ETSI TR 103 910 V1.1.1 (2025-02)

Methods for Testing and Specification (MTS);
AI Testing;
Test Methodology and Test Specification for ML-based Systems

Reference
DTR/MTS-103910

Keywords
AI, ML, testing

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - APE 7112B
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° w061004871

Important notice

The present document can be downloaded from the ETSI Search & Browse Standards application.

The present document may be made available in electronic versions and/or in print. The content of any electronic and/or print versions of the present document shall not be modified without the prior written authorization of ETSI. In case of any existing or perceived difference in contents between such versions and/or in print, the prevailing version of an ETSI deliverable is the one made publicly available in PDF format on ETSI deliver repository.

Users should be aware that the present document may be revised or have its status changed, this information is available in the Milestones listing.

If you find errors in the present document, please send your comments to the relevant service listed under Committee Support Staff.

If you find a security vulnerability in the present document, please report it through our Coordinated Vulnerability Disclosure (CVD) program.

Notice of disclaimer & limitation of liability

The information provided in the present deliverable is directed solely to professionals who have the appropriate degree of experience to understand and interpret its content in accordance with generally accepted engineering or other professional standard and applicable regulations.
No recommendation as to products and services or vendors is made or should be implied.
No representation or warranty is made that this deliverable is technically accurate or sufficient or conforms to any law and/or governmental rule and/or regulation and further, no representation or warranty is made of merchantability or fitness for any particular purpose or against infringement of intellectual property rights.
In no event shall ETSI be held liable for loss of profits or any other incidental or consequential damages.

Any software contained in this deliverable is provided "AS IS" with no warranties, express or implied, including but not limited to, the warranties of merchantability, fitness for a particular purpose and non-infringement of intellectual property rights and ETSI shall not be held liable in any event for any damages whatsoever (including, without limitation, damages for loss of profits, business interruption, loss of information, or any other pecuniary loss) arising out of or related to the use of or inability to use the software.

Copyright Notification

Contents

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The declarations pertaining to these essential IPRs, if any, are publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI IPR online database.

Pursuant to the ETSI Directives including the ETSI IPR Policy, no investigation regarding the essentiality of IPRs, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

DECT™, **PLUGTESTS™**, **UMTS™** and the ETSI logo are trademarks of ETSI registered for the benefit of its Members. **3GPP™**, **LTE™** and **5G™** logo are trademarks of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners. **oneM2M™** logo is a trademark of ETSI registered for the benefit of its Members and of the oneM2M Partners. **GSM**® and the GSM logo are trademarks registered and owned by the GSM Association.

Foreword

This Technical Report (TR) has been produced by ETSI Technical Committee Methods for Testing and Specification (MTS).

Modal verbs terminology

In the present document "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

Executive summary

The present document covers testing of ML-based systems for the purpose of standardization and elaborates on test methodologies and methods for test specification. It identifies requirements for testing and comes forward with proposals to tackle the technical aspects of certifying trustworthiness of ML-based systems in standardization contexts.

Introduction

Machine Learning (ML) and especially the application of Neural Networks (NNs) have led to amazing successes in recent years due to the availability of large amounts of data as well as the increase in computing capacity. These successes include applications from image recognition, which now achieve better results than humans in many areas, the almost human-like abilities of speech recognition and conversation, which were finally demonstrated convincingly by the NLP model GPT-3, or the massive superiority of algorithmic decision systems in learning and playing strategic games such as Go, demonstrated by the Google subsidiary DeepMind.

With the increasing success of ML and NNs, there is a growing need to integrate ML models and NNs into software systems that are developed for critical tasks and operate in critical environments. At this point at the latest, the question arises as to how ML, especially NN as well as their integration into systems can be rigorously tested and quality assured. The present document describes methods and approaches for testing such ML-based applications.

The present document intentionally focuses on ML as the currently most widely spread method in the field of Artificial Intelligence (AI). Other methods, such as symbolic AI, have their justification, but are not used to the same extent as is currently the case with ML.

In summary, the present document provides an introduction as well as procedural understanding of testing ML-based systems. It presents principles and challenges for testing ML-based systems, quality criteria and test items as well as suitable test methods and their integration into the life cycle of typical ML-based applications, with relevance for industry, regulation and research.

1 Scope

The present document describes test types, test items, quality criteria, and testing methodologies associated with testing ML-based systems, with an emphasis on supervised, unsupervised, and reinforcement learning. The present document outlines how these testing practices can be effectively integrated into the life cycle of typical ML-based systems. The present document applies to all types of organizations involved in any of the lifecycle stages of developing and operating ML-based systems as well as to any other stakeholder roles.

2 References

2.1 Normative references

Normative references are not applicable in the present document.

2.2 Informative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

NOTE: While any hyperlinks included in this clause were valid at the time of publication ETSI cannot guarantee their long-term validity.

The following referenced documents are not necessary for the application of the present document, but they assist the user with regard to a particular subject area.

[i.1] ETSI TR 104 066 (V1.1.1) (07-2024): "Securing Artificial Intelligence; Security Testing of AI".

[i.2] Wang, Richard Y., and Diane M. Strong (1996): "Beyond Accuracy: What Data Quality Means to Data Consumers", Journal of Management Information Systems 12 (4): pp. 5-33.

[i.3] Zhang, J. M., Harman, M., Ma, L. & Liu, Y: "Machine Learning Testing: Survey, Landscapes and Horizons". arXiv:1906.10742 [cs, stat] (2019).

[i.4] V. Riccio, G. Jahangirova, A. Stocco, N. Humbatova, M. Weiss, and P. Tonella: "Testing machine learning based systems: a systematic mapping", Empir Software Eng, Bd. 25, Nr. 6, S. 5193-5254, November 2020. doi: 10.1007/s10664-020-09881-0.

[i.5] M. Pol, T. Koomen, and A. Spillner: "Management and optimisation of the testing process: a practical guide to successful software testing with TPI and TMap", updated ed. Heidelberg: dpunkt-Verl, 2002.

[i.6] Poddey A., Brade T., Stellet J. E. & Branz W: "On the validation of complex systems operating in open contexts", arXiv:1902.10517 [cs], 2019.

[i.7] Gal, Yarin: "Uncertainty in Deep Learning", University of Cambridge, October 13th, 2016 .

[i.8] Madry et.al.: "Adversarial Examples Are Not Bugs, They Are Features", arXiv:1905.02175v4.

[i.9] Sahil Verma and Julia Rubin. 2018: "Fairness definitions explained". In Proceedings of the International Workshop on Software Fairness (FairWare '18). Association for Computing Machinery, New York, NY, USA, 1–7. doi: 10.1145/3194770.3194776.

[i.10] M. Borg: "The AIQ Meta-Testbed: Pragmatically Bridging Academic AI Testing and Industrial Q Needs", arXiv:2009.05260 [cs], September 2020, Zugegriffen: October 13, 2021. [Online].

[i.11] ISO 21448:2022: "Road vehicles — Safety of the intended functionality".

[i.12] ISO 26262:2018: "Road vehicles — Functional safety".

[i.13] ISO/IEC/IEEE 24765™:2017: "Systems and software engineering — Vocabulary", in ISO/IEC/IEEE 24765™:2017I , vol., no., pp.1-541, 28 August 2017, doi: 10.1109/IEEESTD.2017.8016712.

[i.14] Beuth Verlag. (2023): "Managing and understanding artificial intelligence: The Practical Guide for Decision Makers, Developers and Regulators", Beuth Verlag.

[i.15] Shen Y., Song K., Tan X., Li D., Lu W., & Zhuang Y. (2024): "Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face". Advances in Neural Information Processing Systems, 36.

[i.16] Islam S., Elmekki H., Elsebai A., Bentahar J., Drawel N., Rjoub G., & Pedrycz W. (2023): "A comprehensive survey on applications of transformers for deep learning tasks", Expert Systems with Applications, 122666.

[i.17] Sterner P., Friemelt B., Goretzko,D., Kraus E., Bühner M., & Pargent F. (2024): "The confidence/significance level implies a certain cost ratio between error 1. type and error 2. type", Diagnostica.

[i.18] Haneke U., Trahasch S., Zimmer M., & Felden C. (2021): "Data science: basics, architectures and applications", dpunkt Verlag.

[i.19] Chicco D., Warrens M. J., & Jurman G. (2021): "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation", Peerj computer science, 7, e623.

[i.20] Willmott C. J., & Matsuura K. (2005): "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance", Climate research, 30(1), pp. 79-82.

[i.21] Mathias S. G., Großmann D., & Sequeira G. J. (August 2019): "A comparison of clustering measures on raw signals of welding production data", In 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML) (pp. 55-60). IEEE.

[i.22] Prasad M., Jagadeeshwar M., & Shanthi D. (2024): "A Comparative Study Of K-Medoids And Fuzzy K-Means Clustering For The Selection Of Optimal Cloud Service Provider", Educational Administration: Theory and Practice, 30(4), pp. 8045-8051.

[i.23] Mahadevan S. (1996): "Average reward reinforcement learning: Foundations, algorithms, and empirical results", Machine learning, 22(1), pp. 159-195.

[i.24] Yu M., Yang Z., Kolar M., & Wang Z. (2019): "Convergent policy optimization for safe reinforcement learning", Advances in Neural Information Processing Systems, 32.

[i.25] Littman M. L., & Szepesvári C. (July 1996): "A generalized reinforcement-learning model: Convergence and applications", In ICML (Vol. 96, pp. 310-318).

[i.26] Guerraoui R., Gupta N., & Pinot R. (2024): "Robust Machine Learning".

[i.27] Wang M., Yang N., Gunasinghe D. H., & Weng N. (2023): "On the Robustness of ML-Based Network Intrusion Detection Systems: An Adversarial and Distribution Shift Perspective", Computers, 12(10), 209.

[i.28] Brown O., Curtis A., & Goodwin J. (2021): "Principles for evaluation of ai/ml model performance and robustness", arXiv preprint arXiv:2107.02868.

[i.29] Braiek H. B., & Khomh F. (2024): "Machine Learning Robustness: A Primer", arXiv preprint arXiv:2404.00897.

[i.30] Thams N., Oberst M., & Sontag D. (2022): "Evaluating robustness to dataset shift via parametric robustness sets", Advances in Neural Information Processing Systems, 35, pp. 16877-16889.

[i.31] Hort M., Chen Z., Zhang J. M., Harman M., & Sarro F. (2023): "Bias mitigation for machine learning classifiers: A comprehensive survey", ACM Journal on Responsible Computing.

[i.32] Mehrabi N., Morstatter F., Saxena N., Lerman K., & Galstyan A. (2021): "A survey on bias and fairness in machine learning", ACM computing surveys (CSUR), 54(6), pp. 1-35.

[i.33] Kersten H., Reuter J., Schröder K. W., & Wolfenstetter K. D. (2008): "IT security management in accordance with ISO 27001 and IT-Grundschutz", Vieweg.

[i.34] Maurer U., Rüedlinger A., & Tackmann B. (2012): "Confidentiality and integrity: A constructive perspective", In Theory of Cryptography: 9th Theory of Cryptography Conference, TCC 2012, Taormina, Sicily, Italy, March 19-21, 2012. Proceedings 9 (pp. 209-229). Springer Berlin Heidelberg.

[i.35] Xu L., Jiang C., Wang J., Yuan J., & Ren Y. (2014): "Information security in big data: privacy and data mining", IEEE Access, 2, pp. 1149-1176.

[i.36] Wiyatno R. R., Xu A., Dia O., & De Berker A. (2019): "Adversarial examples in modern machine learning: A review", arXiv preprint arXiv:1911.05268.

[i.37] Brendel W., Rauber J., & Bethge M. (2017): "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models", arXiv preprint arXiv:1712.04248.

[i.38] Hanif H., Nasir M. H. N. M., Ab Razak M. F., Firdaus A., & Anuar N. B. (2021): "The rise of software vulnerability: Taxonomy of software vulnerabilities detection and machine learning approaches", Journal of Network and Computer Applications, 179, 103009.

[i.39] Akkiraju et al. (2018): "Characterizing machine learning process: A maturity framework", arXiv:1811.04871 [cs], November 2018.

[i.40] Amershi Saleema, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann (2019): "Software Engineering for Machine Learning: A Case Study", in 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP), pp. 291-300. Montreal, QC, Canada: IEEE. doi: 10.1109/ICSE-SEIP.2019.00042.

[i.41] Studer, Thanh, Drescher, Hanuschkin, Winkler, Peters, and Mueller: "Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology", arXiv:2003.05155 [cs, stat], March 2020.

[i.42] ISO/IEC/IEEE 29119-2™:2021: "Software and systems engineering — Software testing — Part 2: Test processes".

[i.43] ISO/IEC 25059:2023: "Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems".

[i.44] ISO/IEC 25010:2011: "Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) —System and software quality models".

[i.45] ISO/IEC/IEEE 29148™:2018: "Systems and software engineering — Life cycle processes — Requirements engineering".

[i.46] T. Y.Chen, S. C. Cheung, and S. M. Yiu (2020): "Metamorphic Testing: A New Approach for Generating Next Test Cases". doi: 10.48550/ARXIV.2002.12543.

[i.47] Gerrard P. and Thompson N. (2002): "Risk-based e-business testing, Artech House Publishers".

[i.48] Großmann Jürgen, Felderer Michael, Viehmann Johannes, Schieferdecker Ina: "A taxonomy to assess and tailor risk-based testing in recent testing standards", In: IEEE Software, vol. 37 (2020), no. 1, pp. 40-49.

[i.49] Felderer Michael, Großmann Jürgen, Schieferdecker Ina: "Recent advances in classifying risk-based testing approaches", In: Ruggeri, Fabrizio (Ed.): "Analytic Methods in Systems and Software Testing". New York: Wiley-Blackwell, 2018, pp. 1-25.

[i.50] Felderer M. and Ramler R. (2016): "Risk orientation in software testing processes of small and medium enterprises: an exploratory and comparative study", Software Quality Journal, 24 (3), pp. 519-548.

[i.51] Erdogan G., Li Y., Runde R., Seehusen F., Stølen K.: "Approaches for the combined use of risk analysis and testing: A systematic literature review", In International Journal on Software Tools for Technology Transfer, volume 16, pp. 627-642, 2014.

[i.52] ETSI EG 203 251 (01-2016): "Methods for Testing & Specification; Risk-based Security Assessment and Testing Methodologies".

[i.53] Paul Schwerdtner et.al.: "Risk Assessment for Machine Learning Models", arXiv:2011.04328v1.

[i.54] Oliveira S. R., & Zaïane O. R. (11-2003): "Protecting sensitive knowledge by data sanitization". In 3rd IEEE International conference on data mining, pp. 613-616.

[i.55] Arniban Charkroboty et. al.: "Adversarial Attacks and tnces: A Survey". Xiv:1810.00069v1.

[i.56] C. Gladisch, C. Heinzemann, M. Herrmann and M. Woehrle: "Leveraging combinatorial testing for safety-critical computer vision datasets", 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 2020, pp. 1314-1321, doi: 10.1109/CVPRW50498.2020.00170.

[i.57] G. Bernot, L. Bouaziz, and P. Le Gall: "A theory of probabilistic functional testing", in Proceedings of the 19th international conference on Software engineering - ICSE '97, Boston, Massachusetts, United States: ACM Press, 1997, S. 216-226. doi: 10.1145/253228.253273.

[i.58] Pietrantuono R., Russo S. (2018): "Probabilistic Sampling-Based Testing for Accelerated Reliability Assessment", in: 2018 IEEE International Conference on Software Quality, Reliability and Security (QRS). Presented at the 2018 IEEE International Conference on Software Quality, Reliability and Security (QRS), IEEE, Lisbon, pp. 35-46. doi: 10.1109/QRS.2018.00017.

[i.59] Wiesbrock H.-W., Großmann J. (2024): "Outline of an Independent Systematic Blackbox Test for ML-based Systems", arXiv, 2024. doi: 2401.17062.

[i.60] Zhang S., Pan Y., Liu Q., Yan Z., Choo K. K. R., & Wang G. (2024): "Backdoor Attacks and Defenses Targeting Multi-Domain AI Models: A Comprehensive Review". ACM Computing Surveys.

[i.61] M. D. Davis and E. J. Weyuker: "Pseudo-oracles for non-testable programs", Proceedings of the ACM '81 conference on - ACM 81, 1981. doi:10.1145/800175.809889.

[i.62] S. Segura, G. Fraser, A. B. Sanchez, and A. Ruiz-Cortes: "A survey on metamorphic testing", IEEE Transactions on Software Engineering, vol. 42, no. 9, pp. 805-824, 2016. doi: 10.1109/tse.2016.2532875.

[i.63] W. M. McKeeman: "Differential Testing for Software", Digit. Tech. J., pp. 100-107, 1998.

[i.64] C. Murphy, G. E. Kaiser, and M. Arias: "An Approach to Software Testing of Machine Learning Applications", International Conference on Software Engineering and Knowledge Engineering, 2007.

[i.65] D. Marijan and A. Gotlieb: "Software testing for Machine Learning", Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 09, pp. 13576-13582, 2020. doi:10.1609/aaai.v34i09.7084.

[i.66] R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner: "Detecting Adversarial Samples from Artifacts", ArXiv, 2017.

[i.67] J. Lin, L. L. Njilla, and K. Xiong: "Secure machine learning against adversarial samples at Test Time", EURASIP Journal on Information Security, vol. 2022, no. 1, 2022. doi: 10.1186/s13635-021-00125-2.

[i.68] I. J. Goodfellow, J. Shlens, and C. Szegedy: "Explaining and Harnessing Adversarial Examples", CoRR, 2014.

[i.69] Wang, Shengrong, Dongcheng Li, Hui Li, Man Zhao, and W. Eric Wong. (2024): "A Survey on Test Input Selection and Prioritization for Deep Neural Networks", In 2024 10th International Symposium on System Security, Safety, and Reliability (ISSSR), 232-43. doi: 10.1109/ISSSR61934.2024.00035.

[i.70] Dang, Xueqi, Yinghua Li, Mike Papadakis, Jacques Klein, Tegawendé F. Bissyandé, and Yves Le Traon (2024): "Test Input Prioritization for Machine Learning Classifiers", IEEE Transactions on Software Engineering 50 (3): 413-42. doi: 10.1109/TSE.2024.3350019.

[i.71] Mosin, Vasilii, Miroslaw Staron, Darko Durisic, Francisco Gomes de Oliveira Neto, Sushant Kumar Pandey, and Ashok Chaitanya Koppisetty (2022): "Comparing Input Prioritization Techniques for Testing Deep Learning Algorithms", In 2022 48th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), 76-83. doi: 10.1109/SEAA56994.2022.00020.

[i.72] Demir, Demet, Aysu Betin Can, and Elif Surer (2023): "Distribution Aware Testing Framework for Deep Neural Networks", IEEE Access 11: 119481-505. doi: 10.1109/ACCESS.2023.3327820.

[i.73] Monika Steidl, Ruth Breu, and Benedikt Hupfauf (2020): "Challenges in Testing Big Data Systems: An Exploratory Survey", In "Software Quality: Quality Intelligence in Software and Systems Engineering", published by Dietmar Winkler, Stefan Biffl, Daniel Mendez, and Johannes Bergsmann, 371:13-27. Lecture Notes in Business Information Processing. Cham: Springer International Publishing. doi: 10.1007/978-3-030-35510-4_2.

[i.74] Michael Felderer, Barbara Russo, and Florian Auer (2019): "On Testing Data-Intensive Software Systems", arXiv:1903.09413 [cs], April. doi: 1903.09413.

[i.75] English L.P.: "Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits", John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1999.

[i.76] Wang, Richard Y., and Diane M. Strong (1996): "Beyond Accuracy: What Data Quality Means to Data Consumers", Journal of Management Information Systems 12 (4): 5-33. doi: 10.1080/07421222.1996.11518099.

[i.77] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016): "'Why Should I Trust You?': Explaining the Predictions of Any Classifier", arXiv:1602.04938 [cs, stat], August. doi: 1602.04938.

[i.78] Bansal, Aayush, Ali Farhadi, and Devi Parikh (2014): "Towards Transparent Systems: Semantic Characterization of Failure Modes", In Computer Vision - ECCV 2014, published by David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, 8694:366-81. Lecture Notes in Computer Science. Cham: Springer International Publishing. doi: 10.1007/978-3-319-10599-4_24.

[i.79] Miller Tim (2018): "Explanation in Artificial Intelligence: Insights from the Social".

[i.80] Pei, Kexin, Yinzhi Cao, Junfeng Yang, and Suman Jana (2017): "DeepXplore: Automated Whitebox Testing of Deep Learning Systems", In Proceedings of the 26th Symposium on Operating Systems Principles, 1-18. Shanghai China: ACM. doi: 10.1145/3132747.3132785.

[i.81] Ma, Lei, Felix Juefei-Xu, Fuyuan Zhang, Jiyuan Sun, Minhui Xue, Bo Li, Chunyang Chen, u. a (2018): "DeepGauge: Multi-Granularity Testing Criteria for Deep Learning Systems", In Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering, 120-31. Montpellier France: ACM. doi: 10.1145/3238147.3238202.

[i.82] Sun, Youcheng, Xiaowei Huang, Daniel Kroening, James Sharp, Matthew Hill, and Rob Ashmore (2019): "Structural Test Coverage Criteria for Deep Neural Networks", In 2019 IEEE/ACM 41st International Conference on Software Engineering: Companion Proceedings (ICSE-Companion), 320-21. Montreal, QC, Canada: IEEE. doi: 10.1109/ICSE-Companion.2019.00134.

[i.83] Matthew Wicker, Xiaowei Huang, and Marta Kwiatkowska (2018): "Feature-Guided Black-Box Safety Testing of Deep Neural Networks", arXiv:1710.07859 [cs], February. doi: 1710.07859.

[i.84] Cheng, Chih-Hong, Georg Nührenberg, Chung-Hao Huang, Harald Ruess, and Hirotoshi Yasuoka (2018): "Towards Dependability Metrics for Neural Networks", arXiv:1806.02338 [cs, stat], June. doi: 1806.02338.

[i.85] Kim, Jinhan, Robert Feldt, and Shin Yoo (2018): "Guiding Deep Learning System Testing using Surprise Adequacy", arXiv:1808.08444 [cs], August. doi: 1808.08444.

[i.86] Huang, Xiaowei, Daniel Kroening, Wenjie Ruan, James Sharp, Youcheng Sun, Emese Thamo, Min Wu, and Xinping Yi (2020): "A Survey of Safety and Trustworthiness of Deep Neural Networks: Verification, Testing, Adversarial Attack and Defence, and Interpretability", arXiv:1812.08342 [cs], May. doi: 1812.08342.

[i.87] Dong, Yizhen, Peixin Zhang, Jingyi Wang, Shuang Liu, Jun Sun, Jianye Hao, Xinyu Wang, Li Wang, Jin Song Dong, and Dai Ting (2019): "There is Limited Correlation between Coverage and Robustness for Deep Neural Networks", arXiv:1911.05904 [cs, stat], November. doi: 1911.05904.

[i.88] Aldahdooh A., Hamidouche W., Fezza S. A., & Déforges O. (2022): "Adversarial example detection for DNN models: A review and experimental comparison". Artificial Intelligence Review, 55(6), 4403-4462.

[i.89] Hugging Face: "AI tasks", Retrieved June 23, 2024.

[i.90] Van der Maaten L., & Hinton G. (2008): "Visualizing data using t-SNE", Journal of machine learning research, 9(11).

[i.91] Anowar F., Sadaoui S., & Selim B. (2021): "Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne)", Computer Science Review, 40, 100378.

[i.92] Platzer A. (2013): "Visualization of SNPs with t-SNE", PloS one, 8(2), e56883.

[i.93] Norvig P. R., & Intelligence S. A. (2002): "A modern approach", Prentice Hall Upper Saddle River, NJ, USA: Rani M., Nayak R., & Vyas OP (2015): "An ontology-based adaptive personalized e-learning system, assisted by software agents on cloud storage", Knowledge-Based Systems, 90, pp. 33-48.

[i.94] Borgonovo E., & Plischke E. (2016): "Sensitivity analysis: A review of recent advances", European Journal of Operational Research", 248(3), pp. 869-887.

[i.95] Fasano G., & Franceschini A. (1987): "A multidimensional version of the Kolmogorov-Smirnov test", Monthly Notices of the Royal Astronomical Society, 225(1), pp. 155-170.

[i.96] Dunkelau J., & Duong M. K. (2022): "Towards equalised odds as fairness metric in academic performance prediction", arXiv preprint arXiv:2209.14670.

[i.97] Lin C. H., Lu M. C., Yang S. F., & Lee M. Y. (2021): "A Bayesian control chart for monitoring process variance", Applied Sciences, 11(6), 2729.

[i.98] Celis L. E., Keswani V. & Vishnoi N. (November 2020): "Data preprocessing to mitigate bias: A maximum entropy based approach", In International conference on machine learning, pp. 1349-1359. PMLR.

[i.99] Jiang B. (March 2018): "Approximate Bayesian computation with Kullback-Leibler divergence as data discrepancy", In International conference on artificial intelligence and statistics, pp. 1711-1721. PMLR.

[i.100] Monteiro R. P., Bastos-Filho C., Cerrada M., Cabrera D. R., & Sánchez R. V. (2021): "Using the Kullback-Leibler Divergence and Kolmogorov-Smirnov test to select input sizes to the fault diagnosis problem based on a CNN model", Learning and Nonlinear Models, 18(2), pp. 16-26.

[i.101] Vieira S. M., Kaymak U., & Sousa J. M. (July 2010): "Cohen's kappa coefficient as a performance measure for feature selection", In International conference on fuzzy systems, pp. 1-8. IEEE.

[i.102] Lakshminarayan K., Harp S. A., Goldman R. P., & Samad T. (August 1996): "Imputation of Missing Data Using Machine Learning Techniques", In KDD, vol. 96.

[i.103] García S., Ramírez-Gallego S., Luengo J., Benítez J. M., & Herrera F. (2016): "Big data preprocessing: methods and prospects", Big data analytics, 1, 1-22.

[i.104] Kalapanidas E., Avouris N., Craciun M., & Neagu D. (November 2003): "Machine learning algorithms: a study on noise sensitivity", In Proc. 1st Balcan Conference in Informatics, pp. 356-365. sn.

[i.105] Kertanah, K., Nurmayanti, W. P., Aini, S. R., Amrullah, L. M., & Sya'roni, M. (2023): "Comparison of Algorithms K-Means and DBSCAN for Clustering Student Cognitive Learning Outcomes in Physics Subject", Kappa Journal, 7(2), pp. 251-255.

[i.106] Zhu Z., Chen M., Zhu C., & Zhu Y. (2024): "Effective defense strategies in network security using improved double dueling deep Q-network", Computers & Security, 136, 103578.

[i.107] Simon D. (2001): "Kalman filtering", Embedded systems programming, 14(6), pp. 72-79.

[i.108] Joy T. T., Rana S., Gupta S., & Venkatesh S. (December 2016): "Hyperparameter tuning for big data using Bayesian optimisation", In 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 2574-2579. IEEE.

[i.109] Bergstra J., & Bengio Y. (2012): "Random search for hyper-parameter optimization", Journal of machine learning research, 13(2).

[i.110] Duesterwald E., Murthi A., Venkataraman G., Sinn M., & Vijaykeerthy D. (2019): "Exploring the hyperparameter landscape of adversarial robustness", arXiv preprint arXiv:1905.03837.

[i.111] Bonet D., Levin M., Montserrat D. M., & Ioannidis A. G. (2024): "Machine Learning Strategies for Improved Phenotype Prediction in Underrepresented Populations", In Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, vol. 29, p. 404. NIH Public Access.

[i.112] Misra P., & Yadav A. S. (2020): "Improving the classification accuracy using recursive feature elimination with cross-validation", Int. J. Emerg. Technol, 11(3), pp. 659-665.

[i.113] Husain H., Ciosek K., & Tomioka R. (March 2021): "Regularized policies are reward robust", In International Conference on Artificial Intelligence and Statistics, pp. 64-72. PMLR.

[i.114] Sinsomboonthong, S. (2022): "Performance Comparison of New Adjusted Min-Max with Decimal Scaling and Statistical Column Normalization Methods for Artificial Neural Network Classification", International Journal of Mathematics and Mathematical Sciences, 2022(1), 3584406.

[i.115] Osugi T., Kim D., & Scott S. (November 2005): "Balancing exploration and exploitation: A new algorithm for active machine learning", In 5th IEEE International Conference on Data Mining (ICDM'05), pp. 8-pp. IEEE.

[i.116] Johnson D. D., Blumstein D. T., Fowler J. H., & Haselton M. G. (2013): "The evolution of error: Error management, cognitive constraints, and adaptive decision-making biases", Trends in ecology & evolution, 28(8), pp. 474-481.

[i.117] Blum A., & Stangl K. (2019): "Recovering from biased data: Can fairness constraints improve accuracy?", arXiv preprint arXiv:1912.01094.

[i.118] Lynn P. (2019): "The advantage and disadvantage of implicitly stratified sampling", Methods, data, analyses: a journal for quantitative methods and survey methodology (mda), 13(2), pp. 253-266.

[i.119] Beretta L., & Santaniello A. (2016): "Nearest neighbor imputation algorithms: a critical evaluation", BMC medical informatics and decision making, 16, pp. 197-208.

[i.120] Ch'ng C. K., & Mahat N. I. (2020): "Winsorize tree algorithm for handling outlier in classification problem", International Journal of Operational Research, 38(2), pp. 278-293.

[i.121] Cheng K., & Young D. S. (2023): "An approach for specifying trimming and Winsorization Cutoffs", Journal of Agricultural, Biological and Environmental Statistics, 28(2), pp. 299-323.

[i.122] Salem A. M. G. (2022): "Adversarial inference and manipulation of machine learning models".

[i.123] Chivukula A. S., Yang X., Liu B., Liu W., & Zhou W. (2023): "Adversarial Machine Learning: Attack Surfaces, Defence Mechanisms, Learning Theories in Artificial Intelligence", Springer International Publishing.

[i.124] Bayani S. V., Prakash S., & Shanmugam L. (2023): "Data guardianship: Safeguarding compliance in AI/ML cloud ecosystems", Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 2(3), pp. 436-456.