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 Confidence Calibration and Selective Abstention Across Heterogeneous Security Decision Tasks

Raji fredo

Rajifredo@outlook.com

milano

Research Article

Keywords: security decision systems, confidence calibration, selective prediction, abstention, biometric authentication, quantum key distribution, digital integrity, uncertainty

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11802873/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 Confidence Calibration and Selective Abstention Across Heterogeneous Security Decision Tasks

Raji Fredo
Independent Researcher

***Abstract*—Security decision systems increasingly combine machine-learning scores with domain-specific integrity evidence, yet raw confidence values are often used without checking whether they remain reliable under operating variation. This paper presents a controlled comparative study of confidence calibration and selective abstention across three representative security settings: biometric identity verification, communication-security monitoring, and digital-content integrity checking. The study does not introduce a new classifier or security architecture. Instead, it evaluates standard logistic regression, random forest, extremely randomized trees, and multilayer perceptron models on reproducible synthetic benchmarks designed to express the types of trust signals commonly reported in these domains. An isotonic calibration stage is additionally applied to the Extra Trees model, and a simple confidence-based abstention policy is evaluated under both development and shifted operating conditions. Across three random seeds and the three domains, logistic regression obtained the highest mean development macro-F1 of 0.915 and AUROC of 0.971, while calibrated Extra Trees achieved macro-F1 of 0.913 and AUROC of 0.969. Calibration reduced the Extra Trees expected calibration error from 0.114 to 0.015 without materially changing discrimination. Under controlled distribution shift, calibrated Extra Trees retained macro-F1 of 0.907 and AUROC of 0.967. With a confidence threshold of 0.90, selective prediction retained about 71.2% of development decisions while increasing accuracy on retained cases to 97.4%. The results support treating calibration and abstention as separate operational properties from classification accuracy when security decisions are made from heterogeneous trust signals.**

***Index Terms*—security decision systems, confidence calibration, selective prediction, abstention, biometric authentication, quantum key distribution, digital integrity, uncertainty**

I. Introduction

Machine-learning components are now routinely embedded in security workflows that were historically based on deterministic checks. Biometric authentication systems estimate whether a presented sample belongs to an enrolled identity; monitoring systems infer whether communication behavior is normal or suspicious; and digital-integrity mechanisms decide whether a document or data object is consistent with an expected cryptographic state. Although the underlying evidence differs, these systems share an operational difficulty: a classifier can be accurate on average while still being overconfident on individual decisions, particularly when operating conditions differ from those seen during development.

This concern is well established in the broader machine-learning literature. Probability estimates from modern classifiers are not automatically calibrated [1], [2], distribution shift can degrade uncertainty estimates even when headline accuracy remains acceptable [3], [4], and explicit rejection or abstention can improve the reliability of automated decision pipelines when uncertain cases are deferred for additional checks [5], [6]. In security contexts, the cost of a confident wrong decision can be substantially higher than the cost of requesting another factor, repeating a measurement, or escalating a case for inspection.

Three application areas illustrate the diversity of such trust signals. First, biometric identity systems combine matching evidence, liveness or quality indicators, and cryptographic binding. Zita and Khalil reported a fingerprint-based recognition pipeline coupled with public/private key generation, demonstrating how identity confidence and cryptographic material can be linked within one workflow [7]. Second, practical quantum key distribution (QKD) monitoring increasingly considers open-set behavior and calibrated uncertainty rather than relying only on a fixed quantum bit error rate threshold. QKD-XGuard is an example of a physics-informed monitoring formulation that explicitly separates supported behavior from previously unseen behavior [14]. Third, digital-content integrity systems combine canonicalization, hashing, signatures, and hierarchical verification. A 2026 framework for Digital Holy Quran verification uses cryptographic integrity mechanisms and structured verification logic to protect highly sensitive digital text [19]. These works address different assets and threat models, but they motivate a common question: how should confidence and uncertainty be handled when security decisions draw on multiple integrity signals?

This paper investigates that question through a comparative benchmark rather than by proposing another domain-specific architecture. Three controlled synthetic tasks represent identity, communication, and content-integrity decisions. Four standard classifiers are compared, the Extra Trees model is calibrated using isotonic regression, and a confidence-based abstention rule is evaluated. The analysis focuses on four practical questions:

1) How consistent are standard classifiers across heterogeneous security-signal distributions?
2) Does post-hoc calibration improve probability reliability without materially changing discrimination?
3) How much performance is retained under a moderate operating shift?

4) What accuracy–coverage trade-off results from abstaining on low-confidence decisions?

The contribution is therefore empirical and comparative. No new classifier, cryptographic primitive, or security protocol is claimed.

II. Background and Related Work

A. Biometric identity and cryptographic binding

Biometric recognition has long been studied as a probabilistic identity decision problem in which acquisition quality, intra-class variation, spoofing, and template security all affect operational reliability [8], [9]. Template-protection methods aim to avoid storing directly exploitable biometric representations, while fuzzy extractors and related constructions connect noisy biometric measurements with reproducible cryptographic material [10]. Recent work has continued to investigate efficient recognition pipelines and cryptographic identity binding. The fingerprint-based system of Zita and Khalil combines handcrafted fingerprint descriptors, LightGBM classification, and key-pair generation and reports evaluation across FVC2000, FVC2002, and FVC2004 [7]. For the present study, this literature motivates identity-domain features such as match consistency, quality, liveness-like evidence, and key-binding consistency; the original fingerprint data are not reused.

B. Communication monitoring and unsupported behavior

QKD provides security guarantees under specified protocol and device assumptions, but implementation imperfections can create exploitable deviations [11]–[13]. Detector-control attacks demonstrated that practical devices can behave outside the simplified models assumed by protocol-level reasoning [12]. This gap has motivated monitoring methods that inspect richer physical telemetry. The QKD-XGuard formulation considers physics-informed residuals, conformal prediction, and explicit open-set rejection for attack behaviors not represented during training [14]. The present benchmark does not simulate a QKD protocol; instead, it uses this application area to motivate a communication-domain task in which residual deviation, timing inconsistency, detector imbalance, and latent-distance-like signals can vary under benign and malicious conditions.

C. Digital-content integrity and hierarchical verification

Cryptographic integrity mechanisms are fundamentally different from biometric matching but encounter a related operational problem when several checks must be interpreted jointly. Secure hash functions provide deterministic digests [15]; digital signatures provide origin authentication and integrity under cryptographic assumptions [16]; and canonicalization is necessary when semantically equivalent text representations can differ at the byte level [17]. Merkle constructions additionally permit hierarchical verification of large collections [18]. The Digital Holy Quran verification framework by Zita et al. combines canonical Arabic-text normalization, hierarchical SHA-256 hashing, and visual authentication within a multilayer verification architecture [19]. In this study, the content-integrity domain abstracts these ideas into signals such as digest consistency, signature or certificate consistency, normalization agreement, and hierarchy consistency.

D. Calibration, distribution shift, and selective prediction

Calibration asks whether predicted probabilities correspond to empirical event frequencies. Platt scaling [20], isotonic regression [21], and temperature scaling [2] are common post-hoc techniques. Niculescu-Mizil and Caruana showed that different supervised learning algorithms produce probability estimates with different calibration characteristics [1]. More recent work has emphasized that calibration quality can change under dataset shift [4].

A related line of work concerns abstention. Chow formalized classification with a reject option [5]; selective classification later reframed the problem as balancing predictive risk against coverage [6]. In security practice, abstention corresponds naturally to step-up authentication, repeated measurement, manual review, link isolation, or another verification step. The present paper therefore evaluates not only classification quality but also the accuracy of retained decisions as the confidence threshold increases.

III. Experimental Methodology

A. Benchmark design

Three binary security-decision domains are generated independently: Identity, Communication, and Content Integrity. Each development set contains 6,000 observations. A separate shifted set contains 1,800 observations per domain. Labels are balanced between benign/supported operation ($y = 0$) and malicious/inconsistent operation ($y = 1$).

Each observation has ten standardized continuous features. The first five features are domain-specific; the remaining five represent nuisance, quality, load, and weak auxiliary evidence. The simulator does not claim to reproduce a specific deployed biometric sensor, QKD testbed, or digital archive. Its purpose is controlled comparison: each domain exposes multiple moderately informative signals, and the shifted condition changes nuisance variables while slightly reducing the separability of core evidence.

For Identity, the principal signals represent match consistency, liveness-like evidence, template consistency, cryptographic binding consistency, and auxiliary anomaly evidence. For Communication, they represent physical residual deviation, timing inconsistency, detector imbalance, error-rate deviation, and latent-distance-like evidence. For Content Integrity, they represent hash disagreement, signature/certificate disagreement, canonicalization disagreement, hierarchy inconsistency, and auxiliary verification evidence.

B. Models

Four standard classifiers are evaluated:

- logistic regression with balanced class weighting;
- random forest with 180 trees;
- extremely randomized trees (Extra Trees) with 220 trees;
- a multilayer perceptron with two hidden layers of 48 and 24 units.

TABLE I
MEAN PERFORMANCE ACROSS THREE DOMAINS AND THREE SEEDS

Model	Split	F1	Bal. Acc.	AUROC	ECE	Brier
Logistic	Dev.	**0.915**	**0.915**	**0.971**	0.014	**0.064**
Random Forest	Dev.	0.909	0.908	0.965	0.047	0.073
Extra Trees	Dev.	0.913	0.912	0.969	0.114	0.081
MLP	Dev.	0.911	0.910	0.969	0.026	0.068
Cal. Extra Trees	Dev.	0.913	0.912	0.969	0.015	0.066
Logistic	Shift	0.905	0.907	**0.968**	0.023	0.069
Random Forest	Shift	0.900	0.901	0.961	0.050	0.077
Extra Trees	Shift	0.907	0.908	0.967	0.115	0.085
MLP	Shift	0.901	0.902	0.965	0.028	0.073
Cal. Extra Trees	Shift	**0.907**	**0.908**	0.967	**0.018**	**0.069**

Logistic regression and the MLP use standardized inputs. Tree models operate on the original generated values. Hyperparameters are held fixed across domains and random seeds to avoid domain-specific tuning.

An additional Calibrated Extra Trees model is constructed using isotonic regression with three-fold internal calibration. This model is included to isolate the effect of post-hoc probability calibration from discrimination performance.

C. Splitting and repeated evaluation

For each domain and seed, 70% of the 6,000 development observations are used for training and 30% for testing, with stratification by class. The experiment is repeated with seeds 17, 31, and 59. The separate shifted set is generated with a different random stream and is never used for model fitting.

D. Metrics

The primary discrimination metrics are macro-F1, balanced accuracy, and area under the receiver operating characteristic curve (AUROC). Probability quality is measured using the Brier score and expected calibration error (ECE). For M probability bins B_m, ECE is

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|. \tag{1}$$

Ten equally spaced bins are used.

For selective prediction, let $p(x)$ denote the predicted probability of the positive class and define confidence as

$$c(x) = \max\{p(x), 1 - p(x)\}. \tag{2}$$

A prediction is retained only if $c(x) \geq \tau$. Coverage is the fraction retained, while selective accuracy is accuracy among retained observations. Thresholds $\tau \in \{0.60, 0.70, 0.80, 0.90\}$ are evaluated.

IV. RESULTS

A. Overall classifier comparison

Table I reports mean results across the three domains and three seeds. Logistic regression produced the highest development macro-F1 (0.915) and AUROC (0.971), closely followed by calibrated Extra Trees. The ranking changed little under the shifted condition, although all models showed a small decline.

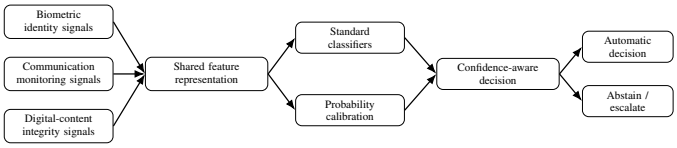

Fig. 1. Study workflow across the three security domains. Domain-specific signals are mapped to a common representation, evaluated by standard classifiers, calibrated where required, and passed to a confidence-aware decision stage.

TABLE II
CALIBRATED EXTRA TREES BY SECURITY DOMAIN

Domain	Split	F1	Bal. Acc.	AUROC	ECE
Identity	Dev.	0.916	0.916	0.972	0.017
Identity	Shift	0.909	0.910	0.969	0.022
Communication	Dev.	0.894	0.893	0.958	0.015
Communication	Shift	0.888	0.888	0.955	0.019
Content Integrity	Dev.	0.929	0.929	0.979	0.013
Content Integrity	Shift	0.926	0.926	0.976	0.015

Two observations are important. First, a relatively simple linear classifier remained competitive because the benchmark was deliberately constructed from multiple moderately informative signals rather than highly nonlinear feature interactions. Second, raw Extra Trees probabilities were poorly calibrated despite strong discrimination. Isotonic calibration reduced ECE from 0.114 to 0.015 in development and from 0.115 to 0.018 under shift, with negligible change in F1 or AUROC.

B. Domain-level results

Table II shows the Calibrated Extra Trees results by domain. Content Integrity was the easiest task, with development AUROC of 0.979 and shifted AUROC of 0.976. Communication was the most difficult, but still retained AUROC above 0.95 under shift. Identity performance fell between these two cases.

The result does not imply that digital-content verification is inherently easier than biometric or communication-security monitoring. Rather, it reflects the signal distributions used in this controlled benchmark. The purpose of the domain comparison is to test whether calibration and abstention behavior remain qualitatively similar when the underlying evidence sources differ.

C. Effect of calibration

Calibration substantially changed probability reliability without changing ranking quality. Extra Trees and Calibrated Extra Trees had almost identical AUROC values, yet the calibrated model's ECE was lower by approximately 0.099 in development. This distinction matters operationally: a system that uses a predicted probability to trigger step-up verification or another action requires probability estimates whose scale is meaningful, not only a ranking that separates classes well.

Logistic regression was already well calibrated in this benchmark, with development ECE of 0.014. This is consistent with prior observations that some generalized linear models can provide reasonable probabilities when their functional assumptions are not severely violated [1]. The comparison



Fig. 2. Calibration and selective-decision mechanism used in the benchmark. Calibration changes the interpretation of the probability scale, while the threshold τ determines whether a prediction is retained or escalated.

TABLE III
SELECTIVE PREDICTION WITH CALIBRATED EXTRA TREES

Split	Threshold τ	Coverage	Selective Acc.
Development	0.60	0.956	0.927
Development	0.70	0.896	0.945
Development	0.80	0.826	0.960
Development	0.90	0.712	0.974
Shifted	0.60	0.951	0.925
Shifted	0.70	0.892	0.943
Shifted	0.80	0.815	0.959
Shifted	0.90	0.699	0.974

also shows why calibration should be measured rather than assumed from model family alone.

D. Selective abstention

Table III reports the confidence–coverage trade-off for Calibrated Extra Trees averaged across all domains and seeds. Raising the threshold monotonically increased accuracy on retained cases while reducing coverage. At $\tau = 0.90$, approximately 71.2% of development observations were retained and selective accuracy increased to 97.4%. Under shift, coverage was approximately 69.9% and retained accuracy remained 97.4%.

In practical systems, the rejected fraction would not necessarily be denied outright. In a biometric system it could trigger another factor or recapture; in a communication monitor it could trigger additional telemetry collection; and in a content-integrity system it could invoke a deeper signature or provenance check. The appropriate threshold therefore depends on the cost of escalation relative to the cost of a wrong automatic decision.

V. DISCUSSION

The experiments support three general conclusions. First, discrimination and calibration should be reported separately. Extra Trees provided competitive F1 and AUROC, but its raw probabilities were substantially less reliable than its calibrated probabilities. Second, modest distribution shift did not destroy discrimination, yet calibration error still changed, reinforcing the need to evaluate probability reliability under conditions beyond the development split. Third, abstention provided a transparent mechanism for trading automated coverage against decision accuracy.

These findings connect naturally to the three security contexts considered in the related work without claiming that the synthetic benchmark validates those specific systems. Fingerprint-based cryptographic identity combines recognition

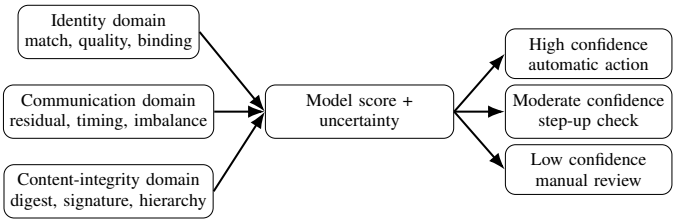

Fig. 3. Domain-specific interpretation of selective prediction. The same confidence mechanism can lead to different operational follow-up actions depending on whether the evidence originates from identity verification, communication monitoring, or digital-content integrity checking.

confidence with security-sensitive key handling [7]; an operational implementation could benefit from escalation when identity confidence is insufficient. Open-set QKD monitoring explicitly treats unsupported behavior as different from a forced known-class label [14]; abstention is therefore conceptually aligned with cautious monitoring. Hierarchical digital-text verification combines deterministic cryptographic checks with layered validation logic [19]; uncertainty management becomes relevant when auxiliary or visual evidence is introduced alongside cryptographic checks.

The study also shows that a more complicated classifier is not necessarily preferable. Logistic regression performed as well as or slightly better than the nonlinear models on several aggregate metrics. For operational security systems, model selection should therefore consider calibration, computational cost, interpretability, and escalation policy in addition to maximum predictive accuracy.

VI. LIMITATIONS

The principal limitation is that the benchmark is synthetic. It captures interpretable patterns motivated by three security domains but does not reproduce the full physical, behavioral, or cryptographic mechanisms of any deployed system. In particular, the communication task is not a QKD simulator; the identity task does not use raw fingerprint images; and the content-integrity task does not process Quranic text or cryptographic signatures directly. The results should therefore be interpreted as a controlled comparison of uncertainty-handling procedures, not as performance claims for those application systems.

A second limitation is the binary task formulation. Real systems often involve multiple attack types, multiple identities, hierarchical error categories, or explicit unknown classes. Third, only one post-hoc calibration method is applied to one tree model. Temperature scaling, Platt scaling, beta calibration, conformal prediction, and model ensembles could be compared in future empirical work. Finally, the shift model is moderate and predefined. Real deployment shift can be gradual, adversarial, device-specific, or correlated across features.

VII. REPRODUCIBILITY

The complete experiment is implemented in Python using NumPy, pandas, and scikit-learn [28]. The supplied script

deterministically regenerates the three development datasets, the shifted datasets, all model fits, calibration results, and selective-prediction tables using seeds 17, 31, and 59. No private, personal, or human-subject data are used.

VIII. Conclusion

This paper compared standard classification, post-hoc calibration, and confidence-based abstention across controlled identity, communication, and content-integrity security tasks. Logistic regression and tree ensembles achieved AUROC near 0.97, but probability reliability differed materially across models. Isotonic calibration reduced Extra Trees ECE from approximately 0.114 to 0.015 in the development condition with negligible change in discrimination. Selective abstention further increased retained-decision accuracy to approximately 97.4% at 71.2% coverage. The main implication is methodological rather than architectural: security studies that rely on learned confidence scores should evaluate calibration and explicit uncertainty handling in addition to classification accuracy.

References

[1] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 625–632, doi: 10.1145/1102351.1102430.

[2] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1321–1330, arXiv doi: 10.48550/arXiv.1706.04599.

[3] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, Eds., *Dataset Shift in Machine Learning*. Cambridge, MA, USA: MIT Press, 2009, doi: 10.7551/mitpress/9780262170055.001.0001.

[4] Y. Ovadia *et al.*, "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, arXiv doi: 10.48550/arXiv.1906.02530.

[5] C. K. Chow, "On optimum recognition error and reject tradeoff," *IEEE Trans. Inf. Theory*, vol. 16, no. 1, pp. 41–46, 1970, doi: 10.1109/TIT.1970.1054406.

[6] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, arXiv doi: 10.48550/arXiv.1705.08500.

[7] W. Zita and M. S. Khalil, "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: 10.63496/ejcs.Vol1.Iss3.112.

[8] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 4–20, 2004, doi: 10.1109/TCSVT.2003.818349.

[9] N. K. Ratha, J. H. Connell, and R. M. Bolle, "Enhancing security and privacy in biometrics-based authentication systems," *IBM Syst. J.*, vol. 40, no. 3, pp. 614–634, 2001, doi: 10.1147/sj.403.0614.

[10] Y. Dodis, R. Ostrovsky, L. Reyzin, and A. Smith, "Fuzzy extractors: How to generate strong keys from biometrics and other noisy data," *SIAM J. Comput.*, vol. 38, no. 1, pp. 97–139, 2008, doi: 10.1137/060651380.

[11] C. H. Bennett and G. Brassard, "Quantum cryptography: Public key distribution and coin tossing," *Theor. Comput. Sci.*, vol. 560, pp. 7–11, 2014, doi: 10.1016/j.tcs.2014.05.025.

[12] L. Lydersen *et al.*, "Hacking commercial quantum cryptography systems by tailored bright illumination," *Nature Photon.*, vol. 4, pp. 686–689, 2010, doi: 10.1038/nphoton.2010.214.

[13] H.-K. Lo, M. Curty, and B. Qi, "Measurement-device-independent quantum key distribution," *Phys. Rev. Lett.*, vol. 108, p. 130503, 2012, doi: 10.1103/PhysRevLett.108.130503.

[14] W. Zita, "QKD-XGuard: Physics-informed conformal open-set monitoring for unseen and adaptive attacks in decoy-state BB84," Research Square, Version 1, 2026, doi: 10.21203/rs.3.rs-11078149/v1.

[15] National Institute of Standards and Technology, "Secure Hash Standard (SHS)," FIPS PUB 180-4, 2015, doi: 10.6028/NIST.FIPS.180-4.

[16] S. Josefsson and I. Liusvaara, "Edwards-Curve Digital Signature Algorithm (EdDSA)," RFC 8032, Internet Engineering Task Force, 2017, doi: 10.17487/RFC8032.

[17] M. Davis and K. Whistler, "Unicode Normalization Forms," Unicode Standard Annex #15, Unicode Consortium, 2023.

[18] R. C. Merkle, "A certified digital signature," in *Advances in Cryptology—CRYPTO '89*, 1990, pp. 218–238, doi: 10.1007/0-387-34805-0_21.

[19] W. Zita, N. Zita, M. S. Khalil, R. I. M. ElShenieky, and B. Şahin, "Verification Method and Framework for the Digital Holy Quran," Research Square, 2026, doi: 10.21203/rs.3.rs-9577829/v1.

[20] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. Cambridge, MA, USA: MIT Press, 1999, pp. 61–74.

[21] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2002, pp. 694–699, doi: 10.1145/775047.775151.

[22] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proc. ICLR*, 2017, arXiv doi: 10.48550/arXiv.1610.02136.

[23] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, arXiv doi: 10.48550/arXiv.1612.01474.

[24] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1757–1772, 2013, doi: 10.1109/TPAMI.2012.256.

[25] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*, 2nd ed. Cham, Switzerland: Springer, 2022, doi: 10.1007/978-3-031-06649-8.

[26] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[27] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, pp. 3–42, 2006, doi: 10.1007/s10994-006-6226-1.

[28] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011, doi: 10.5555/1953048.2078195.

[29] D. Sculley *et al.*, "Hidden technical debt in machine learning systems," in *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[30] E. Tabassi, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, 2023, doi: 10.6028/NIST.AI.100-1.