Provide 20 test cases for evaluating deep research systems.  
   
Wed Apr 16 2025

**Functional Correctness and Accuracy Test**  
   
Verify that the deep research system produces correct and scientifically valid outputs under a set of well-defined, representative input conditions designed to reflect core use cases in the intended domain [26]. This type of test ensures that the model performs the intended tasks accurately, such as correct classifications, predictions, or recommendations, reflecting the correctness of both the algorithms and the data pathways.  
   
**Robustness Against Noisy and Incomplete Data**  
   
Assess how the system handles real-world imperfections by presenting it with datasets containing missing, incomplete, or noisy values [35]. The test evaluates whether the system can maintain accuracy, avoid silent failures, and employ techniques such as missingness embeddings or robust feature extraction to sustain performance when ideal data conditions are not met.  
   
**Scalability Under Increasing Data Volume**  
   
Measure the system’s runtime performance, accuracy, throughput, and resource utilization as the dataset size is incrementally scaled up to the maximum designed capacity [4]. This test identifies how well the system copes with “big data” challenges, ensuring responsiveness and avoiding bottlenecks as the volume of processed data grows.  
   
**User Load Scalability and Concurrent Usage Test**  
   
Simulate an environment with a rapidly increasing number of simultaneous user interactions or service requests to determine whether the system maintains responsiveness, stability, and does not degrade under heavy user or API load [66]. The aim is to establish performance thresholds and to surface resource contention issues inherent in multi-user, real-world deployment scenarios.  
   
**Cross-Platform and Framework Interoperability**  
   
Validate the system’s integration capabilities by deploying it across various operating systems, hardware configurations, or with multiple deep learning frameworks such as TensorFlow, PyTorch, or integrated third-party tools [94]. This evaluation ensures the system can seamlessly operate in diverse computing environments and exchange data with external modules correctly.  
   
**Usability and User Interface Evaluation**  
   
Conduct heuristic evaluations using recognized usability principles to determine the clarity, consistency, and intuitiveness of the user interface, as well as the ease of navigating through workflows for both novice and expert users [84]. Collect both qualitative feedback and quantitative usage metrics to identify design improvements.  
   
**Explainability and Interpretability Assessment**  
   
Examine the system’s ability to provide interpretable and understandable explanations for its predictions or decisions by leveraging frameworks such as LIME, SHAP, or symbolic reasoning modules [44]. This test focuses on user trust and compliance with regulatory or domain requirements for transparency in critical research decisions.  
   
**Adaptability to Novel Domains and Tasks**  
   
Test the extent to which the system can be repurposed for new, unseen domains or tasks—including domain adaptation scenarios where distribution shift is present between training and deployment data [28]. Measure performance, retraining requirements, and knowledge transfer effectiveness when deploying beyond the original problem setting.  
   
**Handling of Adversarial and Unstructured Inputs**  
   
Present carefully crafted adversarial examples or inputs with unexpected structure to challenge the system’s defenses and resilience against malicious manipulation or natural data aberrations [15]. This ensures the system remains reliable, even when facing intentionally misleading or anomalous data.  
   
**Anomaly and Outlier Detection Proficiency**  
   
Evaluate the robustness and sensitivity of the system in detecting unusual, rare, or outlier data points within noisy or high-dimensional research datasets by employing benchmarks from multiple application domains [30]. This capability is critical for scientific discovery and the reliable flagging of anomalous results.  
   
**Version Control and Experiment Reproducibility**  
   
Test the traceability, management, and recreation of experiment results through robust version control integration and experiment tracking [116]. Assessability should be possible even when altering data, code, or environmental parameters, ensuring scientific transparency and repeatability of research findings.  
   
**Security and Privacy Vulnerability Assessment**  
   
Conduct a comprehensive audit for security flaws, including checks for adversarial attacks, unauthorized data access, model theft, and privacy leakage in sensitive or regulated research scenarios [111]. This includes tests for compliance with frameworks like GDPR and application of privacy-preserving techniques.  
   
**Ethical Compliance and Bias Mitigation Testing**  
   
Evaluate system outputs for algorithmic bias, discrimination, and compliance with ethical standards using benchmark datasets featuring diversity in attributes [2]. The test also verifies the implementation of fairness-aware algorithms, audit logs, and user-controlled privacy settings.  
   
**Real-time Response and Latency Benchmarking**  
   
Measure average and worst-case latency for real-time inference tasks under normal and peak loads, ensuring that critical scientific workflows are not delayed and deadlines are met [103]. This includes monitoring round-trip processing times from data submission to result retrieval.  
   
**Fault Tolerance and Recovery**  
   
Simulate partial system failures, including hardware, software, and network disruptions, to test the system’s resilience, ability to isolate faults, and recover or gracefully degrade services without catastrophic loss of data or computation [17].  
   
**Multi-language and Internationalization Support**  
   
Assess the system’s capability to process, classify, or generate outputs in multiple natural languages, including correct text rendering, tokenization, classification, and support for specific domain terminology spanning several linguistic backgrounds [42]. This ensures global applicability and accessibility.  
   
**Documentation and Help Resource Accessibility**  
   
Evaluate the completeness, clarity, and usability of software documentation and help resources by verifying coverage of all features, error handling guides, and ease of access for both technical and non-technical end users [129].  
   
**Collaborative Workflow and Multi-user Scenario Testing**  
   
Test the system’s support for collaborative research workflows involving concurrent edits, access control, synchronization, and history tracing in multi-user scientific environments [83]. The system must facilitate and not hinder efficient group research productivity.  
   
**Benchmarking Against Gold Standard Datasets**  
   
Use well-established benchmark datasets and scenarios to compare system performance against recognized state-of-the-art metrics, enabling a standardized and reproducible evaluation of accuracy, speed, and robustness [24].  
   
**Automated Hypothesis Generation and Innovation Assessment**  
   
Evaluate the capacity of the system to generate, prioritize, and recommend novel, scientifically plausible hypotheses or research directions, including validation through cross-comparison with recent advances or expert review [134][136]. The aim is to quantify the value added by the system for cutting-edge exploration and discovery.  
   
---  
   
These test cases, drawn from a diverse range of technical literature and best practices, collectively provide a comprehensive framework to evaluate the efficacy, reliability, adaptability, and ethical responsibility of deep research systems in complex, real-world scientific environments.  
  
  
**参考文献**  
  
[1] A. J. Mansfield, J. Wayman, Dave Rayner, & Npl J L Wayman. (2002). Best Practices in Testing and Reporting Performance of Biometric Devices Produced for the Biometrics Working Group By. https://www.semanticscholar.org/paper/bef403c136beaa6fd43fc3184d4666512daaf9e5  
  
[2] A. Simkó, A. Garpebring, J. Jonsson, T. Nyholm, & Tommy Löfstedt. (2022). Reproducibility of the Methods in Medical Imaging with Deep Learning. In ArXiv. https://www.semanticscholar.org/paper/a95f2ab03ea71a0ee7044d94132fe114ceed8db7  
  
[3] Adil Mukhtar, Dietmar Jannach, & Franz Wotawa. (2024). Investigating Reproducibility in Deep Learning-Based Software Fault Prediction. In 2024 IEEE 24th International Conference on Software Quality, Reliability and Security (QRS). https://www.semanticscholar.org/paper/86676af22afab87971e278607bb019b88c7e496f  
  
[4] Ahmed Anwar, Brian B. Moser, Dayananda Herurkar, Federico Raue, Vinit Hegiste, T. Legler, & Andreas Dengel. (2024). FedAD-Bench: A Unified Benchmark for Federated Unsupervised Anomaly Detection in Tabular Data. In 2024 2nd International Conference on Federated Learning Technologies and Applications (FLTA). https://www.semanticscholar.org/paper/2b3332986061c120fef75e3a7967ca37e798c010  
  
[5] Aitor Arrieta. (2022). Multi-objective metamorphic follow-up test case selection for deep learning systems. In Proceedings of the Genetic and Evolutionary Computation Conference. https://www.semanticscholar.org/paper/3182bee673948beff52ed66d5699f54f7745f4f5  
  
[6] Ali Mohammed Omar Ali. (2025). Explainability in AI: Interpretable Models for Data Science. In International Journal for Research in Applied Science and Engineering Technology. https://www.semanticscholar.org/paper/d293115703f921708a80dbccaa8e5dffcc126af3  
  
[7] Amirata Ghorbani & James Y. Zou. (2018). Embedding for Informative Missingness: Deep Learning With Incomplete Data. In 2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton). https://www.semanticscholar.org/paper/52c43b42db1eb432f1c6fe65d5706d0c23876252  
  
[8] Amjan Shaik. (2011). A Suggestive Evaluation of System Test Cases in OO Systems Through Carving and Replaying Differential Unit Test Cases : A Metric Context. https://www.semanticscholar.org/paper/58cb8a6d70908c7611f048a1b69b6cdfc9ed0745  
  
[9] Amruta Sudhir Vatare & Pratibha Adkar. (2019). Review Paper on Centralized and Distributed Version Control System. https://www.semanticscholar.org/paper/547bb4669da1c5fefef20214e74ee00a7ed66a47  
  
[10] Anastasia-Maria Leventi-Peetz & T. Östreich. (2022). Deep Learning Reproducibility and Explainable AI (XAI). In ArXiv. https://www.semanticscholar.org/paper/6a1beca93452b40339ac27e1107a140318e4c787  
  
[11] Antoaneta Angelova-Stanimirova. (2023). CRITICAL CRITERIA FOR EVALUATION OF SCIENTIFIC RESEARCH ACTIVITY IN SCIENCE DIRECT. In KNOWLEDGE - International Journal. https://www.semanticscholar.org/paper/ebfed8b9e0d7c7fc9f07ab20faf6938932adee75  
  
[12] Arbi Ghazarian. (2013). Prediction of Functional Requirements Classes in Business Information Systems. https://www.semanticscholar.org/paper/be76b8abbfead5b0959f21ee7ff879d379888ac3  
  
[13] Arne Grobrugge, Nidhi Mishra, Johannes Jakubik, & G. Satzger. (2024). Explainability in AI-Based Applications – A Framework for Comparing Different Techniques. In 2024 26th International Conference on Business Informatics (CBI). https://www.semanticscholar.org/paper/9bab04c0f349115db3e823868593db58df81f038  
  
[14] Arunkumar L. Chandan. (n.d.). USER EXPERIENCE (UX) DESIGN PRINCIPLES FOR SOFTWARE APPLICATIONS. https://www.semanticscholar.org/paper/559e44a0af3d4e1890b93cf9cf2d601e47ed6db5  
  
[15] Asela Hevapathige. (2021). Evaluation of Deep Learning Approaches for Anomaly Detection. In 2021 5th SLAAI International Conference on Artificial Intelligence (SLAAI-ICAI). https://www.semanticscholar.org/paper/375ecf1316092bb603e5fc7a9ed837f89f6a4cc0  
  
[16] Bianca Montrosse-Moorhead. (2023). Evaluation criteria for artificial intelligence. In New Directions for Evaluation. https://www.semanticscholar.org/paper/2fea29dec73a48a3ac3a03a62266ec943b753b73  
  
[17] Bingyu Li, Zhi Li, & Yilong Yang. (2021). NFRNet: A Deep Neural Network for Automatic Classification of Non-Functional Requirements. In 2021 IEEE 29th International Requirements Engineering Conference (RE). https://www.semanticscholar.org/paper/37a50f1cefebcb11d83d0bf016c086fe61469511  
  
[18] Brijesh Singh, Manjula Neti, & Snigdhamayee Choudhury. (2024). Ethical Considerations in the Use of Deep Learning for HR Decision-Making. In 2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT). https://www.semanticscholar.org/paper/e438659307710b4d56ec1ba8c564afd523cd3164  
  
[19] C. landin, L. Hatvani, S. Tahvili, H. Haggrén, M., Längkvist, A. Loutfi, A. Håkansson, & C. landin. (2020). Performance Comparison of Two Deep Learning Algorithms in Detecting Similarities Between Manual Integration Test Cases. https://www.semanticscholar.org/paper/26f9be81c236fef11cc7139f909e576d997eb249  
  
[20] Carlos Alario-Hoyos & Scott Wilson. (2010). Comparison of the main alternatives to the integration of external tools in different platforms. https://www.semanticscholar.org/paper/316223d6cf7d8189cea27a8dd0bb158c099dd479  
  
[21] Chung‐Hyok Oh, Hongjun An, & Jongho Lee. (n.d.). Exploring reproducibility in deep learning-based parallel imaging reconstruction. In ISMRM Annual Meeting. https://www.semanticscholar.org/paper/c64b4eb11d2e7cb97fba4cd806b33cc73125c714  
  
[22] D. Brugali. (2019). Non-Functional Requirements in Robotic Systems: Challenges and State of the Art. In 2019 IEEE International Conference on Real-time Computing and Robotics (RCAR). https://www.semanticscholar.org/paper/305b1a2b6e0386e2f3cf125452ca5972c0000f4a  
  
[23] D. H. A. Ibrahim, Chiew Kang Leng, & Nadianatra Musa. (2017). Capturing Service Versioning in Provenance Trace to Support Reproducibility. In Journal of Telecommunication, Electronic and Computer Engineering. https://www.semanticscholar.org/paper/1d99d62a3dcf30d04dd11f699f6b0f81ccca6f49  
  
[24] D. H. A. Ibrahim, Nadianatra Musa, Chiew Kang Leng, J. Labadin, Johari Abdullah, & Sarina Sulaiman. (2017). Achieving Reproducibility Incorporating Service Versioning into Provenance Model. In Journal of Telecommunication, Electronic and Computer Engineering. https://www.semanticscholar.org/paper/16f3be500f2fe815464ecfad9d259a86aaec4064  
  
[25] D. Hussain & S. Haroon. (2015). Supporting Tool for Collaborative Scientific Workflow. https://www.semanticscholar.org/paper/b89ed20212f29dad2c8da33a0e68d05d5b379a9f  
  
[26] D. Yatsenyak, V. Oleksiuk, & N. Balyk. (2022). Study of ergonomic criteria for evaluating the software user interface. In Journal of Physics: Conference Series. https://www.semanticscholar.org/paper/fbdbd17e89f2d379f92d9fab4177cc88b8ca8f82  
  
[27] Daim Ali, Muhammad Kamran Abid, Muhammad Baqer, Yasir Aziz, Naeem Aslam, & Nasir Umer. (2025). IMPROVING THE EXPLAINABILITY AND TRANSPARENCY OF DEEP LEARNING MODELS IN INTRUSION DETECTION SYSTEMS. In Kashf Journal of Multidisciplinary Research. https://www.semanticscholar.org/paper/0f75eca3ec7ca7270f84cd047ed0308f4f0fd494  
  
[28] Deeksha Bhalla, K. Rangarajan, Tany Chandra, Subhashis Banerjee, & Chetan Arora. (2023). Reproducibility and Explainability of Deep Learning in Mammography: A Systematic Review of Literature. In The Indian Journal of Radiology & Imaging. https://www.semanticscholar.org/paper/9bf965ed145c2a1dd7dfbe34d17a29e6e477873e  
  
[29] Donghuo Chen, Xuandong Li, & Shizhong Zhao. (2010). Auto-generation and redundancy reduction of test cases for reactive systems. In 2010 2nd International Conference on Software Technology and Engineering. https://www.semanticscholar.org/paper/df49cd1574b7ca7035bcfd9991105ea76b9d37a5  
  
[30] E. A. Carvalho, J. O. Gomes, Alessandro Jatobá, Mônica Ferreira da Silva, & P. Carvalho. (2021). Software Requirements Elicitation for Complex Systems with the Functional Resonance Analysis Method (FRAM). In Proceedings of the XVII Brazilian Symposium on Information Systems. https://www.semanticscholar.org/paper/ed170217cea92d8eea119abdc860d88ad8a2ac83  
  
[31] E. Bonawitz & T. Griffiths. (2010). Deconfounding hypothesis generation and evaluation in Bayesian models. https://www.semanticscholar.org/paper/ace2178e8362cb012c459f85cf2541a173ad46cd  
  
[32] E. L. White. (1994). Search Hanford Accessible Reports Electronically system test plan and documentation: Revision 1. https://www.semanticscholar.org/paper/72c946dd080bc3206defd7da3526f7cc67fd4bb3  
  
[33] E. Z. Zeng, Hayden Gunraj, Sheldon Fernandez, & Alexander Wong. (2023). Explaining Explainability: Towards Deeper Actionable Insights into Deep Learning through Second-order Explainability. In ArXiv. https://www.semanticscholar.org/paper/597ff19287cca630fc8fc8af6206abb6411f0e7f  
  
[34] Erhu Liu, Song Huang, Cheng Zong, Changyou Zheng, Yongming Yao, Jing Zhu, Shi-Qi Tang, & Yanqiu Wang. (2021). MTGAN: Extending Test Case set for Deep Learning Image Classifier. In IEICE Trans. Inf. Syst. https://www.semanticscholar.org/paper/3f2a90068c5f4ef1ff6f90bcf0b21b1a83236b68  
  
[35] Erik Wilde. (2005). Augmenting XHTML for Help and Documentation. In International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC’06). https://www.semanticscholar.org/paper/617f209f2f38f563e33a776da9847af293890ec4  
  
[36] Florian Fankhauser & M. Ronniger. (2011). SECURITY TEST ENVIRONMENT FOR VOIP RESEARCH. https://www.semanticscholar.org/paper/b893d877babb9c79739897c39e914c554be8eb2e  
  
[37] G. A. Alencar, Felipe V. De S. Oliveira, Jorge da Silva Correia-Neto, & M. M. Teixeira. (2019). Non-Functional Requirements In Health Information Systems: In 2019 14th Iberian Conference on Information Systems and Technologies (CISTI). https://www.semanticscholar.org/paper/31425b1d2adb1d66f4614fa5c08e9b860b8510a6  
  
[38] G. A. V. Hakim, David Osowiechi, Mehrdad Noori, Milad Cheraghalikhani, Ali Bahri, Moslem Yazdanpanah, Ismail Ben Ayed, & Christian Desrosiers. (2024). CLIPArTT: Adaptation of CLIP to New Domains at Test Time. https://www.semanticscholar.org/paper/ceac663aa9213bfdb457deacba9e5ed9ddd7ae92  
  
[39] G. Marsal, B. Denis, Jean-Marc Faure, & G. Frey. (2006). Evaluation of response time in ethernet-based automation systems. In 2006 IEEE International Workshop on Factory Communication Systems. https://www.semanticscholar.org/paper/187ed1a52705558265ab49d730dad666bd2b83bc  
  
[40] G. Singh, P. Rana, & Parveen Kakkar. (2015). Evaluation of test cases using Ant Colony Optimization with Enhanced Heuristics:An Implementation. https://www.semanticscholar.org/paper/03e8bae8ccaa61bdd3f91e465d8ed4f2a1aaa960  
  
[41] G. Steinke, Ryan C. LaBrie, & Satadipa Sarkar. (2022). Recommendation for Continuous Ethical Analysis of AI Algorithms. In Proceedings of the 2022 European Interdisciplinary Cybersecurity Conference. https://www.semanticscholar.org/paper/08f2d600f4ea4f9a19c213cb075d702a1e8a28d9  
  
[42] Grace Hanna, Nathan Boyar, Nathan Garay, & Mina Maleki. (2024). Revolutionizing Requirements Elicitation: Deep Learning-Based Classification of Functional and Non-Functional Requirements. In 2024 IEEE 5th International Conference on Pattern Recognition and Machine Learning (PRML). https://www.semanticscholar.org/paper/2c6df3f2ee3196620115f34347fa731f2a57bddf  
  
[43] Hanqing Li, Xiang Li, Yuanping Nie, & Jianwen Tian. (2024). A Survey of Security Testing Techniques for Deep Learning Frameworks. In 2024 9th International Conference on Signal and Image Processing (ICSIP). https://www.semanticscholar.org/paper/cecbc3b6c5281ed027578f1dfe811ee8367f34b8  
  
[44] Hao Li, Shihai Wang, Tengfei Shi, Xinyue Fang, & Jian Chen. (2022). TSDTest: A Efficient Coverage Guided Two-Stage Testing for Deep Learning Systems. In 2022 IEEE 22nd International Conference on Software Quality, Reliability, and Security Companion (QRS-C). https://www.semanticscholar.org/paper/f1174431feae6350f5ab89933652ad0f3ae40ba9  
  
[45] I. Ho & Jin-Cherng Lin. (1998). A method of test cases generation for real-time systems. In Proceedings First International Symposium on Object-Oriented Real-Time Distributed Computing (ISORC ’98). https://www.semanticscholar.org/paper/5bd963ac02966321e433edf48e507ce4646adcc4  
  
[46] Isaac Shukurani Mwakabira, L. Byson, T. Manda, Y. Phiri, & Augustine Simwela. (2023). Design Gaps in Configurable Systems: Adaptability of DHIS2 to other Domains. In 2023 IST-Africa Conference (IST-Africa). https://www.semanticscholar.org/paper/aadd6b5627e21095ec997e3cedf23df84851a666  
  
[47] Ivan Jovanovikj, V. Narasimhan, G. Engels, & Stefan Sauer. (2018). Context-specific Quality Evaluation of Test Cases. In International Conference on Model-Driven Engineering and Software Development. https://www.semanticscholar.org/paper/785415fb74128435f7a581559dc47ef7fe9f1c4d  
  
[48] Ivelina L. Stoyanova, Hristina Kukova, Maria Andreeva Todorova, & Tsvetana Dimitrova. (2024). Multilingual Corpus of Illustrative Examples on Activity Predicates. In Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria. https://www.semanticscholar.org/paper/1461fbf9010288cc3b0a531f73ef7e9c3015ddfd  
  
[49] Ivy Osei & Kwame Mensah. (2024). Intelligent anomaly detection in distributed systems via deep learning. In World Journal of Information and Knowledge Management . https://www.semanticscholar.org/paper/23611acf1a1744a59485a2a2232bdb38c6381d19  
  
[50] Jarrett Booz, Wei Yu, Guobin Xu, D. Griffith, & N. Golmie. (2019). A Deep Learning-Based Weather Forecast System for Data Volume and Recency Analysis. In 2019 International Conference on Computing, Networking and Communications (ICNC). https://www.semanticscholar.org/paper/d6955d8bf0aec2d95963573b9125231e116758d5  
  
[51] Jason Li. (2019). GENERATING ARTIFICIAL DATA FOR SCALABILITY TEST OF A FOLLOWEE TWITTER RECOMMENDER SYSTEM. https://www.semanticscholar.org/paper/107c1aec3864fa7d0500b8385588b7693063dc8c  
  
[52] Jeong-Cheol Seo & Seung Jun Lee. (2018). Optimization of Test Cases for Experimental Reliability Evaluation of Digital Reactor Protection System. https://www.semanticscholar.org/paper/c3ca17555b14260645061f14af281e57f645478a  
  
[53] Jianxin Ge & Jiaomin Liu. (2016). Software test cases recommendation system research based on collaborative filtering. In 2016 17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD). https://www.semanticscholar.org/paper/44ac5db5b4afee31b601b72f517ee1ed18feafbc  
  
[54] Jihen Amara, B. König-Ries, & Sheeba Samuel. (2024). Explainability of Deep Learning-Based Plant Disease Classifiers Through Automated Concept Identification. In ArXiv. https://www.semanticscholar.org/paper/282b35a1a9e4bca2b0e22626fff34112a51940ee  
  
[55] Jiyan Salim Mahmud & Imre Lendák. (2024). Enhancing One-Class Anomaly Detection in Unlabeled Datasets Through Unsupervised Data Refinement. In 2024 IEEE 22nd Jubilee International Symposium on Intelligent Systems and Informatics (SISY). https://www.semanticscholar.org/paper/0220fb2e13602b34c01ab29d4c48be182306c783  
  
[56] Julian Wiederer, Julian Schmidt, U. Kressel, K. Dietmayer, & Vasileios Belagiannis. (2022). A Benchmark for Unsupervised Anomaly Detection in Multi-Agent Trajectories. In 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC). https://www.semanticscholar.org/paper/1c989d93b966cdd76734ce31e7ad14d89d418f4b  
  
[57] Jun Zhou. (2004). Functional requirements and non-functional requirements : a survey. https://www.semanticscholar.org/paper/dc2c04dde9a7d6a48725a24a91f59eee79945182  
  
[58] Justin Sybrandt & Ilya Safro. (2018). Validation and Topic-driven Ranking for Biomedical Hypothesis Generation Systems. In bioRxiv. https://www.semanticscholar.org/paper/3079dfff76bcaf3673571fd7ac661d35e1634087  
  
[59] K. Rubinov. (2010). Generating integration test cases automatically. In Fast Software Encryption Workshop. https://www.semanticscholar.org/paper/91c90ffe33eb1805903dabe6233c0b3450b72726  
  
[60] K. Smit, M. Zoet, & M. Berkhout. (2017). Functional Requirements for Business Rules Management Systems. In Americas Conference on Information Systems. https://www.semanticscholar.org/paper/14593acd1faf4c715b0700ea6eac11e68279423c  
  
[61] Kai Zhou. (2022). A Deep Long Short-Term Memory Network for Bearing Fault Diagnosis Under Time-Varying Conditions. In Volume 10: 34th Conference on Mechanical Vibration and Sound (VIB). https://www.semanticscholar.org/paper/2898f81df2ebded7e078055ede7f3febbbea7fe6  
  
[62] Karsten Stocker, H. Washizaki, & Y. Fukazawa. (2017). Closing the Gap between Unit Test Code and Documentation. In 2017 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW). https://www.semanticscholar.org/paper/b3c28839eba25e916ce1c607f95bf469f32bd02a  
  
[63] Kathia Marcal de Oliveira. (2010). New research challenges for user interface quality evaluation. In Proceedings of the 22nd Conference on l’Interaction Homme-Machine. https://www.semanticscholar.org/paper/527d5cfdf6cfb7b2bd69969fb28e18d5b6631f58  
  
[64] Kengo Tajiri, Tomoharu Iwata, Yoichi Matsuo, & Keishiro Watanabe. (2021). Fault Detection of ICT systems with Deep Learning Model for Missing Data. In 2021 IFIP/IEEE International Symposium on Integrated Network Management (IM). https://www.semanticscholar.org/paper/9f884172ac61a1b6521df2c072c676d3fec2e0f6  
  
[65] Kyeongbo Kong, Junggi Lee, Youngchul Kwak, Minsung Kang, Seong Gyun Kim, & Woo‐Jin Song. (2019). Recycling: Semi-Supervised Learning With Noisy Labels in Deep Neural Networks. In IEEE Access. https://www.semanticscholar.org/paper/45762403862fb7729cccffe5116a0ca348a2cddc  
  
[66] L. Ma, Felix Juefei-Xu, Jiyuan Sun, Chunyang Chen, Ting Su, Fuyuan Zhang, Minhui Xue, Bo Li, Li Li, Yang Liu, Jianjun Zhao, & Yadong Wang. (2018). DeepGauge: Comprehensive and Multi-Granularity Testing Criteria for Gauging the Robustness of Deep Learning Systems. In ArXiv. https://www.semanticscholar.org/paper/5563d7e6e5ee659bf26fa25e04d38c8d1a56e204  
  
[67] L. Marques, P. Matsubara, W. Nakamura, I. Wiese, L. Zaina, & T. Conte. (2019). UX-Tips: A UX evaluation technique to support the identification of software application problems. In Proceedings of the XXXIII Brazilian Symposium on Software Engineering. https://www.semanticscholar.org/paper/fdf93b9a47ae77d75fe1dae9942afd4e50b09cd5  
  
[68] Li Jia, Hao Zhong, & Linpeng Huang. (2021). The Unit Test Quality of Deep Learning Libraries: A Mutation Analysis. In 2021 IEEE International Conference on Software Maintenance and Evolution (ICSME). https://www.semanticscholar.org/paper/24a9472dc016dd8f745655ae82cbc9c36fa2f46c  
  
[69] M. Alazzam, Fawaz Alassery, & Ahmed H. Almulihi. (2022). Federated Deep Learning Approaches for the Privacy and Security of IoT Systems. In Wireless Communications and Mobile Computing. https://www.semanticscholar.org/paper/db9162ac7b35342a424da7259d0205317f63a8b3  
  
[70] M. C. Camargo, R. Barros, & Vanessa Tavares de Oliveira Barros. (2018). Visual design checklist for graphical user interface (GUI) evaluation. In Proceedings of the 33rd Annual ACM Symposium on Applied Computing. https://www.semanticscholar.org/paper/a700135212613b809f3db2d6ed131926e67bad1f  
  
[71] M. Tariq, N. Memon, Shakeel Ahmed, S. Tayyaba, M. T. Mushtaq, N. A. Mian, M. Imran, & M. Ashraf. (2020). A Review of Deep Learning Security and Privacy Defensive Techniques. In Mob. Inf. Syst. https://www.semanticscholar.org/paper/cb36b2a823a68fc5a7e7e459cf5006267a095c02  
  
[72] Manar Majthoub, Yousra Odeh, & Mohammed Hijjawi. (2020). Non-Functional Requirements Classification for Aligning Business with Information Systems. In Proceedings of the 2020 International Conference on Big Data in Management. https://www.semanticscholar.org/paper/d06d4b40a25323c9b42713acb4b05ce5dce7e84c  
  
[73] MandhavMundhra, Madhav Khode, Mandar Tare, & Ms. Priyanka Dhasal. (n.d.). Scalability and Performance Optimization in Online Community Platforms. https://www.semanticscholar.org/paper/45cfe73a14dac056c0723bf0f80e73be60f20ccf  
  
[74] Mehil B. Shah, Mohammad Masudur Rahman, & Foutse Khomh. (2024). Towards Enhancing the Reproducibility of Deep Learning Bugs: An Empirical Study. In Empir. Softw. Eng. https://www.semanticscholar.org/paper/aaf3c8d6df4529c9a0ca6bb385a1507fe3a4427b  
  
[75] Miaosen Chai, Emily Herron, Erick Cervantes, & Tirthankar Ghosal. (2024). Exploring Scientific Hypothesis Generation with Mamba. In Proceedings of the 1st Workshop on NLP for Science (NLP4Science). https://www.semanticscholar.org/paper/5ac80c35214f8d49625cb1c1d899846a65ef0599  
  
[76] Min Yan, Li Wang, & Aiguo Fei. (2020). ARTDL: Adaptive Random Testing for Deep Learning Systems. In IEEE Access. https://www.semanticscholar.org/paper/d1798d0394f901007626278e753d59c5c2115f2b  
  
[77] N. Hariri & Y. Norouzi. (2011). Determining evaluation criteria for digital libraries’ user interface: a review. In Electron. Libr. https://www.semanticscholar.org/paper/f3c7b379ed8b2f1d22080a3a0a6ee084fd4748fd  
  
[78] N. Rama & D. K. C. Sekharaiah. (2016). A Methodological Review Based Version Control System with Evolutionary Research for Software Processes. In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies. https://www.semanticscholar.org/paper/78ad72a34bf7223b7e1a6f40fc18221b208e88d1  
  
[79] N. Truong, Young-Joo Suh, & Chansu Yu. (2013). Latency Analysis in GNU Radio/USRP-Based Software Radio Platforms. In MILCOM 2013 - 2013 IEEE Military Communications Conference. https://www.semanticscholar.org/paper/fb21bafa1075e9198d8102cd08f5aba8bdee6188  
  
[80] Nidhi Singh & Craig Olinsky. (2017). Demystifying Numenta anomaly benchmark. In 2017 International Joint Conference on Neural Networks (IJCNN). https://www.semanticscholar.org/paper/541bb7fc1bb733602d3ad9f051b6fd9734ff59d2  
  
[81] Nikita Ishwar Dhanvijay. (2025). AI-Based Smart Utility Management System. In International Journal for Research in Applied Science and Engineering Technology. https://www.semanticscholar.org/paper/c490426180ef923db5f9e3b98a09f0fcc0ce8f77  
  
[82] Noha Mohamed, Sherin M. Moussa, N. Badr, & M. Tolba. (2021). Enhancing Test Cases Prioritization for Internet of Things based systems using Search-based Technique. In International Journal of Intelligent Computing and Information Sciences. https://www.semanticscholar.org/paper/1568cf4bb02f6f0201b4d36f414137ff46315386  
  
[83] O. Shpak, Welf Löwe, Anna Wingkvist, & Morgan Ericsson. (2014). A Method to Test the Information Quality of Technical Documentation on Websites. In 2014 14th International Conference on Quality Software. https://www.semanticscholar.org/paper/24f854e3dfc6896b99f048fac300eec04aa9a554  
  
[84] Obead Alhadreti. (2020). Exploring UX Maturity in Software Development Environments in Saudi Arabia. https://www.semanticscholar.org/paper/ce7aa65ded5359fe9e3a1bb2b0c8f3772493c265  
  
[85] Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, & Hans-Joachim Böhme. (2021). FullStop: Multilingual Deep Models for Punctuation Prediction. In Swiss Text Analytics Conference. https://www.semanticscholar.org/paper/fff1cbfad4d7ff43ac180638e0a7435a70c5beac  
  
[86] P. Chillakanti, S. Ekwaro-Osire, & A. Ertas. (2021). Evaluation of Technology Platforms for Use in Transdisciplinary Research. In Education Sciences. https://www.semanticscholar.org/paper/ca6cd7d1e6dfa17c8d41c112f4b7d82d5f8b1620  
  
[87] Pengwei Wang, Zhongyuan Wang, Lei Ji, Jun Yan, & Lianwen Jin. (2017). Can Machines Intelligently Propose Novel and Reasonable Scientific Hypotheses? In Proceedings of the 26th International Conference on World Wide Web Companion. https://www.semanticscholar.org/paper/f3f7d9dafa2d0fecc8f437ef77fe61e0fd25860e  
  
[88] R. Brillhart. (2002). Integration of Multiple Tools for Efficient Test Data Manipulation. In Sound and Vibration. https://www.semanticscholar.org/paper/c097d8853511438a7b8ee5c42d2675727936f437  
  
[89] Rajitha de Silva, Grzegorz Cielniak, & Junfeng Gao. (2021). Towards agricultural autonomy: crop row detection under varying field conditions using deep learning. In ArXiv. https://www.semanticscholar.org/paper/2bb63b9f053332389d2ca7bc02181bdd92edb9fc  
  
[90] Rajneesh Mahajan, Ramesh Govindarajulu, J. R. Armstrong, F. G. Gray, & rarnahaja. (2002). A multi-language goal-tree based functional test planning system. In Proceedings. International Test Conference. https://www.semanticscholar.org/paper/b1ccd84e7ae44de218bce30615b06a83018a6079  
  
[91] Ruilin Xie, Zhanqi Cui, Minghua Jia, Yuan Wen, & Baoshui Hao. (2020). Testing Coverage Criteria for Deep Forests. In 2019 6th International Conference on Dependable Systems and Their Applications (DSA). https://www.semanticscholar.org/paper/82856be2c8c4a3d5be31eda2fefd4c1f3172458d  
  
[92] Ruoying Wang, Kexin Nie, Tie Wang, Yang Yang, & Bo Long. (2020). Deep Learning for Anomaly Detection. In Proceedings of the 13th International Conference on Web Search and Data Mining. https://www.semanticscholar.org/paper/28631a1eb7cefd38ee5e153970ef04e1f28c13f1  
  
[93] Rustam Saakyan & Ani Saghatelyan. (2024). Development of Tools for Integration Testing of Web Application Interfaces. In Scientific Proceedings of Vanadzor State University: “Natural and Exact Sciences. https://www.semanticscholar.org/paper/9733af8c620a8ec92c4c42cef21f5104e8ddd8a7  
  
[94] S. Atkins, Ishwarradj Badrie, & S. Otterloo. (2021). Applying Ethical AI Frameworks in practice: Evaluating conversational AI chatbot solutions. In Computers and Society Research Journal. https://www.semanticscholar.org/paper/3e1d3287ad2e717133bbb28b8a2e3318a88515d7  
  
[95] S. Jezek & Radim Burget. (2024). Deployment of deep learning-based anomaly detection systems: challenges and solutions. In Proceedings II of the 30st Conference STUDENT EEICT 2024: Selected papers. https://www.semanticscholar.org/paper/fb68e82689b39658db440a2ea7b17e08141301d9  
  
[96] S. Mathur & Gourav Shrivastava. (2024). M2PSC: Multilingual sentiment analysis using improved multi-attention based Deep Learning model. In Intell. Decis. Technol. https://www.semanticscholar.org/paper/1e912a59e70cf912c34a45494e0e8fd26fe06ec3  
  
[97] S. Noh & Kee-chun Bang. (2009). A Study for a Method of Response Time Evaluation Through Latency Test in Intranet System. https://www.semanticscholar.org/paper/749799d13c5243e29dddab9c0b3b427fab470a5b  
  
[98] S. Wagner, C. Matek, Sayedali Shetab Boushehri, M. Boxberg, L. Lamm, A. Sadafi, Dominik Jens Elias Waibel, C. Marr, & T. Peng. (2022). Built to last? Reproducibility and Reusability of Deep Learning Algorithms in Computational Pathology. In Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc. https://www.semanticscholar.org/paper/c0ea6d908d15af46148713dafbc3366833b5f776  
  
[99] S. Zhong, Yan Liu, & K. Hua. (2016). Field Effect Deep Networks for Image Recognition with Incomplete Data. In ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM). https://www.semanticscholar.org/paper/d44f2e6e19b7a006cd30f4f03f4c38c97c9f9eae  
  
[100] Safa Taieb, Layth Sliman, Benoit Charroux, & Yvan Stroppa. (2013). Towards a Total Traceability of Research Collaboration: a New Distributed Research Oriented Version Control System. https://www.semanticscholar.org/paper/846157d894990797858e8cc21c5b285e32089f0c  
  
[101] Salma Messaoudi, Donghwan Shin, Annibale Panichella, D. Bianculli, & L. Briand. (2021). Log-based slicing for system-level test cases. In Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis. https://www.semanticscholar.org/paper/f731c35604f0246ffe343fc41678b9f4aee3bf67  
  
[102] Sandeep Konakanchi. (2024). Next-Generation Low-Latency Architectures for Real-Time AI-Driven Cloud Services. In International Journal of Scientific Research in Computer Science, Engineering and Information Technology. https://www.semanticscholar.org/paper/9f9a76626e7be81215b7b57452e049c64ddf5f12  
  
[103] Santosh Kumar Jawalkar. (2022). Ethical QA Practices: Addressing Bias and Ensuring Compliance in Software Testing Frameworks. In International Journal of Social Science Exceptional Research. https://www.semanticscholar.org/paper/e335dfd6d35ddeeffa31277c7709733ac3138c7d  
  
[104] Selem Charfi, H. Ezzedine, & C. Kolski. (2015). RITA: a useR Interface evaluaTion frAmework. In J. Univers. Comput. Sci. https://www.semanticscholar.org/paper/72545a91c04b293daf8f0152b7272f4195af236a  
  
[105] Seulki Lee & S. Nirjon. (2021). Deep Functional Network (DFN): Functional Interpretation of Deep Neural Networks for Intelligent Sensing Systems. In Proceedings of the 20th International Conference on Information Processing in Sensor Networks (co-located with CPS-IoT Week 2021). https://www.semanticscholar.org/paper/dc44e3791447b84c6fbc1686b20977dda50ac935  
  
[106] Shilei Liu. (2001). Generating Test Cases from Software Documentation Title: Generating Test Cases from Software Documentation. https://www.semanticscholar.org/paper/7e7f0b6f55cd23169a51c65bb4e9858943282c91  
  
[107] Shiyong Lu & Jia Zhang. (2009). Collaborative Scientific Workflows. In 2009 IEEE International Conference on Web Services. https://www.semanticscholar.org/paper/04ea4f1cdba99cb88c57e85d3858a16d2ac8eb9d  
  
[108] Siqi Gu, Jiawei Liu, Zhan-wei Hui, Wenhong Liu, & Zhenyu Chen. (2022). MetaA: Multi-Dimensional Evaluation of Testing Ability via Adversarial Examples in Deep Learning. In 2022 IEEE 22nd International Conference on Software Quality, Reliability and Security (QRS). https://www.semanticscholar.org/paper/4a68b6236af2d197f2e9209c415b609bc43e3fc7  
  
[109] Sriparna Banerjee & Dipika Shende. (2023). MLing-Net: A computationally inexpensive deep neural framework designed to perform multilingual image document classification. In 2023 7th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech). https://www.semanticscholar.org/paper/a0c471159fe2889c15db31cbd52b2d8656a00da9  
  
[110] Stephen Casper, Yuxiao Li, Jiawei Li, Tong Bu, Kevin Zhang, & Dylan Hadfield-Menell. (2023). Benchmarking Interpretability Tools for Deep Neural Networks. In ArXiv. https://www.semanticscholar.org/paper/284e59cad4f3dd613bc81e00eb8e02eac8723530  
  
[111] Su-Jin Jeong & Jee-Yeon Lee. (2011). Usability of Research Information Systems. https://www.semanticscholar.org/paper/6360d867846c46d0b3a2ef8112d9c7bb3a6696ae  
  
[112] Sun Wen-can. (2004). Multi-Language Support for Object-Oriented Language Compiler Automatic Test Tool. In Journal of Beijing Institute of Technology. https://www.semanticscholar.org/paper/b1bca75815ef0d45442e500589b67ba3203b8550  
  
[113] T. Kushwaha & O. Sangwan. (2013). Prediction of usability level of test cases for GUI based application using fuzzy logic. https://www.semanticscholar.org/paper/4835bbde5eaaac38988cfd5063f6deefb670a94d  
  
[114] T. Mayfield, V. Gligor, J. Cugini, J. Boone, & Rob Dobry. (1995). Security Criteria for Distributed Systems: Functional Requirements. https://www.semanticscholar.org/paper/63b08d0466805f7c62b13ff621a68f2bdd3e5058  
  
[115] Tahereh Zohdinasab, Vincenzo Riccio, & P. Tonella. (2023). DeepAtash: Focused Test Generation for Deep Learning Systems. In Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis. https://www.semanticscholar.org/paper/610e42d01387d19625fadbf0e1ec4de4f5a3c5e9  
  
[116] Talissa Dreossi. (2025). Bridging Logic Programming and Deep Learning for Explainability through ILASP. In Electronic Proceedings in Theoretical Computer Science. https://www.semanticscholar.org/paper/f3c55f10ec07f62bff77e5596f0a4cfc1628216d  
  
[117] Tianyang Wang, Ziqian Bi, Yichao Zhang, Ming Liu, Weiche Hsieh, Pohsun Feng, Lawrence K.Q. Yan, Yizhu Wen, Benji Peng, Junyu Liu, Keyu Chen, Sen Zhang, Ming Li, Chuanqi Jiang, Xinyuan Song, Junjie Yang, Bowen Jing, Jintao Ren, Jun-Jie Song, … Qian Niu. (2024). Deep Learning Model Security: Threats and Defenses. In ArXiv. https://www.semanticscholar.org/paper/b9bb06319d1a003e5358a39efcbd2dd656206bc2  
  
[118] Ties Robroek, Aaron Duane, Ehsan Yousefzadeh-Asl-Miandoab, & Pinar Tozun. (2023). Data Management and Visualization for Benchmarking Deep Learning Training Systems. In Proceedings of the Seventh Workshop on Data Management for End-to-End Machine Learning. https://www.semanticscholar.org/paper/89431f1255b42e76e4df3fb399455f796478cdb4  
  
[119] Timo Speith. (2022). How to Evaluate Explainability? - A Case for Three Criteria. In 2022 IEEE 30th International Requirements Engineering Conference Workshops (REW). https://www.semanticscholar.org/paper/2e903d2ccaa15b591db53ee8e89c3d07b628155a  
  
[120] Tong Li & Lu Han. (2023). Dealing with Explainability Requirements for Machine Learning Systems. In 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC). https://www.semanticscholar.org/paper/8955b3947fba5de187f202897b9b8ca434837742  
  
[121] Tri Cao, Minh-Huy Trinh, Ailin Deng, Quoc-Nam Nguyen, Khoa Duong, Ngai-Man Cheung, & Bryan Hooi. (2024). Are Anomaly Scores Telling the Whole Story? A Benchmark for Multilevel Anomaly Detection. In ArXiv. https://www.semanticscholar.org/paper/e4748659ab7d98668b02a5c7d709341190877893  
  
[122] W. K. Cheung, Jeremy M. Kalindjian, Robert Bell, A. Nair, L. Menezes, R. Patel, S. Wan, Kacy Chou, Jiahang Chen, R. Torii, R. Davies, J. Moon, Daniel C. Alexander, & Joseph Jacob. (2023). A 3D deep learning classifier and its explainability when assessing coronary artery disease. In ArXiv. https://www.semanticscholar.org/paper/a4e36a2c40d88a1231f77c6d6934f393ee6bd6c4  
  
[123] W. Karaa & Dridi Kawther. (2023). Classification of Multilingual Medical Documents using Deep Learning. In 2023 IEEE/ACIS 21st International Conference on Software Engineering Research, Management and Applications (SERA). https://www.semanticscholar.org/paper/621841641a8739687ed3d693b984eb2b58854978  
  
[124] Waqas Ahmed, Vamsi Krishna Kommineni, B. König-Ries, Jitendra Gaikwad, Luiz Gadelha, & Sheeba Samuel. (2025). Evaluating the method reproducibility of deep learning models in biodiversity research. In PeerJ Computer Science. https://www.semanticscholar.org/paper/2a775a24f0ade3813c452a4001d7f090dbaaa3c7  
  
[125] Weiyu Fu & Lixia Wang. (2022). Software Security Testing through Coverage in Deep Neural Networks. In Security and Communication Networks. https://www.semanticscholar.org/paper/351af54a9fbfb170c54064502ef7bc7e9e06e2a8  
  
[126] Xian-gang Cao, Xin Xu, Yong Duan, & Xin Yang. (2022). Health Status Recognition of Rotating Machinery Based on Deep Residual Shrinkage Network Under Time-Varying Conditions. In IEEE Sensors Journal. https://www.semanticscholar.org/paper/d446eac3c50f57e598f3012393fbcf58b011ad70  
  
[127] Xubo Fei, Shiyong Lu, & Jia Zhang. (2011). A Granular Concurrency Control for Collaborative Scientific Workflow Composition. In 2011 IEEE International Conference on Services Computing. https://www.semanticscholar.org/paper/1ee867534e7bc8ea1e8879f085397c20e47fb6a4  
  
[128] Yang Cao, Sikun Yang, Chen Li, Haolong Xiang, Lianyong Qi, Bo Liu, Rongsheng Li, & Ming Liu. (2025). TAD-Bench: A Comprehensive Benchmark for Embedding-Based Text Anomaly Detection. In ArXiv. https://www.semanticscholar.org/paper/e7a44cbe780be6bd39df6ceb76cda852d3ef78c3  
  
[129] Yanglong Chen. (2024). Design of English Teaching Quality Evaluation System Based on Deep Learning and Data Mining. In 2024 IEEE 2nd International Conference on Image Processing and Computer Applications (ICIPCA). https://www.semanticscholar.org/paper/037c843cc7f03ea019a5e3f7b1cd9da8000862f9  
  
[130] Yanshan Chen, Ziyuan Wang, Dong Wang, Yongming Yao, & Zhenyu Chen. (2019). Behavior Pattern-Driven Test Case Selection for Deep Neural Networks. In 2019 IEEE International Conference On Artificial Intelligence Testing (AITest). https://www.semanticscholar.org/paper/2fb57d113989bd76eb25ab7f9abe2363e619dde6  
  
[131] Yasmine Djebrouni, Isabelly Rocha, S. Bouchenak, Lydia Chen, Pascal Felber, Vania Marangozova-Martin, & V. Schiavoni. (2023). Characterizing Distributed Machine Learning and Deep Learning Workloads. https://www.semanticscholar.org/paper/7ca9e1bc098d74c0d5c30eeacf68f47172ea98a7  
  
[132] Yingzhe He, Guozhu Meng, Kai Chen, Xingbo Hu, & Jinwen He. (2019). Towards Privacy and Security of Deep Learning Systems: A Survey. In ArXiv. https://www.semanticscholar.org/paper/fb647bbef5d0e8d202b305133d6a5c85bd9462b3  
  
[133] Yining Yin, Yang Feng, Zixi Liu, & Zhihong Zhao. (2023). Practical Accuracy Evaluation for Deep Learning Systems via Latent Representation Discrepancy. In Proceedings of the 14th Asia-Pacific Symposium on Internetware. https://www.semanticscholar.org/paper/f709861e262ee02a3eb068bb9fda8df5fa219cf2  
  
[134] Yoav Arad & Michael Werman. (2023). Beyond the Benchmark: Detecting Diverse Anomalies in Videos. In ArXiv. https://www.semanticscholar.org/paper/6237beef8df07d586c428724fa002359450c0d99  
  
[135] Yuangang Li, Jiaqi Li, Zhuo Xiao, Tiankai Yang, Yi Nian, Xiyang Hu, & Yue Zhao. (2024). NLP-ADBench: NLP Anomaly Detection Benchmark. In ArXiv. https://www.semanticscholar.org/paper/0e204b08a753b9d577173cf94512457a1741eca4  
  
[136] Zhaobin Li. (2025). Enhancing Human-AI Collaboration through Adaptive Interaction and Explainability. In AAAI/ACM Conference on AI, Ethics, and Society. https://www.semanticscholar.org/paper/a2b65e0c011cffabb77123589aad09f89cb4cb27  
  
[137] Zhaoxin Li. (2024). A Survey on Deep Domain Adaptation. In Highlights in Science, Engineering and Technology. https://www.semanticscholar.org/paper/8d36907720cf78056b6c1fb714f103be56a6e675  
  
[138] Zhenyu Liu, Tiejiang Liu, Lizhi Cai, & Genxing Yang. (2010). Test coverage for collaborative workflow application based on Petri Net. In The 2010 14th International Conference on Computer Supported Cooperative Work in Design. https://www.semanticscholar.org/paper/81874f0bae38c77472633c0f0773453b33f8d460  
  
[139] Ziwen Kan, Shahbaz Rezaei, & Xin Liu. (2024). Benchmarking Counterfactual Interpretability in Deep Learning Models for Time Series Classification. In ArXiv. https://www.semanticscholar.org/paper/af87c737044825a863ed5022e0e32a57bda7c00e