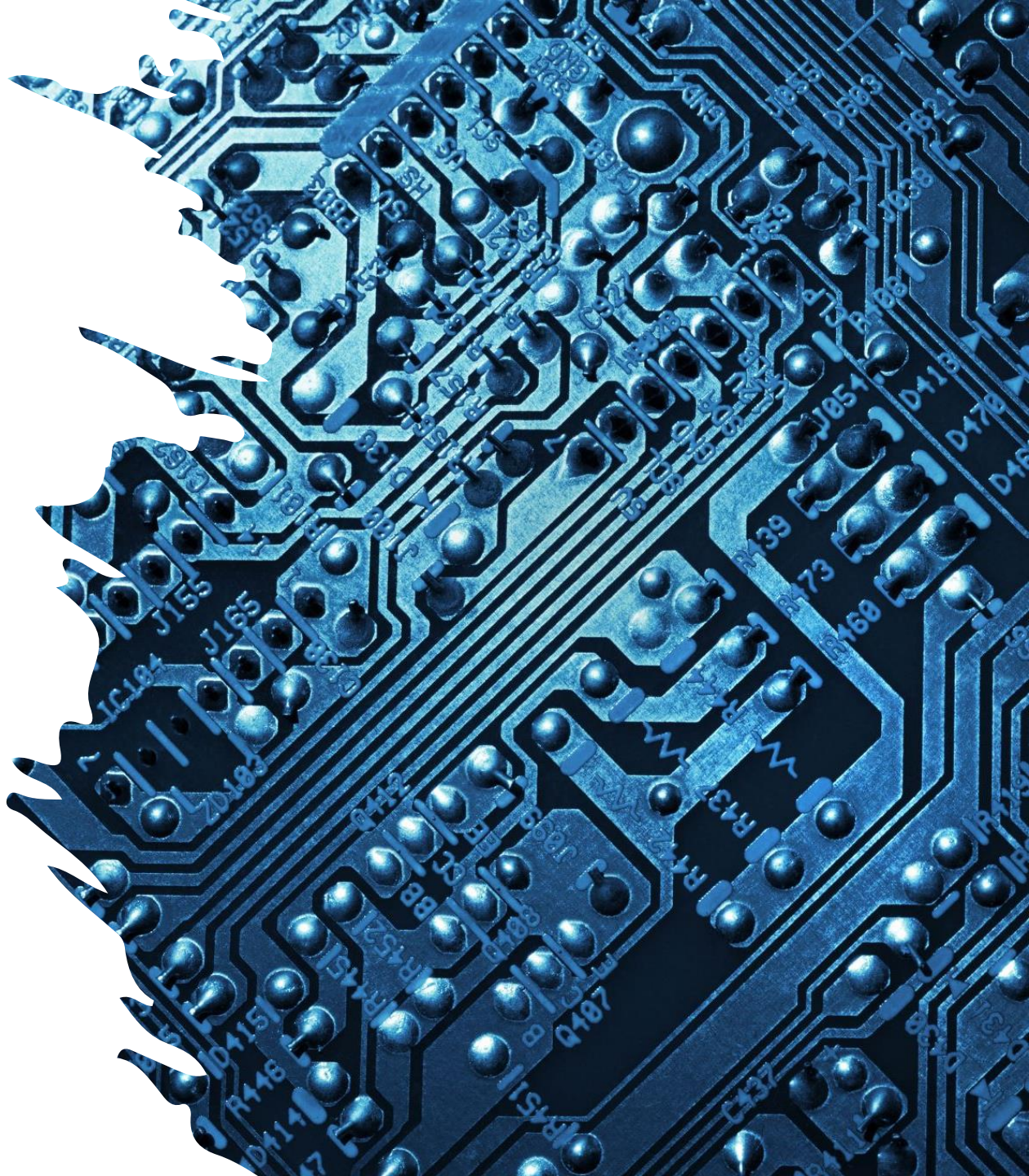
Comparative Evaluation of Data Augmentation Strategies for Rare-Class Intrusion Detection

Assessing Conditional Generative Adversarial Networks Across Benchmark Datasets and Classifiers



The Core Problem

Why Aggregate Accuracy Misleads

High Accuracy ≠ Security Coverage

ML-based IDS can achieve 99%+ overall accuracy while **completely failing** to detect rare, high-impact attack classes. Strong metrics are not evidence of real protection.

Rare Classes Are the Most Dangerous

U2R, R2L, and similar attack types are operationally critical yet **severely underrepresented** in training data — making them the blind spots that matter most.

Class Imbalance = Undetected Threats

Ignoring minority classes is not a statistical nuisance — it maps directly to **real operational risk** and threats that pass through undetected.

⚠ A model that ignores every U2R attack may still report 99%+ accuracy on NSL-KDD.

The Scale of Imbalance

NSL-KDD

52 U2R training samples

0.0413% of 125,973 records

UNSW-NB15

130 Worms samples

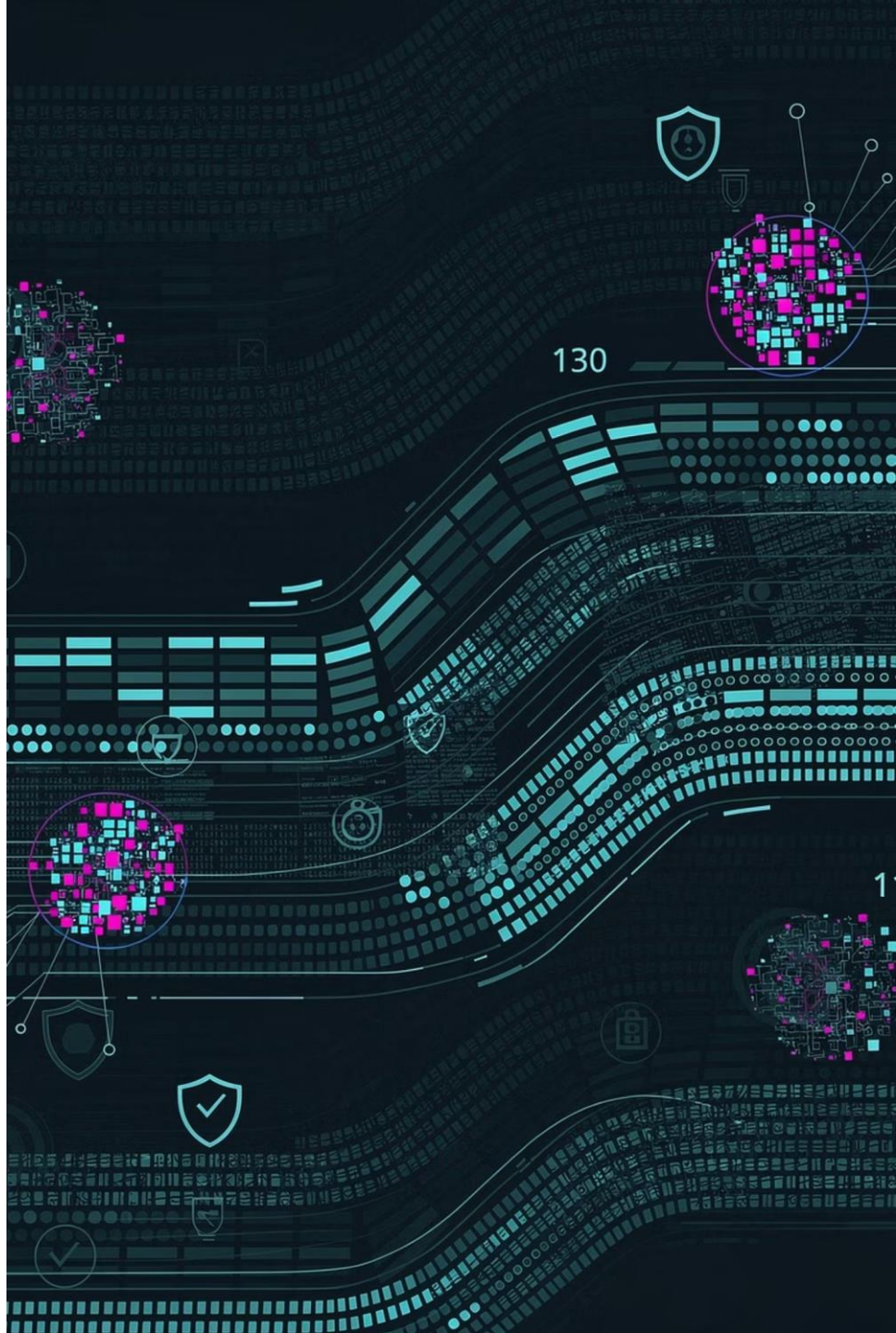
0.0741% of 175,341 records

CICIDS2017

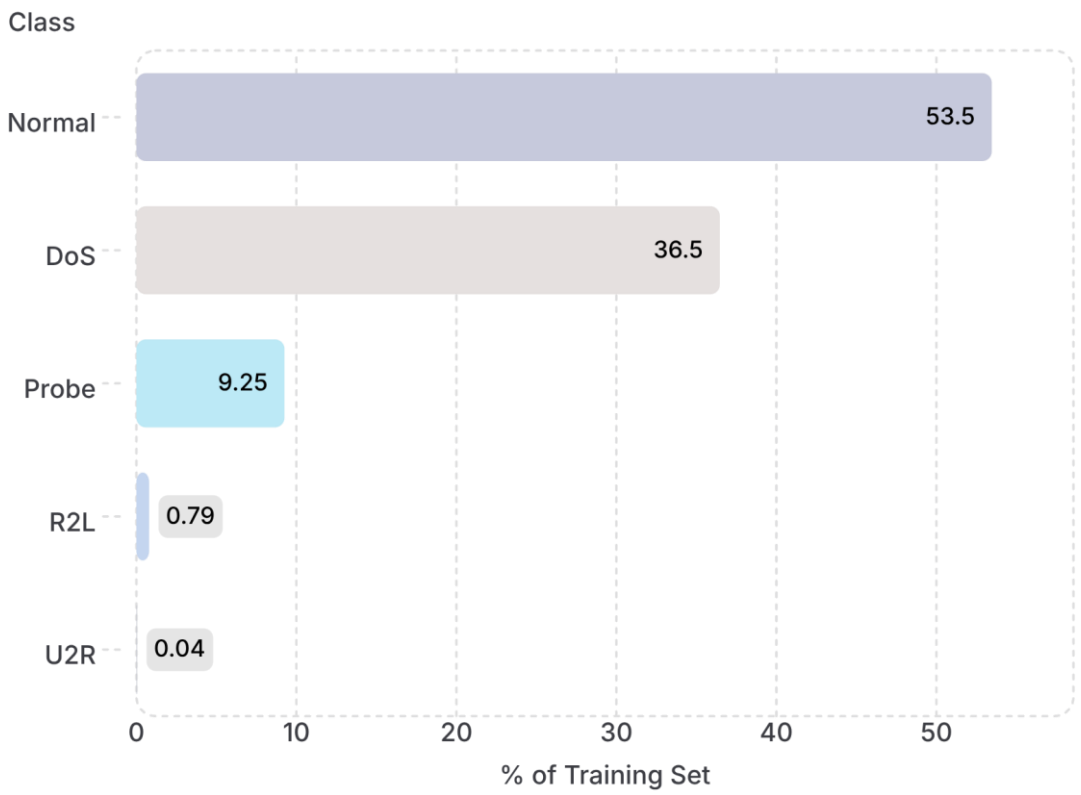
11 Heartbleed samples

Among 2,830,743 total records

The central research question: **Can synthetic augmentation provide useful signal to classifiers from so few real observations?**



Class Distribution in Benchmark Datasets



NSL-KDD Training Set

125,973 total observations. The distribution reveals the extreme skew: U2R and R2L together account for less than 1% of training data, yet represent some of the most operationally critical attack types.

- U2R: only **52 real examples** — barely enough for any generative method to learn from
- R2L: 995 examples — rare but marginally more tractable
- Official test set contains 67 U2R and 2,885 R2L records — a distribution mismatch compounding the challenge

Research Questions

1

RQ1 — Baseline

How did classifiers perform on rare attack detection **without augmentation**?

2

RQ2 — SMOTE & ADASYN

To what extent did **traditional resampling** improve rare-class detection?

3

RQ3 — Standard GAN

Did unconditional GAN augmentation improve on baseline and resampling?

4

RQ4 — cGAN

To what extent did **conditional GAN augmentation** improve upon all prior methods?

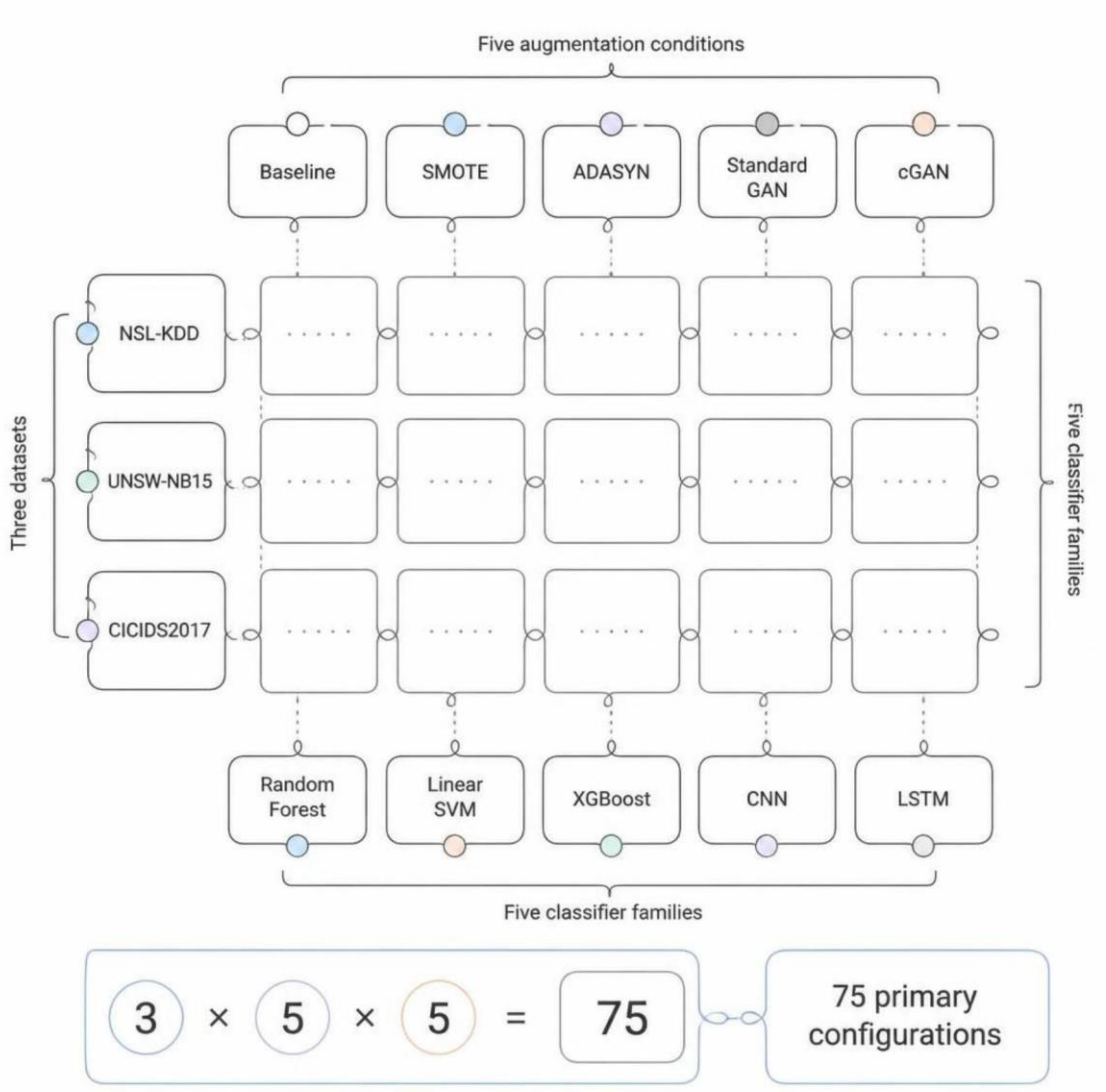
5

RQ5 — Classifier Effect

How did **classifier selection** interact with augmentation strategy?

The study was designed as a **comparative empirical investigation**, not a demonstration of cGAN superiority.

Experimental Design



3 × 5 × 5 = 75 Primary Configurations

Datasets: NSL-KDD, UNSW-NB15, CICIDS2017

Augmentation conditions: Baseline, SMOTE, ADASYN, Standard GAN, cGAN

Classifier families: Random Forest, Linear SVM, XGBoost, CNN, LSTM

Primary evaluation metrics: Macro recall, Macro F1, Target-class recall, Target-class F1. Aggregate accuracy and weighted F1 served as supporting evidence only.

Experimental Controls & Reproducibility

Leakage Prevention

Data were partitioned **before** any preprocessing. Encoders, scalers, and imputers were fitted on training data only. Synthetic records never entered validation or test partitions.

Reserved Test Sets

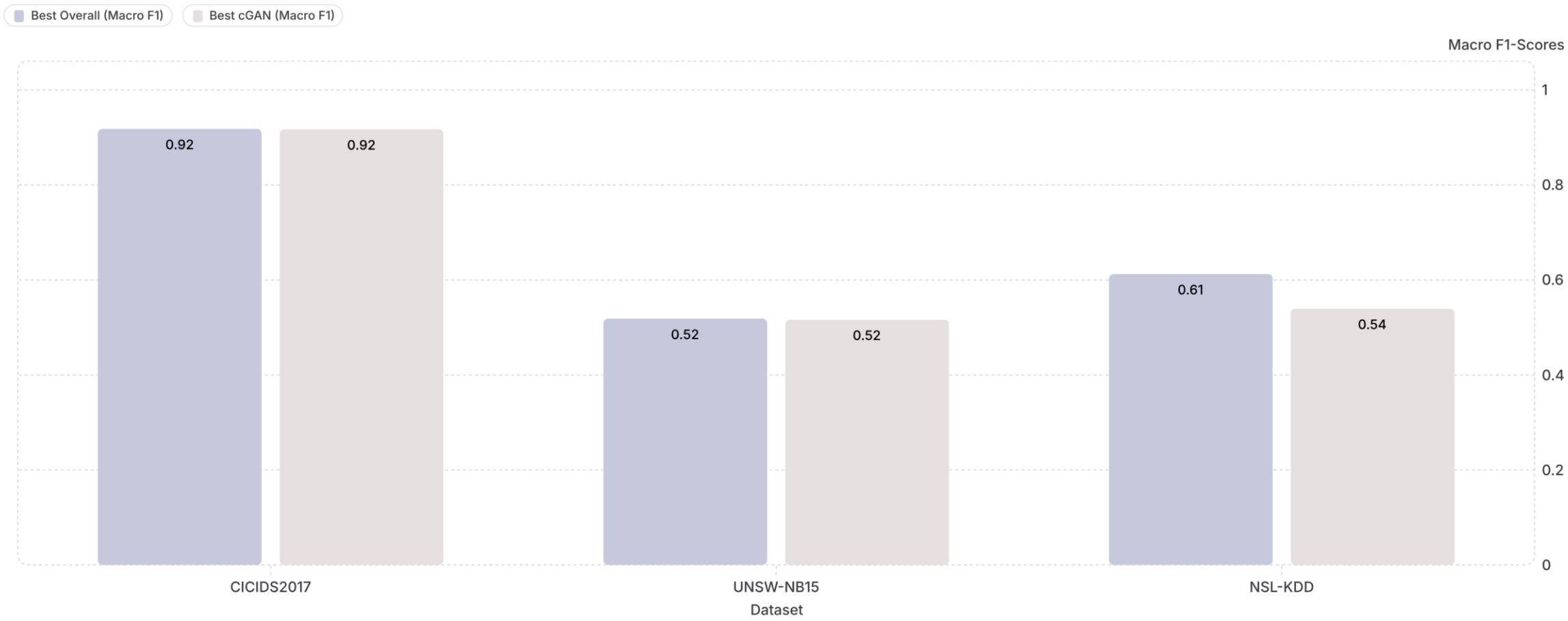
KDDTest+.txt and the 82,332-record UNSW-NB15 official test set were held entirely untouched until final evaluation. CICIDS2017 used a stratified held-out partition.

Reproducibility

Global random seed **42**. Internal split proportions: 70% train / 15% validation / 15% internal test. Experiment identifiers, configurations, and metric outputs were fully archived.

PRIMARY RESULTS

Aggregate Performance Across Datasets



cGAN did not achieve the highest macro F1-score on any of the three datasets. However, margins varied substantially: gaps of 0.0009 and 0.0025 on two datasets, versus 0.0732 on NSL-KDD.

Dataset-Level Results in Detail

CICIDS2017 — Nearly Identical

Best overall: **Baseline XGBoost** — Macro F1: 0.9181 | AUC-ROC: 0.9999
Best cGAN: **cGAN + XGBoost** — Macro F1: 0.9172 | AUC-ROC: 0.9999
Gap: **−0.0009**. cGAN was essentially competitive.

UNSW-NB15 — Marginal Difference

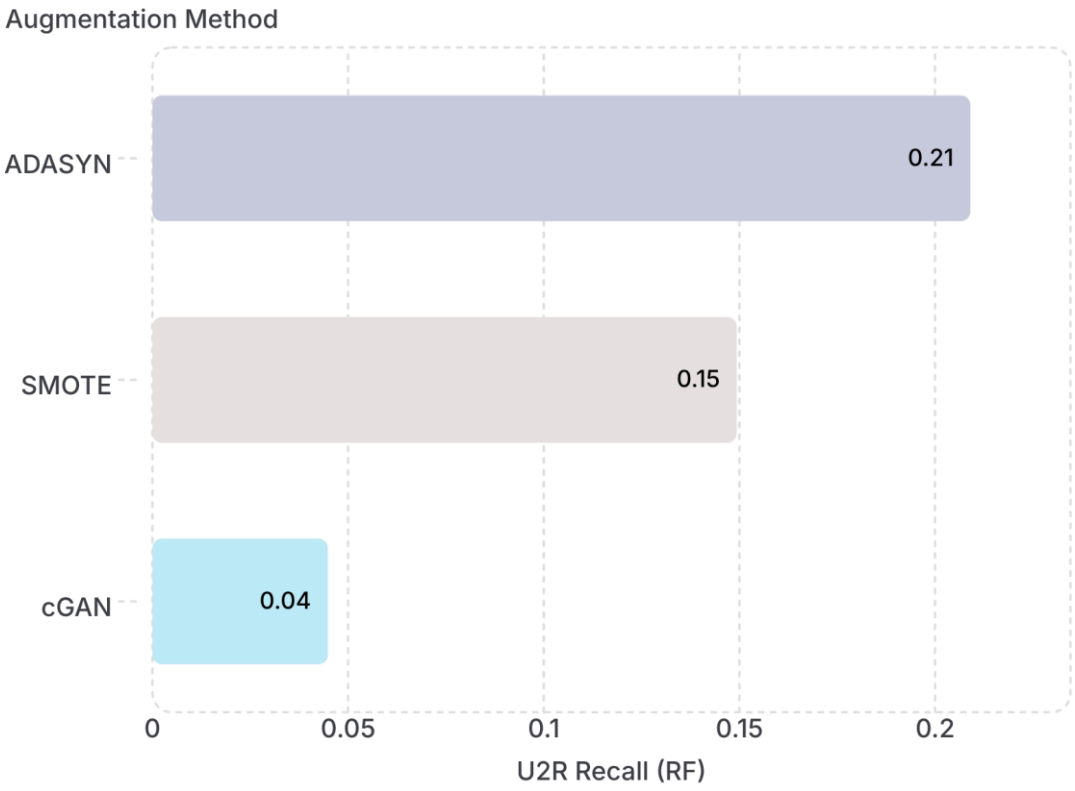
Best overall: **Baseline XGBoost** — Macro F1: 0.5186 | AUC-ROC: 0.9638
Best cGAN: **cGAN + XGBoost** — Macro F1: 0.5161 | AUC-ROC: 0.9632
Gap: **−0.0025**. Again, near parity with the leading configuration.

NSL-KDD — Substantive Gap

Best overall: **SMOTE + SVM** — Macro F1: 0.6125 | AUC-ROC: 0.8875
Best cGAN: **cGAN + LSTM** — Macro F1: 0.5393 | AUC-ROC: 0.9067
Gap: **−0.0732**. cGAN was substantially less competitive here.

MINORITY-CLASS RESULTS

Rare-Class Detection: The Core Question



NSL-KDD U2R — The Motivating Class

U2R was the attack class that originally motivated the dissertation. With only 52 training examples, it represented the hardest augmentation challenge.

ADASYN achieved **recall of 0.2090 and F1 of 0.3256**. SMOTE reached recall of 0.1493. cGAN produced recall of only **0.0448 and F1 of 0.0811**.

cGAN did not improve U2R detection — the class that originally motivated its use.

Class-Specific Outcomes: Mixed Evidence

✗ NSL-KDD R2L

Baseline RF: Recall 0.0582, F1 0.1100
cGAN RF: Recall 0.0475, F1 0.0906
cGAN **reduced** minority-class performance.

✓ UNSW-NB15 Shellcode

Baseline XGBoost: Recall 0.7619, F1 0.4796
cGAN: Recall 0.7884, F1 0.4987
cGAN **improved** detection meaningfully.

✗ UNSW-NB15 Worms

Baseline: Recall 0.5455, F1 0.6234
cGAN: Recall 0.5000, F1 0.5867
Same dataset, same classifier — cGAN **degraded** worms detection.

✓ CICIDS2017 Web Attack XSS

Baseline RF: Recall 0.3367, F1 0.3860
cGAN RF: Recall 0.3571, F1 0.3977
cGAN produced a **modest improvement**.

Key finding: Augmentation effectiveness was class-specific. A method could simultaneously improve one minority class while degrading another.

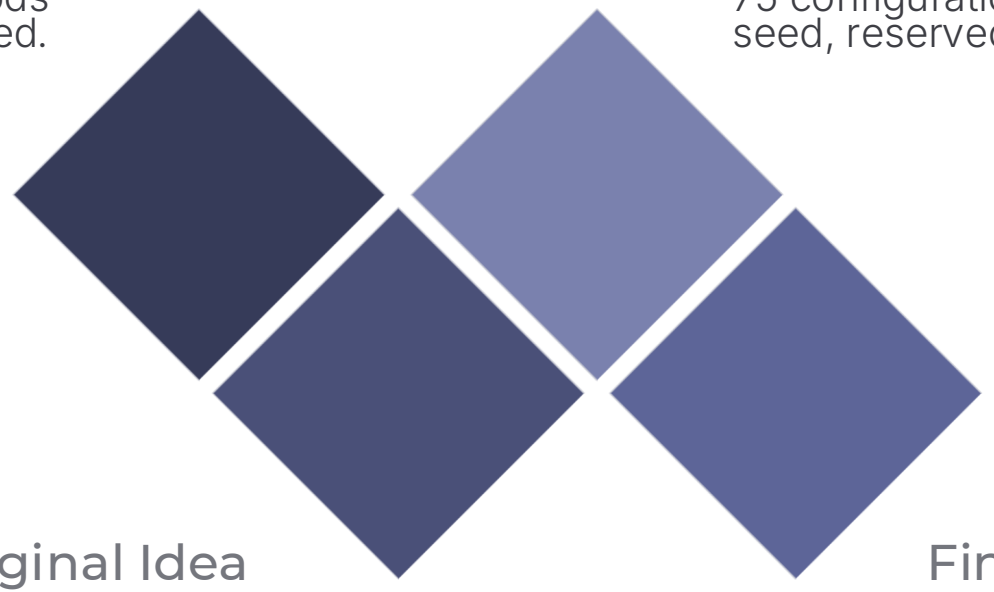
Evolution from Proposal to Final Study

Proposal Design

Three datasets, five methods compared.

Empirical Execution

75 configurations, fixed seed, reserved tests.



Original Idea

Target 20% U2R F1 improvement.

Final Conclusion

Conditional effectiveness; not an enhancement claim.

The core experimental design remained consistent with the approved Proposal. What evolved was the **interpretive framing**: from expected enhancement to rigorous empirical comparison — driven by the results, not by methodological retreat.

Key Methodological Departures

1 Framing: Enhancement → Comparative Evaluation

The Idea Paper targeted ~20% relative improvement in U2R F1. Final cGAN U2R F1 was 0.0811 vs. ADASYN's 0.3256. The dissertation appropriately reframed as empirical comparison rather than forcing a predetermined conclusion.

2 Evaluation: Internal Splits → Reserved Official Test Sets

The Proposal described 70/15/15 internal partitioning. The final study additionally preserved the official NSL-KDD and UNSW-NB15 test files for final evaluation — providing cleaner development/assessment separation.

3 Inferential Statistics: Proposed but Not Applied

The Proposal noted that paired t-tests or Wilcoxon tests *could* be used if sufficient observations existed. The final study used controlled descriptive comparisons — a conservative but appropriate choice given the fixed-seed, limited-trial design.

Principal Contributions & Limitations

Contributions

- **75 primary configurations** across 3 datasets, 5 augmentation conditions, 5 classifier families — one of the more systematic comparative studies in this space
- Demonstrated that greater generative complexity does not guarantee stronger downstream classification
- Confirmed that traditional resampling (SMOTE, ADASYN) remains a competitive baseline
- Showed that **classifier choice materially interacts with augmentation method**: XGBoost led on two datasets unaugmented; SMOTE+SVM led on NSL-KDD
- Reinforced the need to report individual minority-class metrics, not only accuracy or weighted F1

Limitations

- Public benchmark datasets, not live operational IDS traffic
- Extremely small real sample counts for the rarest classes impose hard ceilings on generative learning
- Bounded hyperparameter tuning and GAN/cGAN stochasticity
- Limited independent repeated trials — precludes robust population-level inference
- Offline evaluation only: no latency, throughput, or concept-drift analysis

Final Conclusion

cGAN augmentation was competitive, but not universally superior.

CICIDS2017

cGAN macro F1: **0.9172** vs. best overall: **0.9181** — near parity. Gap of 0.0009.

UNSW-NB15

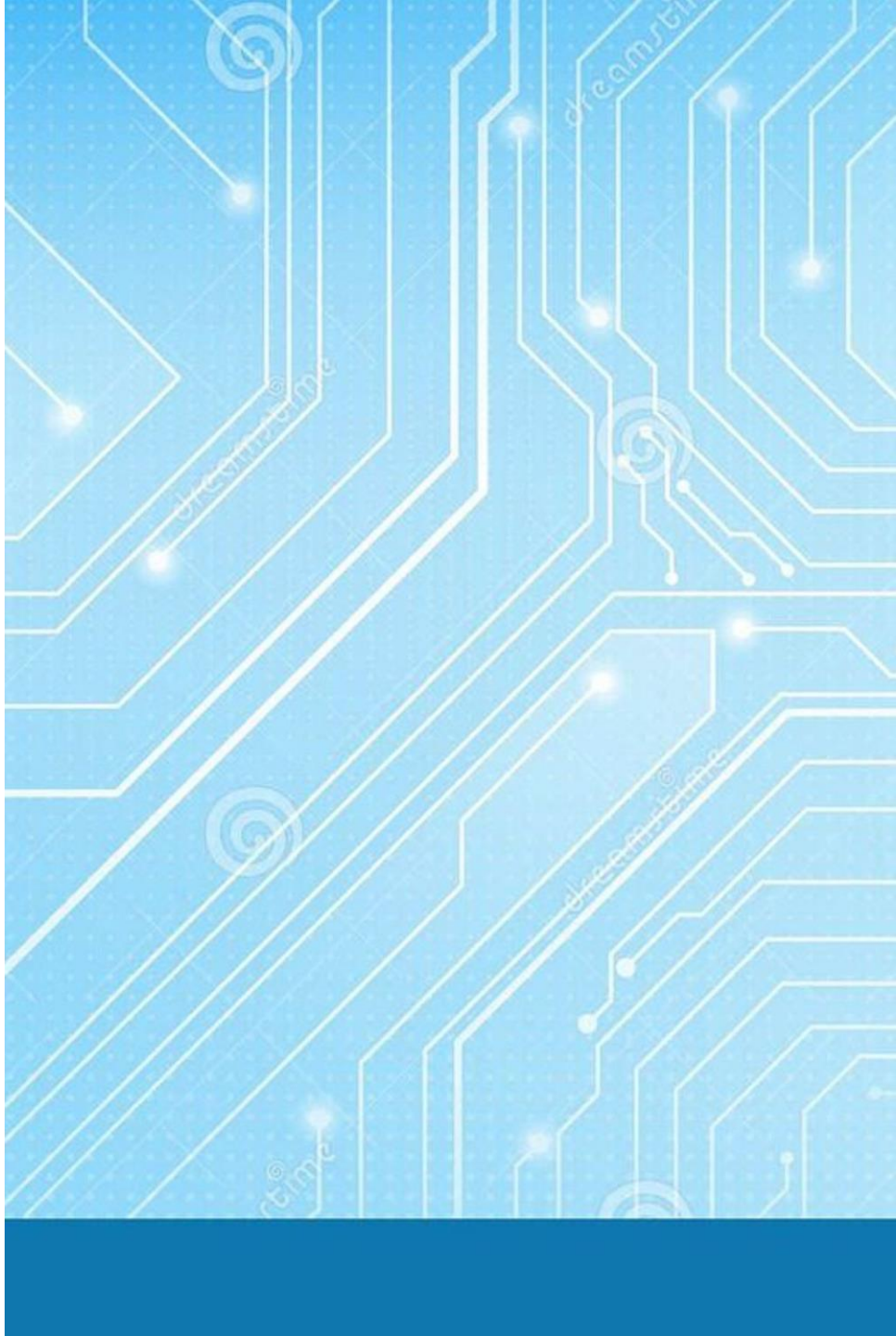
cGAN macro F1: **0.5161** vs. best overall: **0.5186** — marginal difference. Gap of 0.0025. Shellcode improved; worms did not.

NSL-KDD

cGAN macro F1: **0.5393** vs. SMOTE+SVM: **0.6125**. Gap of 0.0732. U2R recall remained near zero under cGAN.

Overall Takeaway

cGAN effectiveness is **conditional on the dataset, target class, classifier, and generation procedure**. Traditional resampling and strong unaugmented classifiers remained difficult to outperform consistently.



References

(1 of 3)

1. Abdelkhalek, A., & Mashaly, M. (2023). Addressing the class imbalance problem in network intrusion detection systems using data resampling and deep learning. *The Journal of Supercomputing, 79*, 1–34. **https://doi.org/10.1007/s11227-023-05073-x**
2. Ahmad, I., Abdullah, A. B., & Alghamdi, A. S. (2010). Applying neural network to U2R attacks. In *2010 IEEE Symposium on Industrial Electronics and Applications (ISIEA)* (pp. 295–299). IEEE. **https://doi.org/10.1109/ISIEA.2010.5679451**
3. Babu, K. S., & Rao, Y. N. (2023). Modified conditional generative adversarial network (MCGAN) for class imbalance problems in network intrusion detection system. *Applied Sciences, 13*(4), Article 2576. **https://doi.org/10.3390/app13042576**
4. Çakır, B. (2019). *Zero-day attack detection with deep learning* [Master's thesis, Middle East Technical University].
5. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. **https://doi.org/10.1613/jair.953**

References

(2 of 3)

1. Divekar, A., Parekh, M., Savla, V., Mishra, R., & Shirole, M. (2018). Benchmarking datasets for anomaly-based network intrusion detection: KDD CUP 99 alternatives. In 2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS) (pp. 1–8). IEEE. **https://doi.org/10.1109/CCCS.2018.8586840**
2. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems, 27.
3. He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE International Joint Conference on Neural Networks (pp. 1322–1328). IEEE. **https://doi.org/10.1109/IJCNN.2008.4633969**
4. Jeya, P. G., Ravichandran, M., & Ravichandran, C. S. (2012). Efficient classifier for R2L and U2R attacks. International Journal of Computer Applications, 45(21), 28–32.
5. Mbow, M., Koide, H., & Sakurai, K. (2022). Handling class imbalance problem in intrusion detection system based on deep learning. International Journal of Networking and Computing, 12(2), 467–492. **https://doi.org/10.15803/ijnc.12.2_467**
6. Mitchell, R., & Chen, I.-R. (2014). A survey of intrusion detection techniques for cyber-physical systems. ACM Computing Surveys, 46(4), Article 55. **https://doi.org/10.1145/2542049**

References

(3 of 3)

1. Patel, M., Wang, X., & Mao, S. (2020). Data augmentation with conditional GAN for automatic modulation classification. In Proceedings of the 2nd ACM Workshop on Wireless Security and Machine Learning (pp. 31–36). ACM. **https://doi.org/10.1145/3395352.3402622**
2. Rahman, S., Pal, S., Mittal, S., Chawla, T., & Karmakar, C. (2024). SYN-GAN: A robust intrusion detection system using GAN-based synthetic data for IoT security. Internet of Things, 26, Article 101212. **https://doi.org/10.1016/j.iot.2024.101212**
3. Rao, Y. N., & Babu, K. S. (2023). An imbalanced generative adversarial network-based approach for network intrusion detection in an imbalanced dataset. Sensors, 23(1), Article 550. **https://doi.org/10.3390/s23010550**
4. Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. In 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications (pp. 1–6). IEEE. **https://doi.org/10.1109/CISDA.2009.5356528**
5. Thana-Aksaneekorn, C., Kosolsombat, S., & Luangwiriya, T. (2024). Machine learning classification for intrusion detection systems using the NSL-KDD dataset. In 2024 IEEE International Conference on Cybernetics and Innovations (ICCI) (pp. 1–6). IEEE. **https://doi.org/10.1109/ICCI60780.2024.10532265**