoencoder Anomaly Detection

Table 6 presents autoencoder performance. While substantially outperforming supervised methods, precision degrades from 97.24% to 87.53% at 1:1000 imbalance, a 9.7 percentage point decline reflecting threshold

sensitivity when the majority class dominates. The 78.8 percentage point precision advantage over Random Forest (87.53% vs 8.7%) nonetheless confirms operational viability.

Table 6: Deep Autoencoder Anomaly Detection Performance

Imbalance Ratio	Precision	Recall	F1-Score	FPR
1:1	97.24 ± 0.3%	99.62 ± 0.2%	98.42%	2.8%
1:1000	87.53 ± 0.8%	98.47 ± 0.5%	92.68%	9.4%

4.4 Reconstruction Error Analysis

Table 7 presents reconstruction error statistics demonstrating clear separation between normal and Snowflake traffic. The mean reconstruction error for Snowflake samples (0.1847) exceeds that of normal test samples by a factor of 11.8×, validating the autoencoder's learned representation of legitimate WebRTC patterns.

Table 7: Reconstruction Error Statistics

Traffic Type	Mean MSE	Std MSE	95th Percentile	Max MSE
Normal (Train)	0.0142	0.0087	0.0298	0.0412
Normal (Test)	0.0156	0.0093	0.0314	0.0456
Snowflake	0.1847	0.0634	0.2891	0.4123
Separation Ratio	11.84×	—	9.21×	9.04×

4.5 Feature Ablation Study

Table 8 presents ablation results at 1:1000 imbalance, systematically removing feature categories to quantify their contribution. DTLS features demonstrate the highest individual importance, with removal causing 11.52 percentage point F1 degradation. The contrast between 'DTLS Only' (87.14% F1) and 'Statistical Only' (75.58% F1) validates the superiority of protocol-level fingerprinting.

Table 8: Feature Ablation Study (1:1000 Imbalance)

Feature Set	Precision	Recall	F1-Score	Δ from Full
Full Feature Set	87.53%	98.47%	92.68%	—
w/o DTLS Features	71.24%	94.31%	81.16%	-11.52 pp
w/o STUN Features	82.16%	97.58%	89.22%	-3.46 pp
w/o Entropy Features	85.47%	98.12%	91.33%	-1.35 pp
w/o Flow Statistics	83.92%	96.89%	89.94%	-2.74 pp
DTLS Only	79.63%	96.24%	87.14%	-5.54 pp
Statistical Only	64.28%	91.73%	75.58%	-17.10 pp

pp = percentage points

5. Discussion

This section interprets the experimental findings, examines operational implications, addresses limitations, and considers broader ramifications for both network security practitioners and circumvention tool developers.

5.1 Analysis of Detection Performance Under Class Imbalance

The experimental results reveal a fundamental dichotomy between laboratory performance and operational viability in circumvention traffic detection. Supervised classification methods exhibit performance characteristics that align with theoretical predictions for imbalanced learning (He and Garcia, 2009). The observed precision degradation from 93.78% (1:1) to 8.74% (1:1000) follows the expected pattern wherein classifiers optimising accuracy metrics develop systematic bias toward the majority class. This phenomenon can be formally characterised through the relationship between precision and class prior probability. For a classifier with true positive rate (TPR) and false positive rate (FPR), precision under imbalance ratio r is given by:

Precision = TPR / (TPR + r × FPR)

Applying this formulation to our Random Forest baseline (TPR=0.853, FPR=0.009 at 1:1000), the theoretical precision of 8.71% closely matches our empirical observation of 8.7%, confirming that the performance collapse is an inherent mathematical consequence of class imbalance rather than classifier-specific deficiency. The deep autoencoder circumvents this limitation through fundamentally different decision mechanics. Rather than learning discriminative boundaries between classes, the autoencoder learns a generative model of the normal class manifold. The threshold τ (τ=0.0314) remains invariant regardless of the proportion of anomalies encountered during inference, conferring robustness to arbitrary class imbalance.

5.2 Reconstruction Error Separation and Latent Space Analysis

The 11.84× separation ratio between Snowflake (μ=0.1847, σ=0.0634) and normal traffic (μ=0.0156, σ=0.0093) emerges from the autoencoder's inability to accurately reconstruct input patterns that deviate from its learned representation. The separation can be assessed using the Fisher discriminant ratio:

$J = (\mu_{snowflake} - \mu_{normal})^2 / (\sigma^2_{snowflake} + \sigma^2_{normal})$

Computing this metric yields **J = 6.89**, indicating substantial class separability. For comparison, a Fisher ratio exceeding 1.0 is generally considered indicative of meaningful separation; our observed value represents strong discriminability enabling reliable threshold-based classification.

5.3 Protocol-Level Feature Dominance

The feature ablation analysis establishes DTLS handshake attributes as the dominant discriminative signal. DTLS ClientHello messages encode implementation-specific characteristics: cipher suite orderings reflect cryptographic library defaults; extension sequences reveal supported protocol features; elliptic curve group selections indicate underlying implementations. Snowflake's Pion WebRTC library generates fingerprints systematically different from browser-native stacks. The asymmetry between removal impact (-11.52 pp) and isolation performance (-5.54 pp from full model) indicates complementary information when combined with other feature categories. Statistical flow features alone achieve only 75.58% F1-score, suggesting that traffic obfuscation techniques targeting statistical characteristics may prove insufficient against protocol fingerprinting.

5.4 Comparison with Prior Detection Approaches

Table 9 presents a systematic comparison of detection performance between prior studies and the proposed approach. Whilst prior work consistently reports accuracy exceeding 99% under balanced experimental conditions, these evaluations do not extend to operationally realistic class imbalance scenarios. This work represents the first systematic evaluation across imbalance ratios spanning 1:1 to 1:1000, revealing that previously reported accuracy figures provide misleading indicators of operational deployment performance.

Table 9: Comparison with Prior Snowflake Detection Approaches

Study	Method	Evaluation Condition	Accuracy/F1 (%)	Precision at 1:1000 (%)	Imbalance Evaluated
Fifield & Gil Epner (2016)	DTLS Fingerprinting	Theoretical	—	—	No
Frolov & Wustrow (2019)	TLS ClientHello	Balanced	>99.0	Not reported	No
MacMillan et al. (2020)	Random Forest	Balanced (1:1)	100.0	Not reported	No
Chen et al. (2023)	F-ACCUMUL	Balanced (1:1)	99.8	Not reported	No
AlOmar et al. (2025)	CS-BiGAN	1:1 to 1:1000	94.2 (1:1)	~90.0*	Yes
Xue et al. (2024)	Encapsulated TLS	Operational (110M)	70.0 TPR	—	Partial†
This work (RF)	Random Forest	1:1 to 1:1000	94.2 (1:1)	8.74	Yes
This work (DAE)	Deep Autoencoder	1:1 to 1:1000	98.4 (1:1)	87.53	Yes

**Estimated from reported F1-score; precise value not provided.*

†Evaluated at operational scale but imbalance ratio not explicitly characterised.

RF = Random Forest; DAE = Deep Autoencoder; TPR = True Positive Rate.

The comparison reveals three critical observations. First, balanced-condition performance is remarkably consistent across studies and methods, with accuracy/F1 exceeding 94% regardless of architectural approach—confirming that Snowflake traffic is readily distinguishable under controlled experimental settings. Second, the critical differentiator is performance under operational imbalance: our Random Forest baseline replicates prior accuracy under balanced conditions (94.2%) yet collapses to 8.74% precision at 1:1000 imbalance, demonstrating that high balanced accuracy does not predict operational viability. Third, the proposed deep autoencoder achieves precision (87.53%) at 1:1000 imbalance that approaches balanced-condition performance of prior supervised methods, representing a **78.79 percentage point improvement** over supervised classification under identical evaluation conditions.

5.5 Limitations and Threats to Validity

Three validity threats constrain generalisability. The dataset derives from controlled infrastructure rather than production networks, limiting internal validity**.** External validity is bounded by use of a single Snowflake configuration, whereas operational deployments span browser extensions (35%), standalone applications (8%), and iptproxy (57%). Finally, temporal validity is constrained by ongoing countermeasure development including covertDTLS and Chrome 129+ fingerprint randomisation.

5.6 Ethical Considerations

The dual-use nature of circumvention detection necessitates explicit ethical consideration. We acknowledge that techniques presented may enable suppression of legitimate privacy-seeking behaviour. We emphasise that publication serves defensive purposes by enabling circumvention developers to remediate weaknesses—consistent with responsible disclosure norms.

6. Conclusion

This paper presented a semi-supervised deep autoencoder framework for detecting Tor Snowflake traffic under extreme class imbalance. The approach addresses fundamental limitations of supervised classification, which collapses to 8.7% precision at 1:1000 imbalance. The autoencoder achieves 87.53% precision, a 78.8 percentage point improvement while maintaining 98.47% recall. The comprehensive evaluation across two imbalance ratios provides practitioners with deployment guidance: precision degrades to 87.53% at extreme 1:1000 ratios. Feature ablation analysis confirms that protocol-level DTLS fingerprinting provides the strongest discriminative signal. Future work will investigate ensemble approaches combining multiple autoencoder architectures, adaptive thresholds based on local traffic distributions, and multi-region evaluation to validate geographic generalisability. The dataset and trained models will be made available to support reproducibility and further research.

Acknowledgement

The authors declare no external funding for this research.

AI Declaration: The authors declare that generative artificial intelligence tools were used during the preparation of this manuscript for language editing and grammatical refinement. The authors reviewed and edited all AI-assisted content and take full responsibility for the accuracy, integrity, and originality of the work presented.

References

AlOmar, B., Trabelsi, Z. and Alrabaee, S. (2025) 'Detection of Tor network obfuscated traffic using bidirectional generative adversarial network', Computer Networks, 271, p. 111586.

Bocovich, C., Breault, A., Fifield, D., Serene and Wang, X. (2024) 'Snowflake, a censorship circumvention system using temporary WebRTC proxies', 33rd USENIX Security Symposium, pp. 2635-2652.

Bock, K., Fax, Y., Reese, K., Singh, J. and Levin, D. (2020) 'Detecting and Evading Censorship-in-Depth: A Case Study of Iran's Protocol Whitelister', 10th USENIX Workshop on Free and Open Communications on the Internet (FOCI '20). Berkeley, CA: USENIX Association.

Chen, J., Cheng, G. and Mei, H. (2023) 'F-ACCUMUL: A protocol fingerprint and accumulative payload length sample-based Tor-Snowflake traffic-identifying framework', Applied Sciences, 13(1), p. 622. doi:10.3390/app13010622.

Davis, J. and Goadrich, M. (2006) 'The Relationship Between Precision-Recall and ROC Curves', Proceedings of the 23rd International Conference on Machine Learning (ICML '06). New York: ACM, pp. 233-240. doi:10.1145/1143844.1143874.

Dingledine, R., Mathewson, N. and Syverson, P. (2004) 'Tor: The second-generation onion router', 13th USENIX Security Symposium, pp. 303-320.

Fifield, D. and Gil Epner, M. (2016) 'Fingerprintability of WebRTC', arXiv preprint, arXiv:1605.08805. Available at: https://arxiv.org/abs/1605.08805

Filastò, A. and Appelbaum, J. (2012) 'OONI: Open Observatory of Network Interference', 2nd USENIX Workshop on Free and Open Communications on the Internet (FOCI '12). Berkeley, CA: USENIX Association.

Freedom House (2024) Freedom on the Net 2024: The Struggle for Trust Online. Washington, DC: Freedom House. Available at: https://freedomhouse.org/report/freedom-net/2024/struggle-trust-online

Frolov, S. and Wustrow, E. (2019) 'The use of TLS in Censorship Circumvention', Proceedings of the Network and Distributed System Security Symposium (NDSS 2019). Reston, VA: Internet Society. doi:10.14722/ndss.2019.23511.

He, H. and Garcia, E.A. (2009) 'Learning from Imbalanced Data', IEEE Transactions on Knowledge and Data Engineering, 21(9), pp. 1263-1284. doi:10.1109/TKDE.2008.239.

International Telecommunication Union (2024) Facts and Figures 2024: ICT Data and Statistics. Geneva: ITU. Available at: https://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx

Khattak, S., Elahi, T., Simon, L., Swanson, C.M., Murdoch, S.J. and Goldberg, I. (2016) 'SoK: Making Sense of Censorship Resistance Systems', Proceedings on Privacy Enhancing Technologies, 2016(4), pp. 37-61. doi:10.1515/popets-2016-0028.

Lashkari, A.H., Draper-Gil, G., Mamun, M.S.I. and Ghorbani, A.A. (2017) 'Characterization of Tor traffic using time based features', Proceedings of ICISSP, pp. 253-262.

MacMillan, K., Holland, J. and Mittal, P. (2020) 'Evaluating Snowflake as an indistinguishable censorship circumvention tool', arXiv preprint, arXiv:2008.03254. Available at: https://arxiv.org/abs/2008.03254

Mani, A., Wilson-Brown, T., Jansen, R., Johnson, A. and Sherr, M. (2018) 'Understanding Tor Usage with Privacy-Preserving Measurement', Proceedings of the Internet Measurement Conference (IMC '18). New York: ACM, pp. 175-187. doi:10.1145/3278532.3278549.

Midtlien, T.S. and Palma, D. (2025) 'Fingerprint-resistant DTLS for usage in Snowflake', Free and Open Communications on the Internet, 2025(2), pp. 1-9.

Mirsky, Y., Doitshman, T., Elovici, Y. and Shabtai, A. (2018) 'Kitsune: An Ensemble of Autoencoders for Online Network Intrusion Detection', Proceedings of the 25th Network and Distributed System Security Symposium (NDSS '18). Reston, VA: Internet Society. doi:10.14722/ndss.2018.23204.

Rescorla, E. (2021) 'WebRTC Security Architecture', RFC 8827. Internet Engineering Task Force. doi:10.17487/RFC8827.

Sun, Z. and Shmatikov, V. (2024) 'Differential degradation vulnerabilities in censorship circumvention systems', arXiv preprint, arXiv:2409.06247. Available at: https://arxiv.org/abs/2409.06247

Sundara Raman, R., Shenoy, P., Kohls, K. and Ensafi, R. (2020) 'Censored Planet: An Internet-wide, Longitudinal Censorship Observatory', Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security (CCS '20). New York: ACM, pp. 49-66. doi:10.1145/3372297.3417883.

Wang, L., Dyer, K.P., Akella, A., Ristenpart, T. and Shrimpton, T. (2015) 'Seeing through network-protocol obfuscation', Proceedings of ACM CCS, pp. 57-69. doi:10.1145/2810103.2813715.

Winter, P. and Lindskog, S. (2012) 'How the Great Firewall of China is Blocking Tor', 2nd USENIX Workshop on Free and Open Communications on the Internet (FOCI '12). Berkeley, CA: USENIX Association.

Wu, M., Sippe, J., Sivakumar, D., Burg, J., Anderson, P., Wang, X., Bock, K., Houmansadr, A., Levin, D. and Wustrow, E. (2023) 'How the Great Firewall of China Detects and Blocks Fully Encrypted Traffic', 32nd USENIX Security Symposium, pp. 2653-2670.

Xue, D., Kallitsis, M., Houmansadr, A. and Ensafi, R. (2024) 'Fingerprinting obfuscated proxy traffic with encapsulated TLS handshakes', 33rd USENIX Security Symposium.

Xue, D., Mixon-Baca, B., ValdikSS, Ablove, A., Kujath, B., Crandall, J.R. and Ensafi, R. (2022) 'TSPU: Russia's Decentralized Censorship System', Proceedings of the 22nd ACM Internet Measurement Conference (IMC '22). New York: ACM, pp. 179-194. doi:10.1145/3517745.3561461.

Zavrak, S. and İskefiyeli, M. (2020) 'Anomaly-Based Intrusion Detection From Network Flow Features Using Variational Autoencoder', IEEE Access, 8, pp. 108346-108358. doi:10.1109/ACCESS.2020.3001350.