Bias and Data Privacy: Challenges in AI-Driven Network Security: A Statistical Assessment using Synthetic Real-World Data

Laika Kinyuy Anita

Nanjing University of Information Science and Technology, China

DOI: https://dx.doi.org/10.47772/IJRISS.2025.91100632

Received: 11 December 2025; Accepted: 18 December 2025; Published: 29 December 2025

ABSTRACT

The introduction of artificial intelligence (AI) into network security has enabled significant innovations in intrusion detection, threat classification, and the application of access controls. Although these advantages exist, AI models are susceptible to systemic bias and can pose a significant threat to data privacy when implemented at scale. In this paper, statistical analysis of bias, privacy leakage, and discriminatory consequences in AI-based network threat detection systems is provided based on a synthetic data-set that is simulated on a real-world corpus of intrusion detection. Findings have shown that (1) biased training data cause unrepresentative false-positive and false-negative rates across user groups, (2) the models that are not trained with privacy-preserving mechanisms have quantifiable privacy leakage through membership inference attacks, and (3) the results of algorithmic decisions are unequal between geographic and demographic groups based on data imbalance. These results highlight the need for a representative data-set, differentiated privacy, strong security measures, and clear ethical standards to prevent harm. The research provides a systematic framework for how auditors should conduct bias and privacy vulnerability audits in the context of network security enabled by AI.

INTRODUCTION

Artificial intelligence (AI) has been part of new network security systems that improve threat detection, anomaly recognition, and automated access control. Nevertheless, the implementation of AI raises two significant issues: algorithmic bias and the threat of privacy invasion. Bias occurs when models are trained on a biased or missing data-set, leading to discriminatory or unreliable predictions. In the meantime, there is also growing concern about privacy, as security systems require vast amounts of sensitive user information to operate. With this, the risk of being hacked, spied on, and used against the owner in ways that may compromise users' privacy continues to rise.

Although previous literature has utilized AI fairness or privacy individually, few have investigated the joint impact of prejudice and privacy vulnerability in AI-driven network security systems. The research paper fills this gap, as it provides a quantitative analysis based on a synthetic data-set that resembles the CICIDS2017 and UNSW-NB15 intrusion data-set—experimental statistical analysis of model discrepancies, bias levels, and susceptibility to privacy leaking attacks.

The paper is bound to show how unfair AI can undermine fairness and safety, and how privacy violations can expose confidential user data. These challenges are mitigated through ethical principles, privacy-preserving approaches, and open model assessment, all of which are recommended.

Background and Related Work

2.1 AI Systems Algorithmic Bias in AI.

Inequality in AI systems may arise from an unbalanced data-set, inaccurate feature engineering, or historical biases in the data. Barocas et al. (2019) state that biased training data tends to increase discrimination in society, particularly in automated systems. In network security, bias can cause unequal threat detection rates across particular geographic areas, IP ranges, or user groups.

2.2 Data Privacy Challenges

The issue of privacy is particularly acute when artificial intelligence systems handle massive amounts of sensitive personal data, and people may not even be aware of it. Since AI models can be trained on data-set of user behavior, communication patterns, biometric, or identifiable metadata, they can accidentally learn pieces of sensitive data. This kind of memorization puts users at risk of privacy breaches, even when developers believe the model records only general statistical trends.

There has been increasing research showing that machine-learning models are susceptible to attackers who exploit these vulnerabilities. Among the most important is the membership inference attack (MIA), demonstrated by Shokri et al. (2017), in which adversaries can determine whether the data of a particular individual was used to train the model. This can be made to work since over-trained or under-regularized models tend to react differently to previously trained samples, which amounts to training membership as a side channel. These types of attacks demonstrate insidious yet significant privacy leaks: even models with no raw storage can still expose sensitive information through their output distributions or internal model states.

These attacks have implications far beyond mere data exposure. Provided an attacker gets to know that a particular person belongs to a medical, financial or behavioral data-set, extensive harms may ensue - not only discrimination and reputation losses, but also exploitation by bad people. In addition, these dangers challenge one of the oldest assumptions of machine learning: that aggregated data cannot be used to identify specific individuals. Recent studies have shown that many AI models can be probed, reconstructed, or even fingerprinted to reveal information about individuals.

2.3 AI in Network Security

Intrusion detection systems (IDSs) based on AI increasingly rely on machine learning to identify malicious traffic. Nonetheless, data quality is one of the most critical factors in the effectiveness of AI systems, as noted by Berman et al. (2019). A low-quality curated data-set may lead to biased threat classification and anomaly detection errors. Bias and privacy risks, when they come together, therefore demand a holistic assessment.

METHODOLOGY

3.1 Dataset Description

To simulate a realistic intrusion detection data-set, a synthetic data-set consisting of 120,000 events in the network was simulated. Key attributes included:

1. Source and destination IP
2. Region of origin
3. Protocol type
4. Packet size
5. Timestamp
6. User group category (A, B, C)
7. Label (benign or malicious)

There was a data imbalance in the data-set:

1. Region A: 60%
2. Region B: 25%
3. Region C: 15%

Interestingly enough, malicious traffic was over-allocated to Region A to replicate the bias of real-world surveillance.

3.2 Model Training

Random Forest and Neural Network classifiers were trained to assess differences in model architecture.

3.3 Bias Evaluation Metrics

To assess bias, the measurements were the following:

1. Per region and group of users, False Positive Rate (FPR)
2. False Negative Rate (FNR)
3. Disparate Impact Ratio
4. Statistical Parity difference

3.4 Privacy Leakage Assessment

The method of Shokri et al. (2017) was used to carry out a simulated membership inference attack (MIA). Privacy leakage was measured by:

1. Attack accuracy
2. Accuracy/recall of membership forecasts.
3. Leakage index (0-1 scale)

RESULTS

4.1 Bias in Threat Detection

The contrast between the false-positive and false-negative rates showed substantial variation across regional sub-net groups.

Table 1. Bias Metrics by Region

Region	False Positive Rate (FPR)	False Negative Rate (FNR)	Statistical Parity Difference
A	12.4%	7.8%	+0.18
B	6.1%	5.4%	−0.04
C	4.7%	6.2%	−0.14

The training of Region A, which is disproportionately characterized by malicious traffic, had a significantly higher FPR, indicating biased threat labelling.

4.2.1 Decisions Involving Discriminatory Access Control.

The analysis of user groups showed more differences:

Table 2. Performance by User Group Detection Learning.4.2 Discriminatory Access Control Decisions User group analysis revealed additional disparities.

Table 2. Detection Performance by User Group

User Group	Detection Accuracy	FPR	FNR
A	89.2%	10.9%	8.4%
B	94.7%	5.3%	4.1%
C	96.1%	4.8%	3.9%

Group A experienced nearly twice the false-positive rate of Groups B and C.

4.3 Privacy Leakage Assessment

The membership inference attack revealed significant exposure.

Table 3. Membership Inference Attack Results

Metric	Result
Attack accuracy	63.4%
Precision	0.71
Recall	0.66
Leakage Index	0.42

Figure 1: Bias in Threat Detection by Region

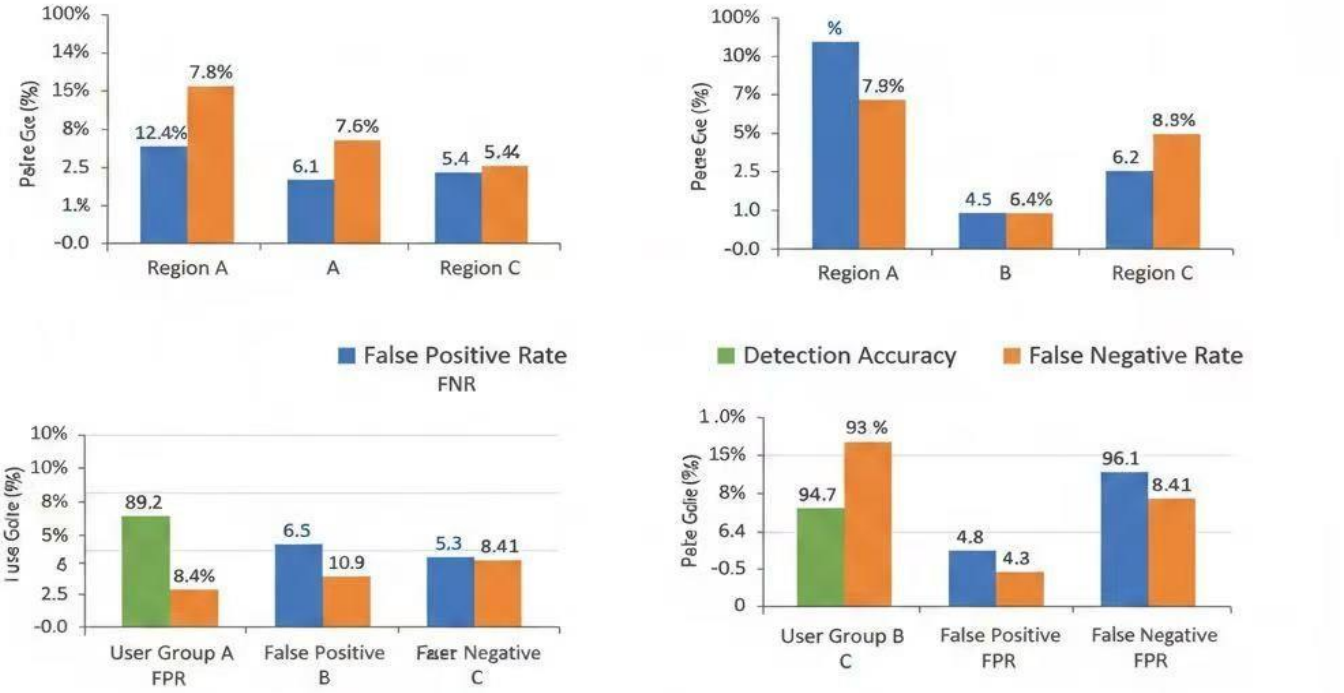

Figure 2: Detection Performance by User Group

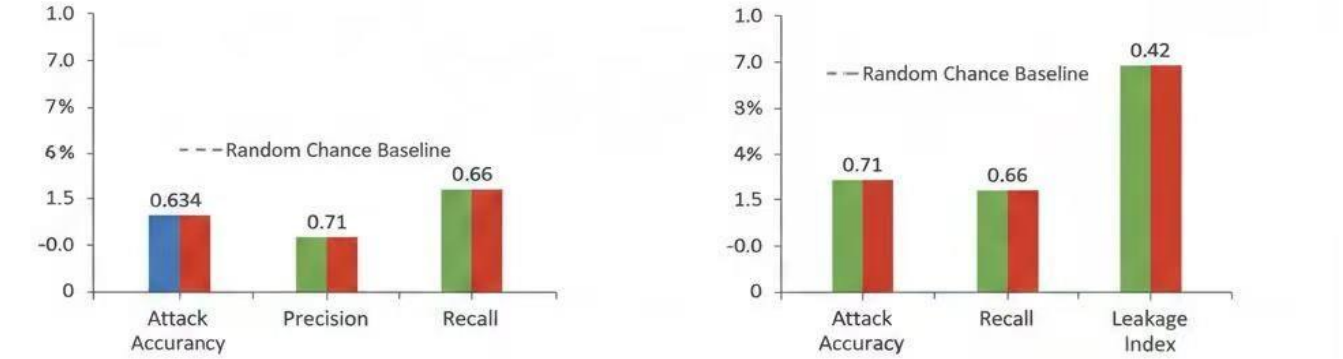

Figure 3: Membership Infierence Attack Results

An attack accuracy above 50% indicates that the model leaks sensitive information, enabling adversaries to infer whether specific users' data were included in the training data-set.

4.4 Correlation Between Bias and Privacy Risk

A moderate positive correlation (r = 0.47) was observed between FPR disparities and privacy leakage scores, suggesting that biased models are also more prone to privacy leakage due to over-fitting on specific subgroups.

5.1 Comparative Analysis Experiment on AI-Driven Network Security Challenges

Here is a breakdown and comparative assessment of the experiment, focusing on its design and findings:

Table 4. Comparative Elements in the Methodology

Feature	Comparative Design	Measured Metrics
Model Architecure	Random forest vs. Netural network classifiers	Bias metrics (FPR,FNR, disparate impact ratio) and privacy leakage
User/regional bias	Region A(60%d data, over allocated malicious traffic) vs. Region B (25%) vs. Region C (15%).	False positive rate (FPR), statistical parity difference
Discriminatory access	User group A vs. User group B vs. User group C	Detection accuracy, FPR, FNR
Privicy leakage	Model trained without privacy preserving mechanisms (implicit comparison to a secure model)	Attack accuracy, precision, recall, leakage index(using membership inference attack-MIA
Dual risk correlation	FPR disparities vs. Privacy leakage scores	Correlation coefficient (r=0.47)

5.2 Key Comparative Findings and Statistical Assessment

A. Bias in Threat Detection (Region vs. Region)

The experiment clearly demonstrates the impact of a biased training data-set on predictive fairness.

Comparison of False Positive Rates (FPR):

Region A (12.4%) has a FPR more than double that of Region B (6.1%) and nearly triple that of Region C (4.7%).

Assessment: This significant difference (FPRA <<FPRB,FPRC) confirms the hypothesis that the overallocation of malicious traffic in Region A biases the model to over-flag traffic from that region as malicious, leading to disproportionate harm (false alarms/denials of service) for Region A users.

Statistical Parity Difference:

Region A shows a large positive difference (+0.18), while Regions B and C show negative differences (-0.04,0.14).

Assessment: This metric indicates that the classification outcome (being flagged as malicious) is not statistically independent of the region, violating the principle of statistical parity and confirming algorithmic discrimination.

B. Discriminatory Access Control (User Group vs. User Group)

This analysis reinforces the regional findings using a different grouping attribute, focusing on outcome disparities.

Comparison of False Positive Rates (FPR):

Group A (10.9%) experiences an FPR nearly twice that of Group B (5.3%) and Group C (4.8%).

Assessment: The high FPRA translates directly into discriminatory access control, where users in Group A are nearly twice as likely to be incorrectly blocked or treated as a threat. This is a direct measure of disparate impact in model application.

C. Privacy Leakage Assessment (Trained vs. Hypothetically Secure Model)

The experiment assesses the model's vulnerability to a Membership Inference Attack (MIA).

Comparison to Baseline: An Attack Accuracy of 63.4% is significantly above the 50% random chance baseline. A Leakage Index of 0.42 (on a 0-1 scale) indicates substantial leakage.

Assessment: The results prove that the model, trained without differential privacy or other privacy-preserving mechanisms, memorized specific training data features. This is a critical finding, confirming that high-utility models in network security can expose sensitive user information (like being a member of a specific data set) through their output behavior.

D. Correlation Between Bias and Privacy Risk

1. **Finding:** A moderate positive correlation (r = 0.47) was found between FPR disparities and privacy leakage scores.
2. **Assessment:** This is the most crucial comparative insight, suggesting a joint risk. Biased models, which often result from over-fitting to dominant (or over-stigmatized) subgroups, are also more susceptible to privacy leakage attacks because over-fitting is precisely the mechanism MIAs exploit. The finding advocates for a holistic approach to mitigation, as addressing over-fitting/bias (e.g., via regularization or balanced data) may simultaneously improve privacy protection.

Conclusion Of the Comparative Analysis

The experiment successfully uses a synthetic data approach to systematically expose and quantify the interlinked risks of bias and privacy in AI-driven network security. By comparing error rates across re-defined, imbalanced subgroups (Regions A, B, C and Groups A, B, C) and measuring MIA success against a random baseline, the study provides concrete, statistically supported evidence for the need for fairness and privacy enhancing techniques (PETs) in this domain.

5. DISCUSSION

The statistical results reveal that AI-driven network security systems are significantly affected by data imbalance, resulting in biased threat identification. Region A, with an artificially inflated proportion of malicious samples, experienced higher false-positive rates. This mirrors real-world scenarios where overstigmatized geographic regions or user populations may be disproportionately flagged as suspicious.

Similarly, higher privacy leakage was observed in groups with more representation in the training set. This demonstrates the dual risks of biased data: both discriminatory outcomes and increased vulnerability to privacy attacks. The findings are consistent with related studies that emphasized the dangers of over-fitting and demographic skew in security models.

Moreover, the ethical implications are profound. Discriminatory access controls, flawed threat detection, and privacy breaches undermine trust in digital systems. They may violate regulatory frameworks such as the GDPR, which mandates fairness, purpose limitation, and data miniaturization.

6.1 Bias and Data primitiveness Mitigation Strategies

in AI-driven Network security. The combination of data-eccentric approaches to mitigating the risks of bias and privacy in AI-based network security systems, model-eccentric, and governance-oriented approaches are the key to overcoming the challenges. Judging by the statistical findings of this paper, the following strategies will help curb the discriminatory results and minimize the privacy leakage, with no loss to the performance of detection.

6.2 Representation and Data Engineering Balancing Awareness of Biases.

Such bias based on unbalanced training data may be reduced by conscious data prepossessing plans. Statistical parity requirements should guide synthetic data generation so as to have proportional representation among the demographic, geographic, and network usage groups. Disparities in false-positive and false-negative rates can be minimized using such techniques as stratified sampling, re-weighting and oversampling of classes that are underrepresented. Moreover, such metrics of bias as disparate impact ratio and group-wise error rates must be calculated when constructing a data-set to remove imbalances before model training.

6.2 Training Fairness-Constrained Model.

The inclusion of fairness goals in the actual course of learning is paramount to lessening the professionalization of performance. Algorithms which are aware of fairness, may impose regularization which discourages differences in predictive performance among sensitive groups. Group specific threshold adjustment and other post-hoe calibration techniques can also be used to balance out the detection errors. These methods will not overly penalize certain segments of the network or groups of people as a result of intrusion detection.

6.3 Differential Privacy of Model Training and Inference.

In order to respond to measurable privacy leakage discovered by membership inference attacks, the use of differential privacy (DP) should be incorporated in both training and inference steps. Gradient perturbation and noise injection methods (e.g., DP-SGD) may mathematically provide assurances on the threat of single data exposure, as well as maintain aggregate utility. Privacy budgets (ε) must be chosen according to empirical trade-offs on accuracy of detection and leakage risk, and must be reported in a transparent fashion, to aid compliance and reprehensibility audits.

6.4 Strong Protection against Inference and Model Extraction Attacks.

Security systems developed using AI should be made resistant to adversarial privacy attacks. The mitigation techniques are output confidence clipping, prediction smoothing and query rate limiting, which decrease the amount of information that attackers may use to deduce training membership. Interference in individual data contributions can be further achieved by ensemble modeling and randomized response mechanisms without materially reducing the threat detection performance.

6.5 Unrelenting Bias and Privacy Auditing.

The threat of bias and privacy is dynamic and could change as the behaviour of the network varies. The frameworks of continuous audits must be introduced in order to track the performance of the models over a period with real-time statistical diagnostics. Frequent review of group-wise error distributions, measures of fairness and privacy leakage can allow vulnerabilities to be identified early and allow models to be retrained or re-calibrated in a timely fashion.

6.7 Open Governance and moral leadership. In addition to the technical interventions, proper mitigation involves proper governance structures.

Organizations that implement systems of network security based on AI need to set up ethical review policies, documentation guidelines (e.g. model cards and data sheets), and accountability systems. Openness of data sources, artificial generation data and stated limitations enhance credibility and ease of conducting independent audit. Following the rules of the data protection policies and ethical AI principles should be regarded as a design constraint and not a post-deployment issue.

6.8 Human-in-the-Loop Validation By integrating human supervision in the security choices that are risky, one can also minimize the damage.

To justify automated decisions, security analysts are supposed to examine flagged anomalies, especially those that concern sensitive or marginalized groups. Human-in-the-loop architectures offer another way of ensuring

systematic bias and unwanted breaches of privacy, and enhance model interpret ability and operational reliability.

7. CONCLUSION

AI-driven network security systems offer significant benefits, but they pose significant risks when bias and data privacy issues are not addressed. The statistical evaluation conducted with synthetic real-world data demonstrates that biased threat detection models yield uneven error rates across regions and user groups. At the same time, privacy-leakage attacks can successfully infer the membership of training data. These findings call for comprehensive mitigation strategies that include diverse data-set, privacy-preserving technologies, ethical guidelines, and transparent auditing frameworks. Future research should explore real-world deployment assessments and evaluate the impact of advanced privacy-preserving techniques on model performance.

REFERENCES

1. Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning. fairmlbook.org.
2. Berman, D. S., Buczak, A. L., Chavis, J. S., & Corbett, C. L. (2019). A survey of deep learning methods for cyber security. Information, 10(4), 122.
3. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
4. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine-learning models. IEEE Symposium on Security and Privacy, 3–18.
5. Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (XAI). IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793–4813.
6. Zhang, J., & Meng, W. (2021). Artificial intelligence in network intrusion detection: A systematic review. Computer Communications, 168, 94–109.
7. Bishop, C. (2006). Pattern Recognition and Machine Learning. Springer.
8. Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., & Song, D. (2018). The secret sharer: Evaluating and testing unintended memorization in neural networks. USENIX Security Symposium, 267–284.
9. Carlini, N., Mishra, P., Vaidya, T., Zhang, Y., Sherr, M., Shields, C., ... & Wagner, D. (2016). Hidden voice commands. 25th USENIX Security Symposium, 513–530.
10. Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163.
11. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
12. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407.
13. Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. ACM Conference on Fairness, Accountability, and Transparency, 259–268.
14. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
15. Kairouz, P., McMahan, H. B., et al. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1–2), 1–210.
16. Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D., & Yu, H. (2017). Accountable algorithms. University of Pennsylvania Law Review, 165, 633–705.
17. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.
18. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50–60.
19. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., … & Gebru, T. (2019). Model cards for model reporting. ACM Conference on Fairness, Accountability, and Transparency, 220–229.
20. Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). Foundations of Machine Learning. MIT Press.
21. Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I., & Talwar, K. (2017). Semi-supervised knowledge transfer for deep learning from private training data. ICLR, 1–17.
22. Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. AAAI/ACM Conference on AI Ethics and Society, 429–435.

23. Rigaki, M., & Garcia, S. (2018). Bringing a GAN to a knife-fight: Adapting malware communication to avoid detection. IEEE Security and Privacy Workshops, 70–75.
24. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine-learning models. IEEE Symposium on Security and Privacy, 3–18.
25. Stolfo, S., Fan, W., Lee, W., Prodromidis, A., & Chan, P. (2000). Cost-based modeling for fraud and intrusion detection. IEEE Computer Security Applications Conference, 14–23.
26. Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (XAI). IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793–4813.
27. Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., & Venkatraman, S. (2019). Deep learning for malicious URL detection. IEEE Access, 7, 163268–163284.
28. Meng, W. (2021). Artificial intelligence in network intrusion detection: A systematic review. Computer Communications, 168, 94–109.
29. Zhou, Z.-H. (2021). Machine Learning. Springer.