ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Techniques for Reliable, Safe and Robust AI Applications

Caroline König[1], Pedro Jesús Copado[1], Cecilio Angulo[1,2] and G. Kumar Venayagamoorthy[3,4] *

1- Computer Science Department and Intelligent Data Science and Artificial Intelligence (IDEAI-UPC) Research Center
Universitat Politècnica de Catalunya, Barcelona, Spain
2- Institut de Robòtica i Informàtica Industrial (IRI), Barcelona, Spain
3- Real-Time Power and Intelligent Systems Lab
Holcombe Department of Electrical and Computer Engineering
Clemson University, Clemson, USA
4- Department of Electrical, Electronic and Computer Engineering
University of Pretoria, Pretoria, South Africa

Abstract.
Reliability, safety, and robustness are essential requirements for safety critical applications. Implementing these properties in artificial intelligence based systems introduces additional challenges, particularly in the development and validation of data driven models. To address these requirements, new techniques are needed for assessing predictive uncertainty, ensuring robustness against intentional or environmental input perturbations, integrating explicit safety constraints into model architectures, and enabling human oversight and interpretability to support auditing and supervision.

1 Introduction

Artificial Intelligence (AI) based systems are nowadays often deployed in safety critical applications such as healthcare, energy, transportation, or critical infrastructures in general [1, 2, 3, 4] due to their ability to solve complex problems efficiently. Nevertheless in such domains failures can have a significant impact on environment or persons and in consequence ensuring the safety and reliability of such AI-based applications has become a central concern to assure they operate reliable even under uncertainty and dynamic conditions [5]. This introduces additional requirements for the development of machine learning (ML) models, including the ability to work under unexpected or degraded inputs, maintaining robustness against attacks, and supporting thorough validation and human oversight. Addressing these challenges is key for the trustworthy adoption of AI in safety-critical domains.

*This research was partially funded by European Union NextGenerationEU as the ERAPERMED 2022-292 research project, supported as grant AC22/0053 by the Instituto de Salud Carlos III (ISCIII) in Spain and U.S. National Science Foundation Grants: CNS 2318612, CNS 2131070 and ECCS 2234032. Prof. Cecilio Angulo is also acknowledging co-funds from the European Union – NextGenerationEU, through the I+D+i PID2021-122835OBC21 research project ARTIFACTS (MCIN/AEI/10.13039/501000011033) and FEDER "Una manera de hacer Europa".

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

In the following sections, this article reviews techniques designed to ensure the reliable operation of ML models under the practical conditions encountered in real-world systems. These include methods to maintain performance under distribution shift, defend against adversarial attacks, and validate robustness and generalization in safety-critical environments. Furthermore, the article summarizes the contributions of the studies accepted for the special session on "Reliable, Safe and Robust AI Applications" of the 34th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2026).

2 Safety-Oriented Machine Learning

2.1 Uncertainty

ML models are trained on datasets that frequently capture only some degree of the variability of real-world environments [6]. As a consequence, their performance can degrade when deployed under conditions that differ from those observed during training [7] representing a challenge for the reliability of ML models in real-world applications. Such discrepancies are known as *distribution shifts*, which describe a mismatch between training and real-world data, or *Concept drift*, which represent changes in data-label based relationships over time [8]. An additional difficulty arises when the model encounters inputs that do not belong to any of the classes represented in the training set, a setting known in the literature as *open-set recognition* [9]. Such application contexts require models not only capable of classifying known categories but also to identify and appropriately handle previously unseen entities [10].

These challenges lead to the need for mechanisms that allow models to assess the reliability of their own predictions [11], enabling them to detect when they are operating under uncertain conditions. Uncertainty-aware mechanisms estimate the reliability or confidence of a model's prediction [11] and include approaches such as Bayesian approximations [12], ensemble-based uncertainty estimation [13], and evidential learning [14]. These methods aim to implement alternative decision-making strategies if model uncertainty is high. For instance, in the medical domain, model uncertainty must be taken into account and require the opinion of a clinical professional [15] if necessary. Similarly, in autonomous driving, systems should activate alternative safe decisions to avoid risk in uncertain situations [16].

Complementary strategies are *out-of-distribution* detection, *anomaly* detection, or *dummy-class* modeling, which aim to identify inputs that fall outside the learned domain [17]. Out-of-distribution and anomaly-detection methods usually rely on methods that focus on the match between the input and the training distribution, for example by analyzing the softmax confidence [18] or measuring distances in feature space [19]. Dummy-class approaches in-turn introduce an explicit unknown class that recognizes samples that do not belong to any known category, for instance by learning a shared prototype that represents unfamiliar inputs [20].

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

2.2 Security constraints

While ML implements data-driven learning, many real-world applications require explicit security constraints to ensure that certain conditions are always satisfied [21]. Such constraints often are domain knowledge, regulatory requirements, or safety rules that must be met. To enforce them, *constraint-aware* systems incorporate mechanisms such as rule-based logic, logical constraints, or formal verification steps that guarantee model predictions remain within safe settings [22].

Several strategies exist for embedding such constraints into machine learning models [23]. One approach is to enforce *hard constraints* through post-processing layers that adapt model predictions which are contrary to predefined rules [24]. Another strategy involves *constrained learning*, where the model is trained under optimization constraints that encode domain-specific rules [25]. Another approach are hybrid architectures that combine ML with symbolic reasoning or rule-based systems implementing an additional mechanism for satisfying security requirements [26]. Such architectures have been successfully applied in several real-world domains. For instance, in autonomous driving to integrate neural network decisions with explicit traffic rules [27] or in healthcare to combine ML models with clinical guidelines integrating rule-based components to ensure that the predictions are consistent with medical protocols [28].

2.3 Privacy preserving learning

The protection of user data is a legal requirement and must be taken into account when developing ML models based on sensitive information. A variety of privacy preserving learning techniques have been proposed. Federated learning [29] enables collaborative model training across distributed systems without transferring raw data reducing the risk of data leakages. Differentially Private Stochastic Gradient Descent (DP-SGD) introduces calibrated noise into the optimization process to provide formal privacy guarantees [30] ensuring that the contribution of a single data point cannot be reverse-engineered from the trained model. These approaches contribute towards privacy safeguarding ML systems.

2.4 Robustness and stability

ML models are not only sensitive to natural variations in the input data but can also be intensionally manipulated through *adversarial attacks* [31]. These attacks introduce slight perturbations in the input to cause the model to produce incorrect predictions. Such vulnerabilities represent a serious concern for the deployment of ML systems in security-critical environments [2].

ML models can be compromised through several classes of adversarial attacks. Gradient-based *white-box* attacks exploit access to model parameters to find effective perturbations [32]. Optimization-driven attacks generate perturbations through searches for the most imperceptible modifications that trigger an incorrect prediction [31]. *Black-box* query attacks operate without access to the

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

internal model parameters and rely on estimating gradients or decision boundaries through exploring the prediction space [33]. These attack strategies focus on different type of model vulnerability and are addressed as a consequence by different type of defensive mechanisms [34]. *Adversarial training* involves training on both clean and adversarial examples and is an effective defense strategy. Other defense techniques include *certified robustness* which examines the model's behavior under bounded perturbations [35], *randomized smoothing*, which trains over noise-perturbed inputs to create more robust models [36] or methods that aim on removing adversarial perturbations before the prediction happens. In domains, such as energy or healthcare such detection based approaches identify adversarially manipulated samples prior to proceed with the downstream processing in the ML application [37].

Beside adversarial manipulations, robustness also encompasses *stability testing*, which evaluates how model predictions change under unintended perturbations such as noise, occlusions or environmental variations. Such evaluations provide a broader view of model reliability by assessing sensitivity to naturally occurring distortions and corruptions [38, 39]. These issues are particularly relevant in autonomous driving, where models must operate reliably for instance under adverse weather conditions or partial visual occlusions [40]. Stability testing is closely related to stress testing, which aim to identify regions of the input space where the model behaves unpredictably [41].

2.5 Validation

Safe deployment of ML models requires validation procedures that extend beyond conventional performance metrics used to evaluate the model's performance. In safety critical domains, AI safety focuses on verifying that models behave predictably across diverse operating conditions and comply with domain-specific safety requirements [42]. In certain sectors such as healthcare, finance, and autonomous systems, these validation strategies are embedded within certification workflows that require traceability, auditability, and compliance with certain safety standards [2].

Many ML models do not make their decision logic explicit and instead operate as opaque black box models so that their validation relies on complementary testing paradigms. Black-box testing evaluates end-to-end behavior, gray-box testing incorporates partial knowledge of the model or data, and white-box testing examines internal decision paths to uncover failure modes [43]. A common approach for testing neural networks is *coverage testing*, where metrics such as *neuron coverage* quantify how thoroughly the test set covers the network's internal logic [44]. Such metrics help identify insufficiently tested neurons that can hint to *corner cases*, situations that are rare and often underrepresented in the training data space [45].

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

2.6 Auditing

Human oversight and interpretability are considered essential for safe, robust, and reliable AI systems. *Human oversight* provides a mechanism that enables monitoring, intervention, and corrective action when models behave unexpectedly while *interpretability* focuses on making model reasoning transparent, supporting tasks such as debugging or system auditing. ML models range from opaque black-box architectures to intrinsically interpretable models with transparent decision logic [46]. For black-box models interpretability approaches comprise *post-hoc* explanation methods such as saliency maps, feature attributions, and counterfactual explanations [47, 48]. For example, counterfactual explanations describe how a small change in an input value impacts the model to predict a different class. In turn intrinsically interpretable models, such as decision trees or rule-based models allow humans to fully understand the underlying decision logic [49]. These mechanisms enable humans to assess whether model outputs align with domain knowledge, safety constraints, and ethical guidelines [47, 48, 50]. Different regulatory frameworks, such as the EU AI Act [51] or the NIST AI Risk Management Framework [52], specifically require appropriate human oversight, transparency, and explainability for high-risk AI systems.

3 Contributions to the ESANN 2026 Special Session

The special session on "Reliable, Safe and Robust AI Applications" at the 34^{th} European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2026) consists of ten paper contributions. This section summarizes these contributions in light of the previous sections.

Three contributions to the session address the use of ML models under uncertainty. Arnold Brosch *et al.*'s paper on "Out-of-Distribution Segmentation via Wasserstein-Based Evidential Uncertainty" [53] introduces an evidential deep learning framework trained with Wasserstein loss to achieve uncertainty-aware semantic segmentation in computer vision tasks. Their approach provides an out of distribution aware segmentation model that captures the underlying probability distribution smoother and provides improved consistency and uncertainty estimates. Benedict Franke *et al.* [54]'s paper entitled "Revisiting Neural Activation Coverage for Uncertainty Estimation" revisits the use of neural activation coverage (NAC), a technique originally proposed for out-of-distribution detection. This work extends NAC as an uncertainty-estimation method for pretrained neural networks in regression settings, demonstrating how activation-based signals can provide meaningful insight into predictive reliability. Jiarui Zhang *et al.* [55] introduce the Neuromodulated Delta Adapter (NDA), a module that enables stable and efficient test-time adaptation for frozen Transformer models under distribution shifts. NDA inserts a "rank-r fast-weight" bottleneck into each layer and updates these fast weights using a gated Delta rule modulated by a three-factor signal. This neuromodulatory mechanism provides adaptive gain control and ensures stable training.

The contribution of Luiz Pereiera *et al.* [56] concerns privacy preserving ML

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

models. The work analyzes the sensitivity of a federated learning approach applied to spiking neural networks, a class of neuromorphic models. The authors investigate how privacy-related factors and data distribution affect the learning dynamics and robustness of these networks within a federated learning framework.

The special session comprises also three contributions related to the topic of adversarial attacks. The contribution of Jacub Hoscilowicz *et al.* [57] investigates a specific class of adversarial attacks referred to as *confusion attack* designed to intentionally destabilize multimodal Large Language Models. This type of attack focuses on maximizing the next-token Shannon entropy of the model's output distribution bringing the model toward highly uncertain predictions. The study shows how this type of attack influence the decoding process and identifies five distinct modes of model failure related to maximized perturbation. The paper by T. Tram Ho *et al.* [58] entitled "SPARC: Superpixel-based Black-Box Adversarial Attack with Regional Confidence" presents an adversarial attack approach using superpixel-guidance in a black box setting. The method segments images into superpixels, identifies the most discriminative regions through occlusion-based confidence scoring, and performs efficiently the perturbation generations based on an optimization process, iteratively updating noise over selected superpixels to maximize the attack objective under L1/L2 norm constraints. Luca Pajola *et al.* [59] introduces an adversarial attack that targets the model selection stage. The study propose "Model Selection Hijacking" (MOSHI), a framework that uses a Variational Autoencoder to generate adversarial validation samples capable of biasing automated model ranking. By perturbing only the validation set this attack influences the model selection process toward suboptimal models, for example those with poorer generalization. The approach is illustrated with computer vision and speech recognition tasks.

A contribution which falls into the scope of interpretability is Fabrizio Leopardi *et al.*'s paper [60] entitled "Interpreting Logical Explanations of Classifying Neural Networks" that presents a technique to make formal, logic-based explanations of neural network decisions more interpretable for humans. Space Explanations express classification pathways as logical formulas, which are often difficult for users to understand. This study proposed using classical geometric interpretations to translate the logical explanations into more intuitive and human understandable forms. Another contribution related to interpretability is Akash Borigi *et al.*'s [61] paper entitled "The Alignment Gate: Intent and Instruction Guardrails for Agentic AI". The article presents an interpretable alignment framework for Agentic AI systems, aimed at reliably mapping user intents to the system instructions that an autonomous agent should execute. Its contribution is the introduction of a probabilistic association model that quantifies how plausibly each instruction corresponds to a given intent, enabling transparent and controllable decision-making.

Muhammad Ahsan *et al.*'s [62] paper relates to robustness techniques and in particular proposed a mitigation technique against spurious correlations in training data that lead to short cut learning in ML models. The proposed

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

shortcut mitigation framework is based on a sparse autoencoder to distinguish core and spurious features in the model's embeddings, allowing them to identify and mute *shortcut neurons*.

References

[1] E. Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25:44–56, 2019.

[2] J. Perez-Cerrolaza et al. Artificial intelligence for safety-critical systems in industrial and transportation domains: A survey. *ACM Computing Surveys*.

[3] J. Park and D. Kang. Artificial Intelligence and Smart Technologies in Safety Management: A Comprehensive Analysis Across Multiple Industries. *Applied Sciences*, 14(24):11934, 2024.

[4] N. Pillay, T. Nyathi, and G. K. Venayagamoorthy. Artificial Intelligence for Critical Infrastructure Systems: Past, Present and Future. In *Proceedings of the 2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, June 2025.

[5] B. Shneiderman. Responsible artificial intelligence governance: A review and research framework. *ACM Transactions on Interactive Intelligent Systems*, 10(4):1–27, 2020.

[6] J. Qu Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. Dataset shift in machine learning. *The MIT Press*, 1(5):1, 2009.

[7] Y. Ovadia et al. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

[8] G. I. Webb, L. K. Lee, B. Goethals, and F. Petitjean. Analyzing concept drift and shift from sample data. *Data Mining and Knowledge Discovery*, 32(5):1179–1199, 2018.

[9] C. Geng, S. Huang, and S. Chen. Recent advances in open set recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3614–3631, 2020.

[10] M. Mancini, Z. Akata, E. Ricci, and B. Caputo. Towards recognizing unseen categories in unseen domains. In *European Conference on Computer Vision*, pages 466–483. Springer, 2020.

[11] H. Tyralis and G. Papacharalampous. A review of predictive uncertainty estimation with machine learning. *Artificial Intelligence Review*, 57(4):94, 2024.

[12] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[13] J. Liu, J. Paisley, M. Kioumourtzoglou, and B. Coull. Accurate uncertainty estimation and decomposition in ensemble learning. *Advances in neural information processing systems*, 32, 2019.

[14] J. Gao, M. Chen, L. Xiang, and C. Xu. A comprehensive survey on evidential deep learning and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[15] B. Kompa, J. Snoek, and A. L. Beam. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1):4, 2021.

[16] M. Khonji, J. Dias, R. Alyassi, F. Almaskari, and L. Seneviratne. A risk-aware architecture for autonomous vehicle operation under uncertainty. In *2020 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pages 311–317. IEEE, 2020.

[17] J. Yang, K. Zhou, Y. Li, and Z. Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

[18] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[19] C. Leys, O. Klein, Y. Dominicy, and C. Ley. Detecting multivariate outliers: Use a robust variant of the mahalanobis distance. *Journal of experimental social psychology*, 74:150–156, 2018.

[20] Z. Liu et al. Psdc: A prototype-based shared-dummy classifier model for open-set domain adaptation. *IEEE Transactions on Cybernetics*, 53(11):7353–7366, 2022.

[21] K. R. Varshney and H. Alemzadeh. On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. *Big data*, 5(3):246–255, 2017.

[22] R. McAllister, Y. Gal, A. Kendall, M Van Der Wilk, A. Shah, R. Cipolla, and A. Weller. Concrete problems for autonomous vehicle safety: advantages of Bayesian deep learning. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 4745–4753. AAAI Press, 2017.

[23] J. Garcıa and F. Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

[24] G. Ciravegna, F. Giannini, S. Melacci, M. Maggini, and M. Gori. A constraint-based approach to learning and explanation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34(04), pages 3658–3665, 2020.

[25] Cotter. A. et al. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research*, 20(172):1–59, 2019.

[26] P. Hitzler et al. Neural-symbolic learning and reasoning: A survey and interpretation. *Frontiers in artificial intelligence and applications*, 342:1–51, 2022.

[27] M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, and U. Topcu. Safe reinforcement learning via shielding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32(1), 2018.

[28] E. H. Shortliffe and M. J. Sepúlveda. Clinical decision support in the era of artificial intelligence. *Jama*, 320(21):2199–2200, 2018.

[29] P. Kairouz and H. B. McMahan. Advances and open problems in federated learning. *Foundations and trends in machine learning*, 14(1-2):1–210, 2021.

[30] M. Abadi et al. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[31] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy*, pages 39–57. Ieee, 2017.

[32] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *Stat*, 1050:9, 2017.

[33] P. Chen, H. Zhang, Y. Sharma, J. Yi, and C. Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 15–26, 2017.

[34] J. C. Costa, T. Roxo, H. Proença, P. Inacio, and M. Ricardo. How deep learning sees the world: A survey on adversarial attacks & defenses. *IEEE Access*, 12:61113–61136, 2024.

[35] E. Wong and Z. Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International conference on machine learning*, pages 5286–5295. PMLR, 2018.

[36] J. Cohen, E. Rosenfeld, and Z. Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR, 2019.

[37] J. Ruan et al. Robust universal adversarial perturbation attacks on renewable energy forecasting. *IEEE Internet of Things Journal*, 2025.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

[38] D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[39] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*, 2018.

[40] J. Ruan, H. Cui, Y. Huang, T. Li, C. Wu, and K. Zhang. A review of occluded objects detection in real complex scenarios for autonomous driving. *Green energy and intelligent transportation*, 2(3):100092, 2023.

[41] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

[42] K. R. Varshney. Engineering safety in machine learning. In *2016 Information Theory and Applications Workshop (ITA)*, pages 1–5. IEEE, 2016.

[43] J. Zhang, M. Harman, L. Ma, and Y. Liu. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering*, 2020.

[44] L. Ma et al. Deepgauge: Multi-level robustness testing for deep learning systems. In *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*, pages 120–131, 2018.

[45] K. Pei, Y. Cao, J. Yang, and S. Jana. Deepxplore: Automated whitebox testing of deep learning systems. In *Proceedings of the 26th Symposium on Operating Systems Principles*, pages 1–18, 2017.

[46] A. Rai. Explainable ai: From black box to glass box. *Journal of the Academy of Marketing Science*, 48:137–141, 2020.

[47] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[48] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3319–3328, 2017.

[49] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 2019.

[50] I. D. Raji et al. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *FAT* (ACM Conference on Fairness, Accountability, and Transparency)*, pages 33–44, 2020.

[51] Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (AI Act), 2024. European Union.

[52] NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0), 2023. National Institute of Standards and Technology.

[53] A. Brosch, A. Eldesokey, M. Felsberg, and K. Maag. Out-of-Distribution Segmentation via Wasserstein-Based Evidential Uncertainty. In *Procs. of the 34^{th} European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2026)*, 2026.

[54] B. Franke, N. Förster, F. Köster, A. Fischer, M. Lange, and A. Raulf. Revisiting Neural Activation Coverage for Uncertainty Estimation. In *Procs. of the 34^{th} European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2026)*, 2026.

[55] I. Zhang and D. Yifan. Neuromodulated Delta Adapters: Stabilizing Test-Time Adaptation via Gated Error Correction. In *Procs. of the 34^{th} European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2026)*, 2026.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

[56] L. Pereira, D. Valadares, M. Perkusich, and K. Gorgônio. On the Impact of Differential Privacy on Federated Neuromorphic Learning Accuracy. In *Procs. of the 34th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2026)*, 2026.

[57] J. Hoscilowicz and A. Janicki. Adversarial Confusion Attack: Disrupting Multimodal Large Language Models. In *Procs. of the 34th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2026)*, 2026.

[58] T. Ho, N. Nguyen, and B. Le. SPARC: Superpixel-based Black-Box Adversarial Attack with Regional Confidence. In *Procs. of the 34th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2026)*, 2026.

[59] L. Pajola, R. Petrucci, F. Marchiori, L. Pasa, and M. Conti. Model Selection Hijacking Adversarial Attack. In *Procs. of the 34th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2026)*, 2026.

[60] F. Leopardi, F. Labbaf, T. Kolarik, M. Wand, and N. Sharygina. Interpreting Logical Explanations of Classifying Neural Networks. In *Procs. of the 34th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2026)*, 2026.

[61] A. Borigi, P. Lindner, A. Schlager, S. Shoaib, R. Lekkala, S. S. Bhamidipati, and A. Lendasse. The Alignment Gate: Intent and Instruction Guardrails for Agentic AI. In *Procs. of the 34th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2026)*, 2026.

[62] M. Ahsan, D. Tawadros, S. Sadiya, P. Q. Le, J. Schlötterer, C. Seifert, and G. Roig. Weakly Supervised Shortcut Learning Mitigation Using Sparse Autoencoders. In *Procs. of the 34th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2026)*, 2026.