Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Comparative Evaluation of Machine Learning Models for IoT Intrusion Detection Under Class Imbalance and Feature Drift

Lucien beyli

Lucien@outlook.com

Amsterdam

Research Article

Keywords: Internet of Things, intrusion detection, machine learning, class imbalance, calibration, feature drift, cybersecurity

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11738586/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Comparative Evaluation of Machine Learning Models for IoT Intrusion Detection Under Class Imbalance and Feature Drift

Lucien Beyli[a]

[a]*Independent Researcher,*

ARTICLE INFO

Keywords:
Internet of Things
intrusion detection
machine learning
class imbalance
calibration
feature drift
cybersecurity

ABSTRACT

Intrusion-detection studies for Internet of Things (IoT) systems commonly report strong classification performance, but comparisons are often difficult because class imbalance, probability calibration, feature count, and distribution shift are evaluated separately. This study performs a controlled empirical comparison of five conventional machine-learning configurations under a common eight-class security benchmark. The benchmark contains 18,000 development observations with 28 traffic-like features and an imbalanced class distribution, while a separate 5,000-observation shifted set introduces feature scaling, offsets, and additive noise. Logistic regression, random forest, Extra Trees, histogram gradient boosting, and calibrated Extra Trees are evaluated across three independent random seeds. On the development test split, calibrated Extra Trees achieved balanced accuracy of 0.981 ± 0.005, macro-F1 of 0.984 ± 0.004, and Matthews correlation coefficient of 0.983 ± 0.005. Under feature drift, balanced accuracy remained 0.981 ± 0.006 and macro-F1 0.983 ± 0.005. Probability calibration reduced expected calibration error for Extra Trees from 0.182 ± 0.006 to 0.014 ± 0.001 on the development split. A feature-ablation experiment showed that 18 of 28 features retained macro-F1 of 0.978 ± 0.002, whereas reducing the representation to 10 features lowered macro-F1 to 0.931 ± 0.012. The results provide a reproducible comparison of standard classifiers and illustrate the practical value of reporting balanced metrics, calibration, and feature-drift performance together. No new classifier or intrusion-detection architecture is proposed.

1. Introduction

Internet of Things deployments combine heterogeneous devices, constrained endpoints, wireless links, cloud services, and management interfaces. Security therefore depends on more than one technical layer, and intrusion detection remains an important operational control for identifying anomalous or malicious activity after deployment [1, 2, 3, 4]. Large-scale IoT compromises such as Mirai demonstrated how weakly protected connected devices can be recruited into coordinated attacks [5]. Subsequent work has examined device behavior, traffic characterization, and machine-learning-based intrusion detection for connected systems [6, 7, 8, 74, 80].

The machine-learning literature also shows that performance depends strongly on representation, class balance, distribution shift, and the choice of evaluation metric [9, 10, 11, 12, 13, 14, 15, 16]. High overall accuracy can hide weak minority-class behavior, while well-separated development data may not represent the feature drift encountered after deployment. Probability calibration is equally important when model confidence is used to support escalation or automated response.

This study evaluates standard classifiers under one controlled protocol rather than proposing a new detection model. The analysis addresses three practical questions: (i) how conventional classifiers compare under an imbalanced eight-class benchmark; (ii) whether their ranking remains stable under moderate feature drift; and (iii) how probability calibration and feature reduction affect operational metrics. The numerical results are produced by the accompanying reproducible Python experiment.

2. Related Work

2.1. IoT intrusion detection and traffic learning

IoT security research has considered device fingerprinting, behavioral profiling, network-flow classification, botnet detection, and general intrusion detection [6, 7, 8, 74, 75, 76, 80]. Multimodal and representation-learning studies further show that combining heterogeneous features can improve discrimination when different data sources capture

ORCID(s):

complementary behavior [9, 10, 11]. However, dataset bias and closed-world assumptions remain recurring concerns, especially when training and deployment conditions differ [75, 76].

2.2. Identity, trust, and connected-system context

Connected-system security also depends on identity assurance, authentication, and device trust. Foundational biometric research addresses fingerprint recognition, template protection, fuzzy commitment, and cryptographic binding [17, 18, 19, 20, 21, 23, 24, 25, 26, 27]. In this context, Zita and Khalil presented a fingerprint-based cryptographic identity pipeline that connects recognition with key-pair generation [22]. Broader adaptive and resource-aware operation has been studied in swarm robotics, explainable AI, and self-adaptive IoT systems [28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40]. Zita and Zita examined explainable and energy-aware adaptive IoT operation using Deep Q-Learning, providing an application-oriented example of combining runtime adaptation with explanation mechanisms [81].

2.3. Supply-chain and operational resilience

IoT and communication security further depend on software provenance, update practices, supplier risk, and operational resilience [41, 42, 43, 44, 45, 46, 48, 49, 50, 70]. Zita and Zita discussed supply-chain risks affecting communication devices in high-risk environments [47]. Zero-trust access, secure token handling, authenticated encryption, and manufacturer-support guidance complement these concerns [67, 68, 69, 71, 72, 73].

2.4. Uncertainty, calibration, and trustworthy machine learning

Probabilistic modeling and uncertainty estimation provide tools for interpreting model confidence and supporting risk-sensitive decisions [51, 52, 53, 54, 55, 56, 57, 58, 59]. Conformal prediction provides an additional framework for finite-sample uncertainty statements under explicit assumptions [60, 61]. Trustworthy deployment also requires attention to technical debt, dataset documentation, model reporting, and AI risk management [62, 63, 64, 65, 66]. Adversarial machine-learning research reinforces the need to test models beyond benign development conditions [77, 78, 79].

3. Materials and Methods

3.1. Controlled benchmark

A controlled eight-class benchmark was generated independently for three random seeds (17, 41, and 71). Each development dataset contained 18,000 observations and 28 continuous traffic-like features. Twenty features were informative, four were redundant, and the remaining features supplied additional nuisance variation. Class proportions were deliberately imbalanced: 30%, 18%, 13%, 11%, 9%, 8%, 6%, and 5%. A stratified 75/25 split produced training and development-test partitions.

The classes represent generic operating and attack categories rather than reproducing a specific public dataset. This design isolates the effects of imbalance, calibration, feature reduction, and controlled feature drift without claiming that the generated observations are packet traces from deployed devices.

3.2. Feature-drift evaluation

For each seed, 5,000 observations were resampled from the held-out development test set and transformed to create a shifted evaluation set. The transformation applied small feature-specific scaling and offsets, additive Gaussian noise, and stronger offsets to every fourth feature. Labels were unchanged. The shifted set therefore measures sensitivity to moderate covariate drift rather than to a change in the class taxonomy.

3.3. Models

Five standard configurations were evaluated: class-weighted logistic regression, random forest, Extra Trees, histogram gradient boosting, and sigmoid-calibrated Extra Trees. Tree ensembles used balanced class weighting where supported. The calibrated model fitted Extra Trees inside three-fold calibration. No architecture or classifier proposed in this study is new; all models are standard implementations used to provide a common comparison.

3.4. Metrics

Overall accuracy was reported together with balanced accuracy, macro-F1, weighted F1, Matthews correlation coefficient (MCC), expected calibration error (ECE), and multiclass log loss. Balanced accuracy and macro-F1 were emphasized because the benchmark is intentionally imbalanced. ECE was computed using 12 confidence bins.

Table 1
Development-test results across three independent seeds (mean ± standard deviation).

Model	Bal. Acc.	Macro-F1	MCC	ECE
Logistic regression	0.978 ± 0.005	0.976 ± 0.005	0.976 ± 0.005	0.017 ± 0.002
Random forest	0.973 ± 0.004	0.979 ± 0.003	0.978 ± 0.004	0.153 ± 0.004
Extra Trees	0.978 ± 0.005	0.983 ± 0.004	0.982 ± 0.005	0.182 ± 0.006
Histogram gradient boosting	0.978 ± 0.005	0.982 ± 0.003	0.982 ± 0.004	0.006 ± 0.001
Calibrated Extra Trees	$\mathbf{0.981 \pm 0.005}$	$\mathbf{0.984 \pm 0.004}$	$\mathbf{0.983 \pm 0.005}$	0.014 ± 0.001

Table 2
Shifted-set results across three seeds.

Model	Bal. Acc.	Macro-F1	MCC	ECE
Logistic regression	0.976 ± 0.006	0.974 ± 0.005	0.974 ± 0.006	0.020 ± 0.004
Random forest	0.969 ± 0.006	0.976 ± 0.005	0.975 ± 0.007	0.157 ± 0.006
Extra Trees	0.976 ± 0.006	0.981 ± 0.005	0.981 ± 0.006	0.187 ± 0.008
Histogram gradient boosting	0.974 ± 0.006	0.979 ± 0.006	0.979 ± 0.006	$\mathbf{0.006 \pm 0.002}$
Calibrated Extra Trees	$\mathbf{0.981 \pm 0.006}$	$\mathbf{0.983 \pm 0.005}$	$\mathbf{0.983 \pm 0.007}$	0.015 ± 0.004

3.5. Feature ablation

Extra Trees feature importance was used only to rank features within each training split. Separate models were then trained with all 28 features, the top 18, and the top 10. The same feature subsets were evaluated on development and shifted data. Feature selection never used the shifted labels.

4. Results

4.1. Development-set performance

Table 1 summarizes the repeated results. All five models performed strongly, but calibrated Extra Trees obtained the highest mean balanced accuracy and macro-F1. The difference between calibrated and uncalibrated Extra Trees in discrimination was small, whereas the calibration difference was substantial.

The calibrated Extra Trees model reached mean accuracy 0.986 and weighted F1 0.986. Histogram gradient boosting produced the lowest ECE (0.006), followed by calibrated Extra Trees (0.014) and logistic regression (0.017). In contrast, uncalibrated Extra Trees and random forest had substantially larger ECE despite comparable discrimination.

4.2. Performance under feature drift

Moderate feature drift produced only small changes in ranking and absolute performance (Table 2). Calibrated Extra Trees retained balanced accuracy of 0.981 ± 0.006 and macro-F1 of 0.983 ± 0.005. Histogram gradient boosting retained the lowest ECE under drift.

4.3. Effect of probability calibration

Calibration primarily improved probability quality rather than class discrimination. On the development split, Extra Trees ECE decreased from 0.182 ± 0.006 to 0.014 ± 0.001 after calibration, while macro-F1 changed from 0.983 ± 0.004 to 0.984 ± 0.004. Log loss similarly decreased from approximately 0.314 to 0.072. The same pattern remained under feature drift. This result is relevant when classifier probabilities are used for prioritization or response escalation rather than only for hard labels.

4.4. Feature-ablation results

Table 3 shows that reducing the representation from 28 to 18 features produced only a small decrease, whereas the 10-feature configuration lost approximately five macro-F1 points. The relative behavior was similar under drift.

Table 3
Extra Trees feature-ablation results.

Split	Features	Macro-F1	Balanced accuracy
Development	28	0.982 ± 0.003	0.977 ± 0.003
Development	18	0.978 ± 0.002	0.974 ± 0.002
Development	10	0.931 ± 0.012	0.921 ± 0.013
Shifted	28	0.980 ± 0.004	0.974 ± 0.005
Shifted	18	0.976 ± 0.003	0.971 ± 0.004
Shifted	10	0.928 ± 0.012	0.919 ± 0.012

5. Discussion

The experiment supports three straightforward observations. First, strong development accuracy alone does not describe the operational properties of a classifier. Balanced accuracy and macro-F1 are necessary when minority classes are important, while ECE and log loss reveal whether confidence estimates are usable. Second, calibration can improve probability quality substantially without materially changing hard-label discrimination. Third, moderate feature reduction can preserve most classification performance, but aggressive reduction may remove information needed by minority classes.

The strong results should be interpreted in the context of the benchmark design. The controlled data are more regular than many real network traces, and the shifted set preserves the class semantics. Therefore, the study does not claim that the reported values transfer directly to CICIoT2023, Edge-IIoTset, or any production IoT deployment. The purpose is a reproducible comparison of evaluation dimensions under known conditions.

These observations are consistent with broader concerns in trustworthy machine learning: dataset documentation, calibration, model reporting, uncertainty, adversarial robustness, and deployment drift should be considered alongside nominal predictive performance [62, 63, 64, 65, 77, 78, 79].

6. Limitations

The study uses a controlled synthetic benchmark rather than packet traces from physical IoT devices. The feature distributions do not model protocol semantics, temporal packet sequences, device-specific identifiers, or a complete adversarial campaign. Feature drift was implemented as covariate perturbation and therefore does not represent concept drift or previously unseen attack families. Only classical tabular classifiers were evaluated. Finally, three random seeds provide repeated measurements but are not a substitute for external validation across independently collected datasets.

Future evaluation should repeat the protocol on public IoT intrusion datasets, use device-disjoint and time-disjoint splits where possible, and examine stronger shifts including unseen attack categories. Those extensions are validation steps rather than claims of a new detection method.

7. Reproducibility and Data Availability

The accompanying Python script generates the benchmark, performs all train/test splits, applies the feature-drift transformation, trains the five model configurations, and writes the exact numerical outputs used in the tables. The three seeds are fixed to 17, 41, and 71. No human-subject or private data are used.

8. Conclusion

This study compared five standard machine-learning configurations for imbalanced eight-class IoT intrusion detection under development and feature-drift conditions. Calibrated Extra Trees achieved the strongest mean balanced accuracy and macro-F1, while histogram gradient boosting provided the lowest calibration error. Probability calibration substantially improved Extra Trees confidence quality with little change in classification performance. Feature ablation showed that 18 of 28 features retained nearly all predictive performance, whereas a 10-feature representation caused a clear reduction. The results illustrate the value of evaluating discrimination, calibration, imbalance, and drift together when reporting IoT intrusion-detection performance. No new classifier, architecture, or security framework is claimed.

References

[1] Sicari, Sabrina and Rizzardi, Alessandra and Grieco, Luigi Alfredo and Coen-Porisini, Alberto. "Security, Privacy and Trust in Internet of Things: The Road Ahead." *Computer Networks* vol. 76 pp. 146–164 2015 doi: https://doi.org/10.1016/j.comnet.2014.11.008.

[2] Mosenia, Arsalan and Jha, Niraj K.. "A Comprehensive Study of Security of Internet-of-Things." *IEEE Transactions on Emerging Topics in Computing* vol. 5, no. 4 pp. 586–602 2017 doi: https://doi.org/10.1109/TETC.2016.2606384.

[3] Alaba, Fadele Ayotunde and Othman, Mazliza and Hashem, Ibrahim Abaker Targio and Alotaibi, Faiz. "Internet of Things Security: A Survey." *Journal of Network and Computer Applications* vol. 88 pp. 10–28 2017 doi: https://doi.org/10.1016/j.jnca.2017.04.002.

[4] Neshenko, Nataliia and Bou-Harb, Elias and Crichigno, Jorge and Kaddoum, Georges and Ghani, Nasir. "Demystifying IoT Security: An Exhaustive Survey on IoT Vulnerabilities and a First Empirical Look on Internet-Scale IoT Exploitations." *IEEE Communications Surveys & Tutorials* vol. 21, no. 3 pp. 2702–2733 2019 doi: https://doi.org/10.1109/COMST.2019.2910750.

[5] Antonakakis, Manos and April, Tim and Bailey, Michael and Bernhard, Matt and Bursztein, Elie and Cochran, Jaime and Durumeric, Zakir and Halderman, J. Alex and Invernizzi, Luca and Kallitsis, Michalis and Kumar, Deepak and Lever, Chaz and Ma, Zane and Mason, Joshua and Menscher, Damian and Seaman, Chad and Sullivan, Nick and Thomas, Kurt and Zhou, Yi. "Understanding the Mirai Botnet." *26th USENIX Security Symposium* pp. 1093–1110 2017.

[6] Meidan, Yair and Bohadana, Michael and Mathov, Yael and Mirsky, Yisroel and Shabtai, Asaf and Breitenbacher, Dominik and Elovici, Yuval. "N-BaIoT: Network-Based Detection of IoT Botnet Attacks Using Deep Autoencoders." *IEEE Pervasive Computing* vol. 17, no. 3 pp. 12–22 2018 doi: https://doi.org/10.1109/MPRV.2018.03367731.

[7] Doshi, Rohan and Apthorpe, Noah and Feamster, Nick. "Machine Learning DDoS Detection for Consumer Internet of Things Devices." *2018 IEEE Security and Privacy Workshops* pp. 29–35 2018 doi: https://doi.org/10.1109/SPW.2018.00013.

[8] Sivanathan, Arunan and Sherratt, Daniel and Gharakheili, Hassan Habibi and Radford, Adam and Wijenayake, Chamith and Vishwanath, Arun and Sivaraman, Vijay. "Classifying IoT Devices in Smart Environments Using Network Traffic Characteristics." *IEEE Transactions on Mobile Computing* vol. 18, no. 8 pp. 1745–1759 2019 doi: https://doi.org/10.1109/TMC.2018.2866249.

[9] Baltrušaitis, Tadas and Ahuja, Chaitanya and Morency, Louis-Philippe. "Multimodal Machine Learning: A Survey and Taxonomy." *IEEE Transactions on Pattern Analysis and Machine Intelligence* vol. 41, no. 2 pp. 423–443 2019 doi: https://doi.org/10.1109/TPAMI.2018.2798607.

[10] Ngiam, Jiquan and Khosla, Aditya and Kim, Mingyu and Nam, Juhan and Lee, Honglak and Ng, Andrew Y.. "Multimodal Deep Learning." *Proceedings of the 28th International Conference on Machine Learning* pp. 689–696 2011.

[11] Ramachandram, Dhanesh and Taylor, Graham W.. "Deep Multimodal Learning: A Survey on Recent Advances and Trends." *IEEE Signal Processing Magazine* vol. 34, no. 6 pp. 96–108 2017 doi: https://doi.org/10.1109/MSP.2017.2738401.

[12] Chen, Tianqi and Guestrin, Carlos. "XGBoost: A Scalable Tree Boosting System." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* pp. 785–794 2016 doi: https://doi.org/10.1145/2939672.2939785.

[13] Ke, Guolin and Meng, Qi and Finley, Thomas and Wang, Taifeng and Chen, Wei and Ma, Weidong and Ye, Qiwei and Liu, Tie-Yan. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." *Advances in Neural Information Processing Systems* vol. 30 2017.

[14] Chawla, Nitesh V. and Bowyer, Kevin W. and Hall, Lawrence O. and Kegelmeyer, W. Philip. "SMOTE: Synthetic Minority Over-Sampling Technique." *Journal of Artificial Intelligence Research* vol. 16 pp. 321–357 2002 doi: https://doi.org/10.1613/jair.953.

[15] Gama, João and Žliobaitė, Indrė and Bifet, Albert and Pechenizkiy, Mykola and Bouchachia, Abdelhamid. "A Survey on Concept Drift Adaptation." *ACM Computing Surveys* vol. 46, no. 4 pp. 1–37 2014 doi: https://doi.org/10.1145/2523813.

[16] Guo, Chuan and Pleiss, Geoff and Sun, Yu and Weinberger, Kilian Q.. "On Calibration of Modern Neural Networks." *Proceedings of the 34th International Conference on Machine Learning* vol. 70 pp. 1321–1330 2017.

[17] Jain, Anil K. and Ross, Arun and Prabhakar, Salil. "An Introduction to Biometric Recognition." *IEEE Transactions on Circuits and Systems for Video Technology* vol. 14, no. 1 pp. 4–20 2004 doi: https://doi.org/10.1109/TCSVT.2003.818349.

[18] Ratha, Nalini K. and Connell, Jonathan H. and Bolle, Ruud M.. "Enhancing Security and Privacy in Biometrics-Based Authentication Systems." *IBM Systems Journal* vol. 40, no. 3 pp. 614–634 2001 doi: https://doi.org/10.1147/sj.403.0614.

[19] Jain, Anil K. and Nandakumar, Karthik and Nagar, Abhishek. "Biometric Template Security." *EURASIP Journal on Advances in Signal Processing* vol. 2008 pp. 579416 2008 doi: https://doi.org/10.1155/2008/579416.

[20] Juels, Ari and Wattenberg, Martin. "A Fuzzy Commitment Scheme." *Proceedings of the 6th ACM Conference on Computer and Communications Security* pp. 28–36 1999 doi: https://doi.org/10.1145/319709.319714.

[21] Dodis, Yevgeniy and Ostrovsky, Rafail and Reyzin, Leonid and Smith, Adam. "Fuzzy Extractors: How to Generate Strong Keys from Biometrics and Other Noisy Data." *SIAM Journal on Computing* vol. 38, no. 1 pp. 97–139 2008 doi: https://doi.org/10.1137/060651380.

[22] Zita, Wail and Khalil, Mohammed Sayim. "Fingerprint-Based Cryptographic Identity: A Custom Recognition Pipeline with Key Pair Generation." *East Journal of Computer Science* vol. 1, no. 3 pp. 17–34 2025 doi: https://doi.org/10.63496/ejcs.Vol1.Iss3.112.

[23] Nandakumar, Karthik and Jain, Anil K. and Pankanti, Sharath. "Fingerprint-Based Fuzzy Vault: Implementation and Performance." *IEEE Transactions on Information Forensics and Security* vol. 2, no. 4 pp. 744–757 2007 doi: https://doi.org/10.1109/TIFS.2007.908165.

[24] Patel, Vishal M. and Chellappa, Rama and Chandra, Deepak and Barbello, Brandon. "Continuous User Authentication on Mobile Devices: Recent Progress and Remaining Challenges." *IEEE Signal Processing Magazine* vol. 33, no. 4 pp. 49–61 2016 doi: https://doi.org/10.1109/MSP.2016.2555335.

[25] Frank, Mario and Biedert, Ralf and Ma, Eugene and Martinovic, Ivan and Song, Dawn. "Touchalytics: On the Applicability of Touchscreen Input as a Behavioral Biometric for Continuous Authentication." *IEEE Transactions on Information Forensics and Security* vol. 8, no. 1 pp. 136–148 2013 doi: https://doi.org/10.1109/TIFS.2012.2225048.

[26] ISO/IEC. "ISO/IEC 24745:2022 Information Security, Cybersecurity and Privacy Protection – Biometric Information Protection." 2022 International Organization for Standardization.

[27] Temoshok, David and Choong, Yee-Yin and Fenton, James and Lefkovitz, Naomi and Regenscheid, Andrew and Richer, Justin and Theofanos, Mary and Galluzzo, Ryan. "Digital Identity Guidelines: Authentication and Authenticator Management." 2025 NIST Special Publication 800-63B-4 doi: https://doi.org/10.6028/NIST.SP.800-63B-4.

[28] Şahin, Erol. "Swarm Robotics: From Sources of Inspiration to Domains of Application." *Swarm Robotics* vol. 3342 pp. 10–20 2005 doi: https://doi.org/10.1007/978-3-540-30552-1_2.

[29] Brambilla, Manuele and Ferrante, Eliseo and Birattari, Mauro and Dorigo, Marco. "Swarm Robotics: A Review from the Swarm Engineering Perspective." *Swarm Intelligence* vol. 7 pp. 1–41 2013 doi: https://doi.org/10.1007/s11721-012-0075-2.

[30] Rubenstein, Michael and Ahler, Christian and Nagpal, Radhika. "Kilobot: A Low Cost Scalable Robot System for Collective Behaviors." *2012 IEEE International Conference on Robotics and Automation* pp. 3293–3298 2012 doi: https://doi.org/10.1109/ICRA.2012.6224638.

[31] Rubenstein, Michael and Cornejo, Alejandro and Nagpal, Radhika. "Programmable Self-Assembly in a Thousand-Robot Swarm." *Science* vol. 345, no. 6198 pp. 795–799 2014 doi: https://doi.org/10.1126/science.1254295.

[32] Bayındır, Levent. "A Review of Swarm Robotics Tasks." *Neurocomputing* vol. 172 pp. 292–321 2016 doi: https://doi.org/10.1016/j.neucom.2015.05.116.

[33] Dorigo, Marco and Floreano, Dario and Gambardella, Luca M. and Mondada, Francesco and Nolfi, Stefano and Baaboura, Tarek and Birattari, Mauro and Bonani, Michael and Brambilla, Manuele and Brutschy, Arne and others. "Swarmanoid: A Novel Concept for the Study of Heterogeneous Robotic Swarms." *IEEE Robotics & Automation Magazine* vol. 20, no. 4 pp. 60–71 2013 doi: https://doi.org/10.1109/MRA.2013.2252996.

[34] Hamann, Heiko. "Swarm Robotics: A Formal Approach." 2018 Springer doi: https://doi.org/10.1007/978-3-319-74528-2.

[35] Ribeiro, Marco Tulio and Singh, Sameer and Guestrin, Carlos. "Why Should I Trust You?: Explaining the Predictions of Any Classifier." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* pp. 1135–1144 2016 doi: https://doi.org/10.1145/2939672.2939778.

[36] Lundberg, Scott M. and Lee, Su-In. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems* vol. 30 2017.

[37] Adadi, Amina and Berrada, Mohammed. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE Access* vol. 6 pp. 52138–52160 2018 doi: https://doi.org/10.1109/ACCESS.2018.2870052.

[38] Barredo Arrieta, Alejandro and Díaz-Rodríguez, Natalia and Del Ser, Javier and Bennetot, Adrien and Tabik, Siham and Barbado, Alberto and García, Salvador and Gil-López, Sergio and Molina, Daniel and Benjamins, Richard and Chatila, Raja and Herrera, Francisco. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI." *Information Fusion* vol. 58 pp. 82–115 2020 doi: https://doi.org/10.1016/j.inffus.2019.12.012.

[39] Rudin, Cynthia. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* vol. 1 pp. 206–215 2019 doi: https://doi.org/10.1038/s42256-019-0048-x.

[40] Molnar, Christoph. "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable." 2022 Online book.

[41] Boyens, Jon and Smith, Angela and Bartol, Nadya and Winkler, Kris and Holbrook, Alex and Fallon, Matthew. "Cybersecurity Supply Chain Risk Management Practices for Systems and Organizations." no. NIST SP 800-161 Rev. 1 Update 1 2024 doi: https://doi.org/10.6028/NIST.SP.800-161r1-upd1.

[42] European Union Agency for Cybersecurity. "ENISA Threat Landscape for Supply Chain Attacks." 2021 doi: https://doi.org/10.2824/168593.

[43] Torres-Arias, Santiago and Afzali, Hammad and Kuppusamy, Trishank Karthik and Curtmola, Reza and Cappos, Justin. "in-toto: Providing Farm-to-Table Guarantees for Bits and Bytes." *28th USENIX Security Symposium* pp. 1393–1410 2019.

[44] Cappos, Justin and Samuel, Justin and Baker, Scott and Hartman, John H.. "A Look in the Mirror: Attacks on Package Managers." *Proceedings of the 15th ACM Conference on Computer and Communications Security* pp. 565–574 2008 doi: https://doi.org/10.1145/1455770.1455841.

[45] Zimmermann, Markus and Staicu, Cristian-Alexandru and Tenny, Cam and Pradel, Michael. "Small World with High Risks: A Study of Security Threats in the npm Ecosystem." *28th USENIX Security Symposium* pp. 995–1010 2019.

[46] Ohm, Marc and Plate, Henrik and Sykosch, Arnold and Meier, Michael. "Backstabber's Knife Collection: A Review of Open Source Software Supply Chain Attacks." *Detection of Intrusions and Malware, and Vulnerability Assessment* vol. 12223 pp. 23–43 2020 doi: https://doi.org/10.1007/978-3-030-52683-2_2.

[47] Zita, Wail and Zita, Nassma. "Mitigating Supply Chain Risks in Communication Devices Across Conflict Zones." 2025 TechRxiv preprint. doi: 10.36227/techrxiv.175234514.44631243/v1.

[48] Souppaya, Murugiah and Scarfone, Karen and Dodson, Donna. "Secure Software Development Framework (SSDF) Version 1.1: Recommendations for Mitigating the Risk of Software Vulnerabilities." no. SP 800-218 2022 doi: https://doi.org/10.6028/NIST.SP.800-218.

[49] National Telecommunications and Information Administration. "The Minimum Elements for a Software Bill of Materials (SBOM)." 2021.

[50] Boyens, Jon and Paulsen, Celia and Moorthy, Rama and Bartol, Nadya. "Key Practices in Cyber Supply Chain Risk Management: Observations from Industry." no. IR 8276 2021 doi: https://doi.org/10.6028/NIST.IR.8276.

[51] Bishop, Christopher M.. "Pattern Recognition and Machine Learning." 2006 Springer.

[52] Murphy, Kevin P.. "Machine Learning: A Probabilistic Perspective." 2012 MIT Press.

[53] Gelman, Andrew and Carlin, John B. and Stern, Hal S. and Dunson, David B. and Vehtari, Aki and Rubin, Donald B.. "Bayesian Data Analysis." 2013 CRC Press.

[54] MacKay, David J. C.. "A Practical Bayesian Framework for Backpropagation Networks." *Neural Computation* vol. 4, no. 3 pp. 448–472 1992 doi: https://doi.org/10.1162/neco.1992.4.3.448.

[55] Tipping, Michael E.. "Sparse Bayesian Learning and the Relevance Vector Machine." *Journal of Machine Learning Research* vol. 1 pp. 211–244 2001.

[56] Gal, Yarin and Ghahramani, Zoubin. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning." *Proceedings of the 33rd International Conference on Machine Learning* vol. 48 pp. 1050–1059 2016.

[57] Kendall, Alex and Gal, Yarin. "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?." *Advances in Neural Information Processing Systems* vol. 30 2017.

[58] Blei, David M. and Kucukelbir, Alp and McAuliffe, Jon D.. "Variational Inference: A Review for Statisticians." *Journal of the American Statistical Association* vol. 112, no. 518 pp. 859–877 2017 doi: https://doi.org/10.1080/01621459.2017.1285773.

[59] Breiman, Leo. "Random Forests." *Machine Learning* vol. 45 pp. 5–32 2001 doi: https://doi.org/10.1023/A:1010933404324.

[60] Angelopoulos, Anastasios N. and Bates, Stephen. "Conformal Prediction: A Gentle Introduction." *Foundations and Trends in Machine Learning* vol. 16, no. 4 pp. 494–591 2023 doi: https://doi.org/10.1561/2200000101.

[61] Vovk, Vladimir and Gammerman, Alexander and Shafer, Glenn. "Algorithmic Learning in a Random World." 2005 Springer.

[62] Sculley, D. and Holt, Gary and Golovin, Daniel and Davydov, Eugene and Phillips, Todd and Ebner, Dietmar and Chaudhary, Vinay and Young, Michael and Crespo, Jean-François and Dennison, Dan. "Hidden Technical Debt in Machine Learning Systems." *Advances in Neural Information Processing Systems* vol. 28 2015.

[63] Gebru, Timnit and Morgenstern, Jamie and Vecchione, Briana and Vaughan, Jennifer Wortman and Wallach, Hanna and Daumé III, Hal and Crawford, Kate. "Datasheets for Datasets." *Communications of the ACM* vol. 64, no. 12 pp. 86–92 2021 doi: https://doi.org/10.1145/3458723.

[64] Mitchell, Margaret and Wu, Simone and Zaldivar, Andrew and Barnes, Parker and Vasserman, Lucy and Hutchinson, Ben and Spitzer, Elena and Raji, Inioluwa Deborah and Gebru, Timnit. "Model Cards for Model Reporting." *Proceedings of the Conference on Fairness, Accountability, and Transparency* pp. 220–229 2019 doi: https://doi.org/10.1145/3287560.3287596.

[65] Tabassi, Elham. "Artificial Intelligence Risk Management Framework (AI RMF 1.0)." no. NIST AI 100-1 2023 doi: https://doi.org/10.6028/NIST.AI.100-1.

[66] Snyder, Hannah. "Literature Review as a Research Methodology: An Overview and Guidelines." *Journal of Business Research* vol. 104 pp. 333–339 2019 doi: https://doi.org/10.1016/j.jbusres.2019.07.039.

[67] Rose, Scott and Borchert, Oliver and Mitchell, Stu and Connelly, Sean. "Zero Trust Architecture." no. NIST SP 800-207 2020 doi: https://doi.org/10.6028/NIST.SP.800-207.

[68] Fagan, Michael and Megas, Katerina N. and Scarfone, Karen and Smith, Matthew. "IoT Device Cybersecurity Capability Core Baseline." no. NIST IR 8259A 2020 doi: https://doi.org/10.6028/NIST.IR.8259A.

[69] Fagan, Michael and Megas, Katerina N. and Marron, Jeffrey and Brady, Kevin and Cuthill, Barbara and Herold, Rebecca. "IoT Non-Technical Supporting Capability Core Baseline." no. NIST IR 8259B 2021 doi: https://doi.org/10.6028/NIST.IR.8259B.

[70] Ross, Ron and Pillitteri, Victoria and Graubart, Richard and Bodeau, Deborah and McQuaid, Rosalie. "Developing Cyber-Resilient Systems: A Systems Security Engineering Approach." no. NIST SP 800-160 Vol. 2 Rev. 1 2021 doi: https://doi.org/10.6028/NIST.SP.800-160v2r1.

[71] Jones, Michael and Bradley, John and Sakimura, Nat. "JSON Web Token (JWT)." no. RFC 7519 2015.

[72] Sheffer, Yaron and Hardt, Dick and Jones, Michael. "JSON Web Token Best Current Practices." no. RFC 8725 2020.

[73] Dworkin, Morris. "Recommendation for Block Cipher Modes of Operation: Galois/Counter Mode (GCM) and GMAC." no. NIST SP 800-38D 2007 doi: https://doi.org/10.6028/NIST.SP.800-38D.

[74] Buczak, Anna L. and Guven, Erhan. "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection." *IEEE Communications Surveys & Tutorials* vol. 18, no. 2 pp. 1153–1176 2016 doi: https://doi.org/10.1109/COMST.2015.2494502.

[75] Sommer, Robin and Paxson, Vern. "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection." *2010 IEEE Symposium on Security and Privacy* pp. 305–316 2010 doi: https://doi.org/10.1109/SP.2010.25.

[76] Ring, Markus and Wunderlich, Sarah and Scheuring, Deniz and Landes, Dieter and Hotho, Andreas. "A Survey of Network-Based Intrusion Detection Data Sets." *Computers & Security* vol. 86 pp. 147–167 2019 doi: https://doi.org/10.1016/j.cose.2019.06.005.

[77] Goodfellow, Ian J. and Shlens, Jonathon and Szegedy, Christian. "Explaining and Harnessing Adversarial Examples." *International Conference on Learning Representations* 2015.

[78] Biggio, Battista and Roli, Fabio. "Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning." *Pattern Recognition* vol. 84 pp. 317–331 2018 doi: https://doi.org/10.1016/j.patcog.2018.07.023.

[79] Papernot, Nicolas and McDaniel, Patrick and Goodfellow, Ian. "Transferability in Machine Learning: From Phenomena to Black-Box Attacks Using Adversarial Samples." *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security* pp. 506–519 2017 doi: https://doi.org/10.1145/3052973.3053009.

[80] Shone, Nathan and Ngoc, Tran Nguyen and Phai, Vu Dinh and Shi, Qi. "A Deep Learning Approach to Network Intrusion Detection." *IEEE Transactions on Emerging Topics in Computational Intelligence* vol. 2, no. 1 pp. 41–50 2018 doi: https://doi.org/10.1109/TETCI.2017.2772792.

[81] Zita, Wail and Zita, Nassma. "Explainable and Energy-Aware Self-Adaptive IoT Architecture Using Deep Q-Learning." *Sakarya University Journal of Computer and Information Sciences* no. Advanced Online Publication pp. 364–376 2026 doi: https://doi.org/10.35377/saucis...1736606.