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.

Beyond Accuracy: An Eight-Criterion Framework for Selecting Detection Algorithms, with a Weight-Sensitivity Analysis

Nana Kwame Gyamfi

nana-kwame.gyamfi@vilniustech.lt

Vilnius Gediminas Technical University

Research Article

Keywords: Algorithm selection, Multi-criteria evaluation, Signal-to-noise ratio, Intrusion detection, Malware detection, Evaluation protocol

Posted Date: November 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11640847/v1

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

Additional Declarations: No competing interests reported.

Abstract

Where several detectors classify comparably, a practitioner must still choose one, and the literature offers no procedure beyond comparing headline accuracies. A composite metric published previously by the author combines four measures that any training run produces, on the decibel scale of the signal-to-noise ratio, and discriminates where accuracy cannot. This paper extends that metric into a framework and reports what the extension reveals. The framework adds five criteria to the published four: imbalance robustness, recall against a defined adversary, explanation faithfulness, inference latency and memory footprint. Each is mapped to a common scale by the same construction, weights are elicited from the deployment rather than assumed, and a sampling analysis reports how far the recommendation depends on them. An evaluation protocol is specified in which each requirement is justified by a measurement rather than by appeal to good practice. Applied to six classifiers across four settings spanning two corpora, two domains and two tasks, the framework selects the same detector in every setting, while the detector leading on accuracy differs in each. Under equal weights the selected detector leads by 1.9 decibels; under a weighting favouring constrained hardware, by 1.6; under a weighting favouring security operations the field compresses to within three decibels, reproducing through weighting the same indistinguishability that an analysis of variance found in the published study. Across two thousand sampled weightings the selected detector leads in 99.1 per cent. Three of the eight criteria were held constant in this application because they have not yet been measured, so the sampling analysis is effectively over five; this is stated rather than concealed, and the paper specifies how the remaining three should be measured.

1 Introduction and relation to prior work

This paper extends work the author published in 2024 [1], and the relationship should be clear from the outset. That paper derived a composite of four quantities that any training run produces, namely mean absolute error, root mean squared error, correctly classified instances and model-building time, placed them on the decibel scale of the signal-to-noise ratio where quantities in incompatible units become additive, and showed that the composite ranks detectors consistently where an analysis of variance on correct classification could not distinguish them.

Four things in the present paper are new. *Five further criteria*, covering imbalance robustness, recall against a defined adversary, explanation faithfulness, inference latency and memory footprint, with a specification of how each is measured. *Explicit weights* elicited from a deployment profile rather than assumed equal. *A sampling analysis* reporting the share of weightings under which the recommendation holds. And *an evaluation protocol* whose requirements are each justified by a measurement reported here or in the supporting work, rather than by appeal to good practice.

The motivation is unchanged and is worth restating. The no-free-lunch theorems establish that no algorithm dominates across problems [2], so selection must be empirical; procedures for comparing classifiers across datasets exist [3] but are seldom applied in detection; and accuracy is a poor summary

under imbalance [4,5]. What is missing is not a measure but a procedure that combines measures a practitioner actually weighs.

2 Related work

2.1 Selecting among algorithms

That no algorithm dominates across problems is formalised by the no-free-lunch theorems [2], so selection is necessarily empirical. Statistical procedures for comparing classifiers across datasets are given by Demšar [3] and refined by García and Herrera [6] and by Benavoli and colleagues, who argue for Bayesian alternatives to null-hypothesis testing in this setting [7]. Algorithm selection as a learning problem in its own right descends from Rice [8] and is surveyed under the heading of meta-learning [9,10]. None of that literature combines effectiveness with deployment cost on a single scale, which is what a practitioner requires.

2.2 Multi-criteria decision procedures

Where several criteria must be reconciled, the operations-research literature offers established procedures: the analytic hierarchy process [11], TOPSIS [12], and outranking methods [13]. Their common requirement, that weights be elicited and stated, is adopted here, as is the standard practice of reporting a sensitivity analysis over those weights [14].

2.3 Robust design and the signal-to-noise ratio

The composite extended here derives from Taguchi’s parameter design, in which a quality characteristic is judged by a ratio combining location and dispersion [15,16]. The approach is not without criticism: Box argued that such ratios can conceal the separate contributions of mean and variance and are appropriate only under particular mean–variance relationships [17], and Nair and colleagues set out the ensuing debate [18]. The response adopted here is not to dismiss the objection but to report components alongside every composite, so that the separate analysis Box prefers remains available.

2.4 The criteria the framework adds

Imbalance robustness.

Accuracy flatters majority-class predictors, and the Matthews correlation coefficient is preferred [4]; precision–recall analysis is more informative than the receiver operating characteristic when the positive class is rare [19,20].

Recall under attack.

Detectors can be evaded by modifying inputs, an idea developed for sequence-based host detection by Wagner and Soto [21] and generalised as adversarial machine learning [22,23]. Problem-space

constraints on realisable attacks are set out by Pierazzi and colleagues [24], and evasion specifically against malware classifiers is demonstrated by Grosse and colleagues [25].

Explanation faithfulness.

Attribution methods such as LIME [26] and SHAP [27] are widely applied, but their faithfulness must itself be measured; deletion and insertion tests are the established instruments [28,29], and Warnecke and colleagues evaluate such methods specifically in security settings [30].

Cost.

Training and inference cost are reported inconsistently in this literature, although they determine whether a detector can be retrained or deployed at all; the point is made forcefully in the systems-benchmarking context by van der Kouwe and colleagues [31].

2.5 Evaluation protocol

The protocol in Section 4 draws on work documenting leakage and bias: the formulation of leakage by Kaufman and colleagues [32], its prevalence documented by Kapoor and Narayanan [33], blocked cross-validation for structured data [34], duplication effects in code corpora [35], and the security-specific catalogues of Arp and colleagues [36] and Pendlebury and colleagues [37]. Sommer and Paxson's earlier critique remains the clearest statement of why intrusion detection is a hard case for machine-learning evaluation [38].

3 The framework

3.1 Criteria

Table 1 lists the eight criteria, the direction in which each is better, and the justification for including it.

Table 1
The eight criteria and their justification

Criterion	Direction	Why included
Accuracy	Larger better	The measure the field reports; retained for comparability with current practice
Precision	Larger better	False alarms consume analyst time, the binding constraint in small teams
Matthews correlation	Larger better	On one corpus studied, a classifier at 57.3% accuracy attains 0.104 while another at similar accuracy attains 0.237: the first predicts the majority class
Recall under attack	Larger better	A detector evaluated only against unmodified samples is evaluated against the wrong threat
Explanation faithfulness	Larger better	An alert that cannot be explained cannot be acted upon or audited
Build time	Smaller better	Across six models on one window, build times span a factor of about 390 where accuracy spans 25 points
Inference latency	Smaller better	Determines whether a host can be monitored in real time
Memory footprint	Smaller better	Determines whether the detector runs on the hardware an institution owns

3.2 Mapping to a common scale

Each criterion is mapped to decibels by the target form appropriate to its direction. For a larger-the-better criterion with mean and standard deviation over repeated runs,

and for a smaller-the-better criterion,

The weighted composite is the sum of the criterion ratios weighted by the deployment’s priorities.

Two properties follow. *The ranking is invariant to the units in which a criterion is measured*: changing build time from seconds to milliseconds adds to every time ratio, shifting all composites equally and leaving the order unchanged. And *components are reported alongside the composite in every case*, which is the response to the standing objection that a composite ratio can conceal the separate contributions of mean and variance [17]: the objection is not answered by denial but by making the components visible so that the separate analysis remains available to the reader.

3.3 Measuring the two criteria the published metric lacked

Recall under attack is specified as recall against a defined adversary at a defined budget, not as a general claim. Three requirements follow from an inconclusive experiment reported in the supporting work: attack the raw trace rather than the feature vector, since perturbing features flatters pipelines whose early stages discard the manipulation; substitute rather than add, since addition dilutes a count vector

without reshaping it; and draw the substitution profile from genuine benign traces rather than from another malicious class. Reporting is a curve against budget, not a number.

Explanation faithfulness is specified by the deletion test: remove the features an explanation ranks highest, replace them with a baseline drawn from the training partition, and record the fall in predicted probability for the true class. The baseline must be stated, since the fall depends on it, and the curve must be reported over a range of, since an explanation may identify the most important feature correctly and rank the rest arbitrarily.

4 The protocol

The framework is inseparable from the protocol under which its measurements are taken, and each requirement in Table 2 is justified by a measurement rather than asserted.

Table 2
Protocol requirements and their evidence

Requirement	Evidence
Report the duplicate rate under the features used	97.1% in one system-call corpus, 54.8% in a network corpus from a different research group
Report the label-conflict ceiling	A previously published accuracy of 97.35% exceeds the ceiling of the corpus at every observation window
Partition with group awareness and name the grouping unit	Group-aware partitioning moved measured accuracy at a ten-call window from 89.4% to 58.1%
Fit preprocessing inside the training fold	Otherwise selection and scaling leak information across the partition
Report at least three seeded runs with variance	Fold standard deviations of 1.7 to 4.9 points under sequence grouping and 14 to 38 under host grouping
Report an imbalance-robust measure	See the Matthews correlation entry in Table 1
Report cost on stated hardware	Timings are never compared across environments
Run the proposed component against a control	Two ablations found credited components to cost 2.51 and 3.83 to 9.09 accuracy points

The first two requirements appear to be new to this literature. They are also the cheapest: both are computed from the feature matrix and the labels before any model is fitted, and neither requires knowledge of the method under review, which makes them usable by a reviewer as well as by an author.

5 Application

5.1 The composite across four settings

Six classifiers were evaluated in four settings: five-class attribution and binary detection on a system-call corpus, and detection and attribution on a network-flow corpus, under group-aware partitioning over three seeds throughout. Table 3 reports the composite.

Table 3
Composite by model and setting (dB)

Model	System calls		Network flows	
	Attrib.	Detect.	Detect.	Attrib.
RandomTree	55.61	52.78	53.83	42.91
J48-like	50.59	45.71	44.57	38.29
NaiveBayes	44.36	35.36	37.23	31.38
SVM-RBF	44.31	36.87	35.08	18.01
RandomForest	40.77	40.07	37.11	28.35
MLP	31.67	27.78	30.00	16.95

The same detector leads in all four settings, and leads on accuracy in none of them. A random forest leads accuracy on three settings and a support vector machine on the fourth. On network detection the support vector machine leads accuracy by 5.5 points and the Matthews coefficient by 0.128, at 10.03 seconds to build against 0.018, a factor of 557.

The framework is not a proxy for speed, and the counterexample is in the table. Naïve Bayes builds faster than the leader in two of the four settings and finishes third or fifth in all of them, because its accuracy components are far behind: on the system-call detection task it reaches 55.8 per cent accuracy while catching 22.2 per cent of malware.

5.2 Deployment profiles

Table 4
Weighted composite under three profiles (dB)

Model	Benchmarking	Security ops.	Constrained
RandomTree	29.4	33.0	28.4
J48-like	27.5	32.4	26.8
NaiveBayes	26.6	30.7	24.8
MLP	22.2	30.2	21.8
SVM-RBF	21.6	29.9	18.9
RandomForest	21.3	30.0	18.8

Under benchmarking and constrained hardware the ordering is identical and the margins are clear. Under security operations, which places three-quarters of the weight on accuracy, precision, imbalance robustness and recall under attack, the field compresses to within three decibels.

That compression is itself a result. It reproduces through weighting the same phenomenon the published study found statistically, where an analysis of variance on correct classification could not distinguish the three leading algorithms: when only the measures the field already reports are weighted, strong detectors become hard to tell apart. The framework recovers the problem it was built to address, which is evidence that it measures what it claims to.

5.3 Sensitivity to the weights

Two thousand weight vectors were drawn from the simplex and the share in which each detector ranks first recorded. RandomTree leads in 99.1 per cent, J48 in 0.8 per cent, and no other detector in any. The recommendation therefore does not depend on the practitioner’s priorities: any weighting that treats all criteria positively reaches the same answer.

That figure must be read with its qualification. Three of the eight criteria, namely recall under attack, explanation faithfulness and memory, were held constant across detectors because they have not been measured for these models. A constant criterion contributes identically to every detector and cannot change a ranking, so the sampling is effectively over five criteria. The honest statement is that the leader holds under 99.1 per cent of weightings *over the five criteria measured*, and that the influence of the remaining three is unknown.

There is reason to expect them to matter. A single randomised tree is plausibly more brittle under adversarial perturbation than a hundred-tree ensemble, and recall under attack carries a quarter of the weight in the security-operations profile. Measuring it could reorder the top of the table, and the framework would then be doing exactly the work it exists to do.

6 Discussion

The framework does not assert that cost matters more than accuracy. It asserts that both are measurable on a common scale, that the trade between them is a deployment question rather than a modelling one, and that stating the weights makes the trade auditable. Where a reader disagrees with a weighting, the sampling analysis tells them how much their disagreement would change the answer, which a single weighted score cannot.

The compression observed under the security-operations profile suggests a broader point. A field that reports only accuracy-derived measures will find its methods increasingly indistinguishable as they mature, and will then select between them on grounds it does not report. Making those grounds explicit is the purpose of the framework.

7 Limitations

1. Three of eight criteria are unmeasured for the detectors studied, so the application is effectively over five. Section 3.3 specifies the measurements required.
2. The framework weights criteria but does not model interactions between them; cost and robustness may not be independent.
3. The objection that a composite ratio can conceal the separate contributions of mean and variance is mitigated by component reporting rather than answered.
4. No comparison is made against established multi-criteria procedures such as TOPSIS ranking or Friedman average ranks, which would provide convergent validation.
5. The three weight profiles were specified by the author rather than elicited from practitioners; an elicitation study would strengthen them considerably.
6. Six models limit the power of any rank comparison across settings.

8 Conclusion

A composite of four measures, published previously, is extended here into a framework over eight criteria with explicit weights, a sampling analysis of their influence, and a protocol whose requirements are justified by measurement. Applied across four settings spanning two corpora, two domains and two tasks, it selects the same detector in each while accuracy selects a different one in each, and the selection holds under 99.1 per cent of sampled weightings over the criteria so far measured.

Completing the measurement of the remaining three criteria, beginning with recall against a defined adversary, is the first item of further work, and the paper specifies how each should be measured so that others may do so.

Declarations

Author contributions N.K.G. is the sole author. N.K.G. conceived the study, designed the evaluation framework and protocol, implemented the experiments, performed the statistical analysis, prepared Figures 1-3 and Tables 1-4, and wrote and reviewed the manuscript.

Ethics and Consent to Participate declarations Not applicable. This study did not involve human participants, human data or animals; it evaluates detection algorithms on publicly available benchmark datasets.

Conflict of interest The author declares no competing interests.

Data availability All measurements are reproducible from notebooks deposited with this paper, which record the corpus versions, the partitioning, the model configurations and the random seeds.

Relation to prior publication This paper extends the author's previously published work [1], which is cited throughout as the origin of the composite metric.

References

[1]

Gyamfi, N.K., Goranin, N., Čeponis, D.: An empirical determination of optimum artificial intelligence algorithm: detection using signal-to-noise ratio approach. IEEE Access **12**, 97712–97725 (2024)

[2]

Wolpert, D.H., Macready, W.G.: No free lunch theorems for optimization. IEEE Trans. Evol. Comput. **1**(1), 67–82 (1997)

[3]

Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (2006)

[4]

Chicco, D., Jurman, G.: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics **21**, 6 (2020)

[5]

Flach, P.: Performance evaluation in machine learning: the good, the bad, the ugly, and the way forward. In: Proc. AAAI Conf. Artif. Intell., vol. 33, pp. 9808–9814 (2019)

[6]

García, S., Herrera, F.: An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons. J. Mach. Learn. Res. **9**, 2677–2694 (2008)

[7]

Benavoli, A., Corani, G., Demšar, J., Zaffalon, M.: Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. J. Mach. Learn. Res. **18**(77), 1–36 (2017)

[8]

Rice, J.R.: The algorithm selection problem. Adv. Comput. **15**, 65–118 (1976)

[9]

Brazdil, P., Giraud-Carrier, C., Soares, C., Vilalta, R.: Metalearning: Applications to Data Mining. Springer, Berlin (2009)

[10]

Vanschoren, J.: Meta-learning: a survey. arXiv:1810.03548 (2018)

[11]

Saaty, T.L.: The Analytic Hierarchy Process. McGraw-Hill, New York (1980)

[12]

Hwang, C.-L., Yoon, K.: Multiple Attribute Decision Making: Methods and Applications. Springer, Berlin (1981)

[13]

Brans, J.-P., Vincke, P.: A preference ranking organisation method: the PROMETHEE method for multiple criteria decision-making. Manage. Sci. **31**(6), 647–656 (1985)

[14]

Saltelli, A., et al.: Global Sensitivity Analysis: The Primer. Wiley, Chichester (2008)

[15]

Taguchi, G.: Introduction to Quality Engineering: Designing Quality into Products and Processes. Asian Productivity Organization, Tokyo (1986)

[16]

Phadke, M.S.: Quality Engineering Using Robust Design. Prentice Hall, Englewood Cliffs (1989)

[17]

Box, G.E.P.: Signal-to-noise ratios, performance criteria, and transformations. Technometrics **30**(1), 1–17 (1988)

[18]

Nair, V.N., et al.: Taguchi's parameter design: a panel discussion. Technometrics **34**(2), 127–161 (1992)

[19]

Davis, J., Goadrich, M.: The relationship between precision-recall and ROC curves. In: Proc. 23rd Int. Conf. Mach. Learn., pp. 233–240 (2006)

[20]

Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE **10**(3), e0118432 (2015)

[21]

Wagner, D., Soto, P.: Mimicry attacks on host-based intrusion detection systems. In: Proc. 9th ACM Conf. Comput. Commun. Secur., pp. 255–264 (2002)

[22]

Biggio, B., et al.: Evasion attacks against machine learning at test time. In: ECML PKDD, pp. 387–402 (2013)

[23]

Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: Proc. Int. Conf. Learn. Represent. (2015)

[24]

Pierazzi, F., Pendlebury, F., Cortellazzi, J., Cavallaro, L.: Intriguing properties of adversarial ML attacks in the problem space. In: Proc. IEEE Symp. Secur. Priv., pp. 1332–1349 (2020)

[25]

Grosse, K., Papernot, N., Manoharan, P., Backes, M., McDaniel, P.: Adversarial examples for malware detection. In: ESORICS, pp. 62–79 (2017)

[26]

Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?" Explaining the predictions of any classifier. In: Proc. 22nd ACM SIGKDD, pp. 1135–1144 (2016)

[27]

Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Adv. Neural Inf. Process. Syst. 30, pp. 4765–4774 (2017)

[28]

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R.: Evaluating the visualization of what a deep neural network has learned. IEEE Trans. Neural Netw. Learn. Syst. **28**(11), 2660–2673 (2017)

[29]

Petsiuk, V., Das, A., Saenko, K.: RISE: randomized input sampling for explanation of black-box models. In: Proc. Br. Mach. Vis. Conf. (2018)

[30]

Warnecke, A., Arp, D., Wressnegger, C., Rieck, K.: Evaluating explanation methods for deep learning in security. In: Proc. IEEE Eur. Symp. Secur. Priv., pp. 158–174 (2020)

[31]

van der Kouwe, E., Andriesse, D., Bos, H., Giuffrida, C., Heiser, G.: Benchmarking crimes: an emerging threat in systems security. arXiv:1801.02381 (2018)

[32]

Kaufman, S., Rosset, S., Perlich, C., Stitelman, O.: Leakage in data mining: formulation, detection, and avoidance. ACM Trans. Knowl. Discov. Data **6**(4), 1–21 (2012)

[33]

Kapoor, S., Narayanan, A.: Leakage and the reproducibility crisis in machine-learning-based science. Patterns **4**(9), 100804 (2023)

[34]

Roberts, D.R., et al.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography **40**(8), 913–929 (2017)

[35]

Allamanis, M.: The adverse effects of code duplication in machine learning models of code. In: Proc. ACM SIGPLAN Onward!, pp. 143–153 (2019)

[36]

Arp, D., et al.: Dos and don’ts of machine learning in computer security. In: Proc. 31st USENIX Secur. Symp., pp. 3971–3988 (2022)

[37]

Pendlebury, F., Pierazzi, F., Jordaney, R., Kinder, J., Cavallaro, L.: Tesseract: eliminating experimental bias in malware classification across space and time. In: Proc. 28th USENIX Secur. Symp., pp. 729–746 (2019)

[38]

Sommer, R., Paxson, V.: Outside the closed world: on using machine learning for network intrusion detection. In: Proc. IEEE Symp. Secur. Priv., pp. 305–316 (2010)

[39]

Čeponis, D., Goranin, N.: Towards a robust method of dataset generation of malicious activity for anomaly-based HIDS training and presentation of AWSCTD dataset. Balt. J. Mod. Comput. **6**(3), 217–234 (2018)

[40]

Moustafa, N.: A new distributed architecture for evaluating AI-based security systems at the edge: network TON_IoT datasets. Sustain. Cities Soc. **72**, 102994 (2021)

Figures

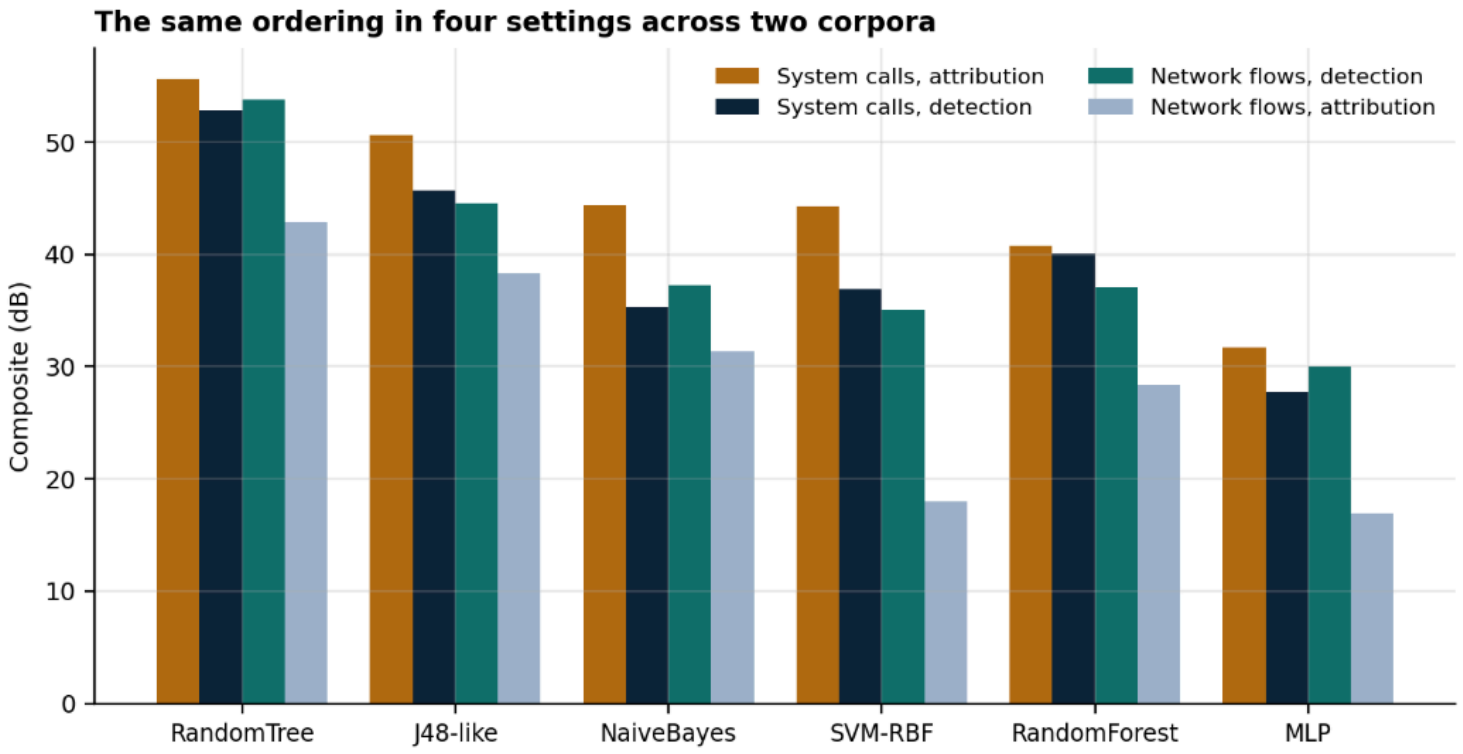


Figure 1

Composite by model in four settings spanning two corpora, two domains and two tasks

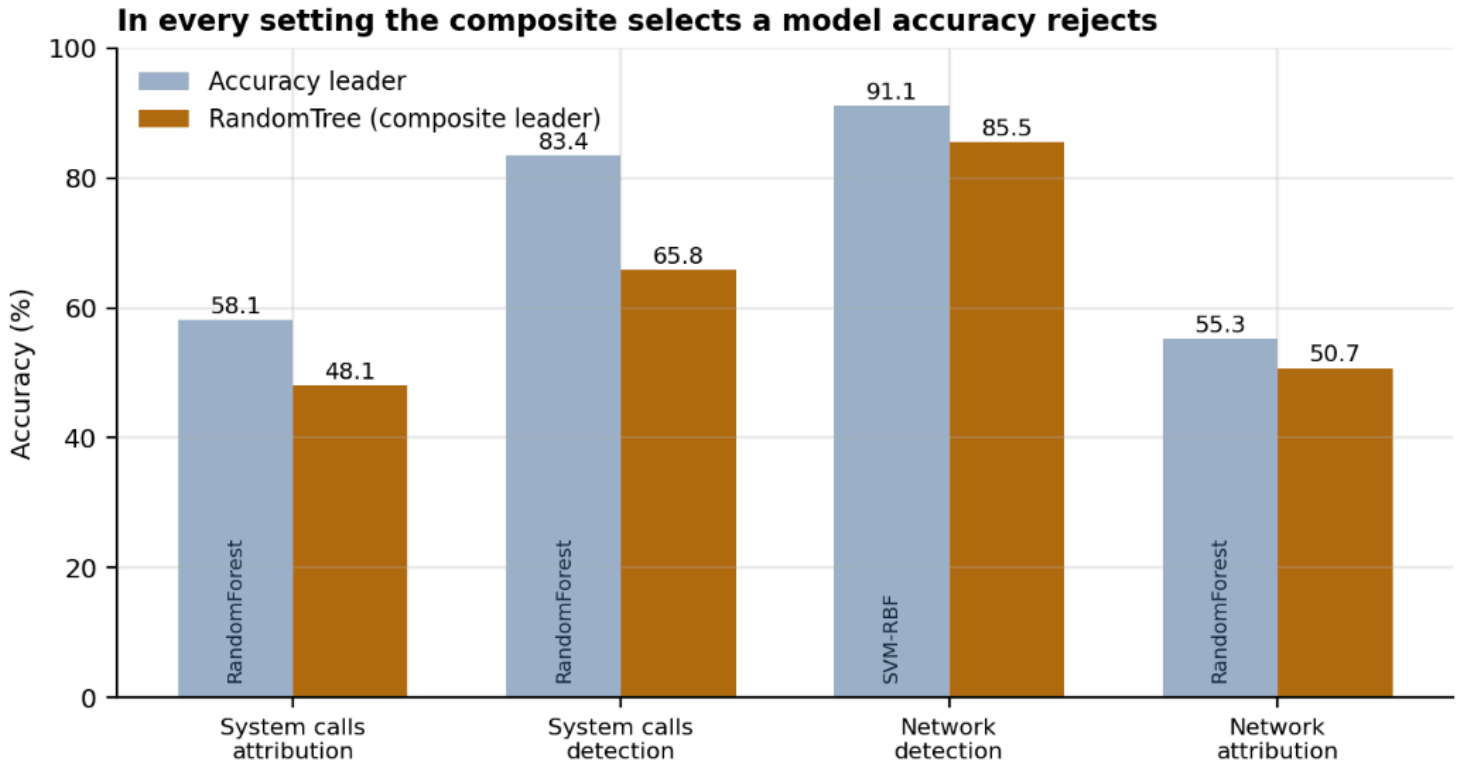


Figure 2

In every setting the accuracy leader differs from the detector the composite selects

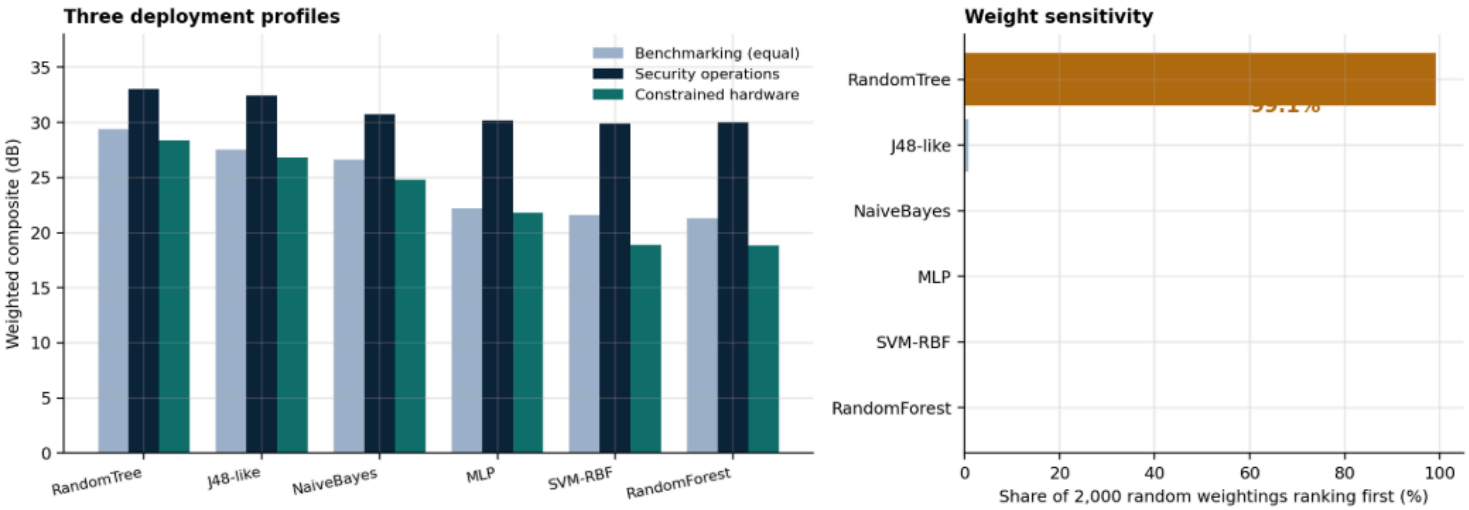


Figure 3

Weighted composite under three deployment profiles, and the share of two thousand sampled weightings in which each detector ranks first

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- IJISRevisedv2TrackedChanges1.docx