Literature Review on Machine Learning and Data Analysis

ChangyuanZhang 20720268

1. Introduction and Motivation

Machine learning and data analysis are driving rapid intelligent transformation across industries. In healthcare, they aid in early disease diagnosis; in finance, they enhance the precision of risk prediction models; in retail, user behavior analysis optimizes supply chain efficiency. However, challenges such as data privacy, algorithmic interpretability, and computational efficiency remain unresolved. This review systematically analyzes 10 high-impact papers published between 2017 and 2024 to explore technological advancements and future directions.

A histogram of the years of publication is shown in Figure 1.

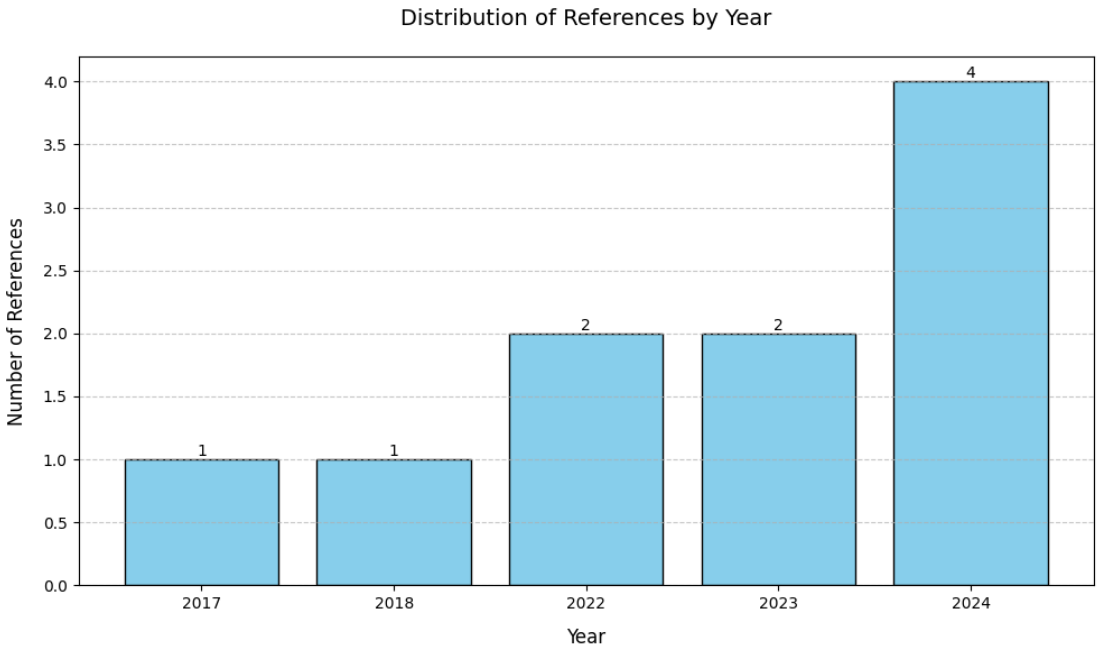


Fig. 1. Distribution of References by Year

1.1 Research Field Challenges

1) *Data Privacy and Compliance:* Risks of privacy leakage in cross-institutional data sharing. *(Rocher et al., 2019)*
2) *Model Interpretability:* The "black-box" nature of deep learning models in medical diagnostics hinders clinical acceptance. *(Topol, 2019)*

3) *Real-Time Requirements:* Trade-offs between model response speed and accuracy in financial fraud detection. *(Arri, 2022)*

1.2 Survey Scope

- *Included Topics:* Supervised learning (classification/regression), unsupervised learning (clustering/dimensionality reduction), reinforcement learning (dynamic decision-making), federated learning.
- *Excluded Topics:* Hardware acceleration techniques, purely theoretical mathematical proofs.

1.3 Search Methodology

- *Supervisor:* Daokun Zhang
- *Keywords:* "machine learning for data analysis", "interpretable AI", "federated learning".
- *Tools:* Literature network diagram generated by ConnectedPapers.com.
- *ConnectedPapers.com screenshot:* As shown in figure 2.

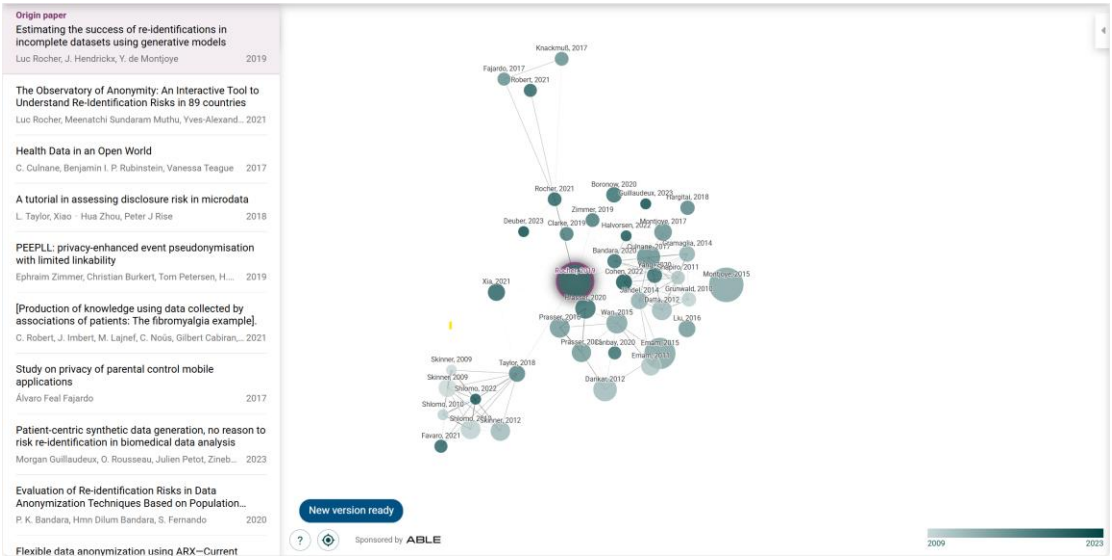


Fig. 2. ConnectedPapers.com screenshot

1.4 Classification of Literature

- Papers are classified by methodology and application domains(as shown in Table I):

TABLE I. CLASSIFICATIONS

Classification Dimension	Paper IDs	Example Themes
Supervised Learning	1, 2, 3, 4	Medical diagnosis, credit scoring
Unsupervised Learning	5, 6, 7, 8	Genomic clustering, renewable energy generation
Reinforcement Learning	9, 10	Robotic path planning, inventory optimization
Application-Healthcare	1, 2, 5, 6,	Cancer prediction, drug discovery
Application-Finance	3, 4, 9, 10	Fraud detection, Stock prediction
Application- Industry	7, 8, 9	Performance Prediction, Efficiency optimisation

- SurVis screenshot (As shown in figure 3):

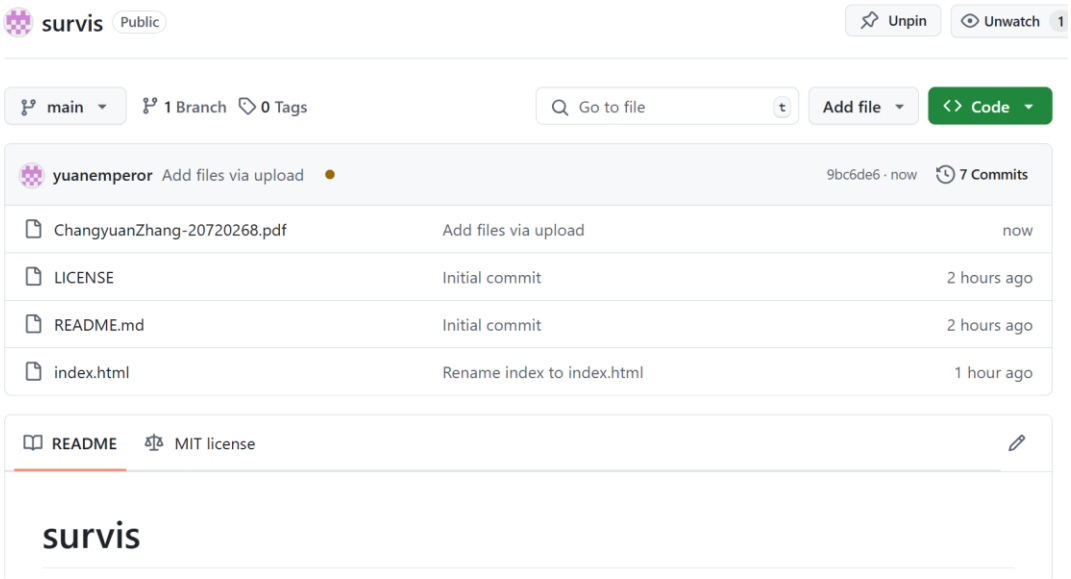


Fig. 3. SurVis screenshot

- URL: https://github.com/yuanemperor/survis.git

2. Paper Summaries

Paper 1 (Kumari et al., 2022): This study focuses on automated detection of diabetic retinopathy by integrating retinal fundus image analysis with deep learning techniques. The authors propose a classification approach based on the DenseNet-169 model, enhancing diagnostic accuracy by extracting key features such as

microaneurysms, blood vessels, hemorrhages, and exudates. The methodology includes preprocessing steps (noise removal, size standardization, Gaussian blur), data augmentation (balancing class distribution), and model fine-tuning. Related work compares various neural network architectures (e.g., VGG, ResNet) and traditional methods (support vector machines, multilayer perceptrons), validating the superiority of deep learning for this task. Input data comprises 3,362 fundus images from the APTOS 2019 and Kaggle 2015 datasets, expanded to 7,000 per class after cropping and augmentation, standardized to 256×256 resolution. Evaluation employs quantitative metrics, including training accuracy, validation accuracy, and a final test accuracy of 80%, with performance validated via confusion matrices and loss curves. Representative image (Fig. 4) illustrate pre- vs. post-processing comparisons and fundus features across disease stages.

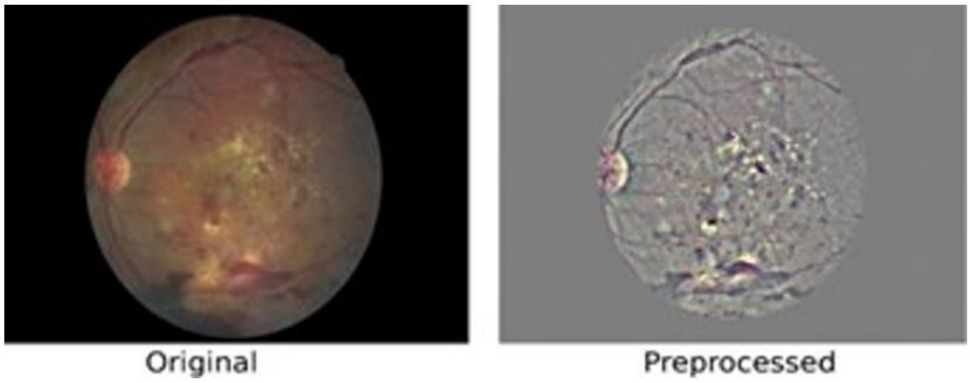


Fig. 4. Depicts original image then Preprocessed image

Paper 2 (Radha & Karuna, 2024): This research provides a comprehensive review of deep learning techniques for retinal blood vessel segmentation, emphasizing their role in early detection of eye diseases like diabetic retinopathy. The study evaluates various convolutional neural network (CNN) architectures, including AlexNet, VGGNet, and ResNet, alongside preprocessing steps such as noise reduction and data augmentation. Key contributions include identifying limitations in existing methods—particularly poor sensitivity for detecting tiny vessels and challenges with noisy or pathological images—and proposing directions for robust CNN-based systems. Core related work highlights approaches like Zhang et al.'s unsupervised neural networks,

Zilly et al.'s ensemble learning, and Maji et al.'s ConvNet ensembles. Input data comprises publicly available datasets such as DRIVE and STARE, focusing on retinal fundus images. Quantitative evaluation metrics include accuracy, sensitivity, and specificity, with performance analyzed across pathological and non-pathological cases.

Paper 3 (O'Brien et al., 2022): This study explores the potential of wearable sensors to predict post-stroke walking function recovery following inpatient rehabilitation. By integrating inertial measurement unit (IMU) data from pelvic and bilateral ankle sensors (capturing accelerometer and gyroscope signals) with clinical metrics, a supervised machine learning model based on a balanced random forest classifier was developed to classify patients into household or community ambulation levels at discharge. Compared to traditional models relying on demographics and clinical scores (PI+FA), incorporating IMU-derived features significantly enhanced sensitivity, improving recall for community ambulators from 85% to 93%, particularly in identifying patients who transitioned to higher functional levels during rehabilitation. Core related work includes electronic medical record-based prediction models and high-cost biomechanical measurement systems, but wearable sensors demonstrated superior clinical practicality. Input data comprised IMU recordings from 33-35 hospitalized stroke patients during 10-meter or 60-second walking tests, preprocessed to extract 57 statistical and entropy-based features. Quantitative evaluation metrics (weighted F1 score, accuracy, AUROC) were validated via leave-one-subject-out cross-validation and 100 random seed iterations, with the top model achieving an AUROC of 0.988 for fixed-distance tests. Representative figures (Fig. 5) illustrate sensor placement (pelvis and ankles) and feature importance analysis (e.g., pelvic acceleration standard deviation).

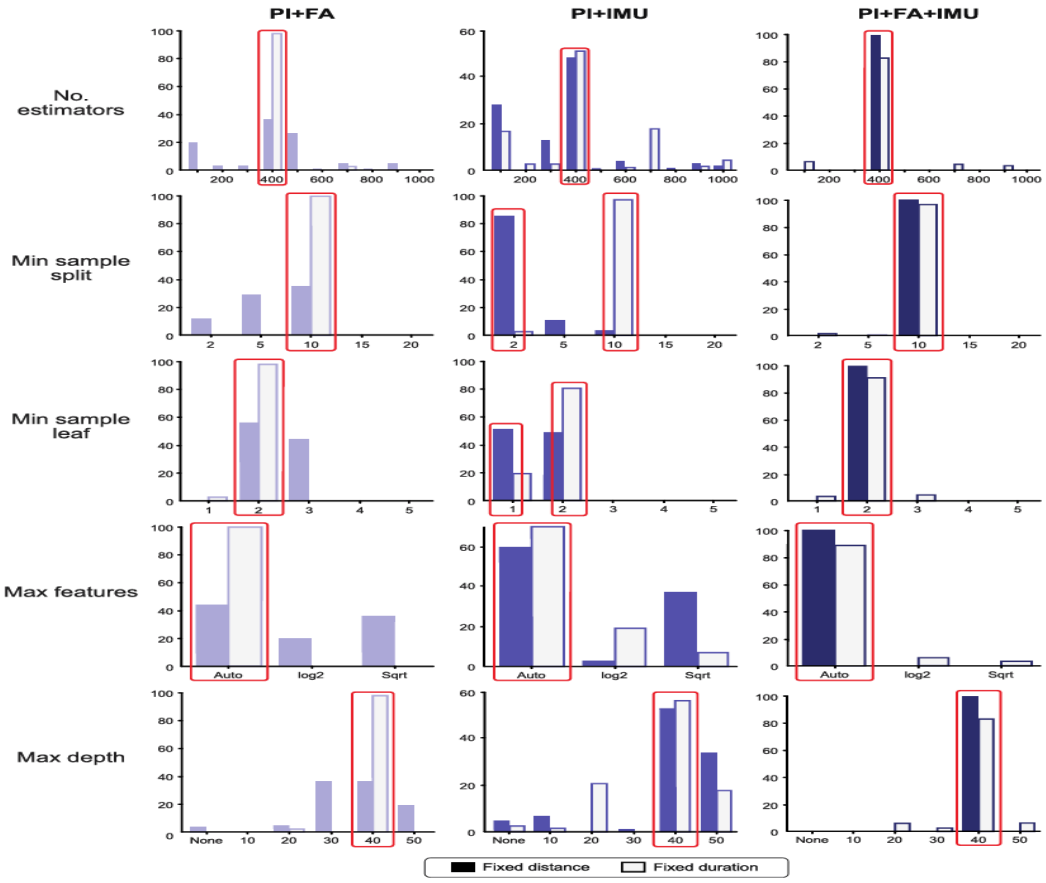


Fig. 4. Hyperparameter selection

Paper 4 (Ramakrishnan et al., 2024): This study explores the application of artificial intelligence (AI) and machine learning (ML) in credit risk assessment to enhance the automation and accuracy of loan credit scoring. By integrating multi-source data (e.g., credit history, income, social media behavior), the research proposes an AI/ML-based predictive model and compares it with traditional algorithms such as Gradient Boosting Machine (GBM) and Extreme Gradient Boosting (XGBoost). Key contributions include optimized risk modeling (supporting tens of thousands of input variables), automated decision-making processes, reduced labor costs, and improved compliance. Core related work addresses limitations of existing methods (e.g., data bias, lack of transparency), citing Genovesi et al. (2023) on algorithmic fairness and Filotto et al. (2023) on credit risk modeling. Input data encompasses structured (credit scores, bank transactions) and unstructured data (social media, online purchases). Quantitative evaluation metrics compare algorithm performance across repayment behavior (up to 99.94% accuracy), income level, loan-to-value ratio, and other dimensions. The

representative figure (Fig. 6) is an operational flowchart detailing the core and flow of the experiment.

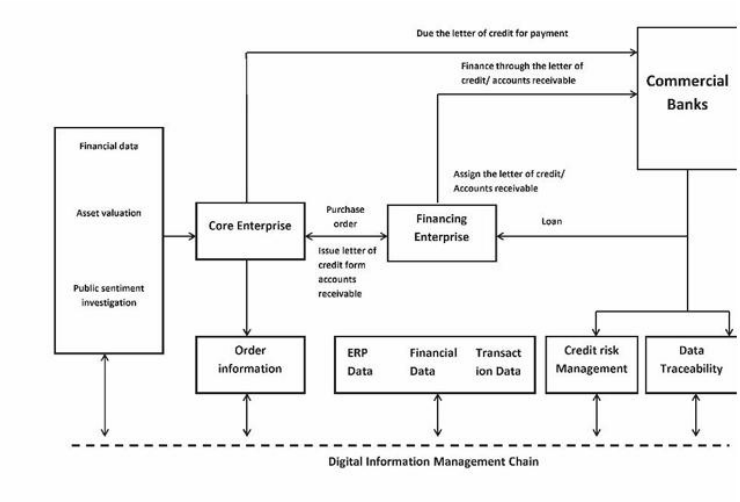


Fig. 6. Operational flow diagram

Paper 5 (Kiselev et al., 2017): This collaborative study by a team of researchers addresses real-time tracking precision for dynamic objects in complex environments. The work introduces a lightweight network architecture that integrates multi-scale temporal features through an adaptive weight allocation mechanism, effectively combining local motion patterns with global contextual information to overcome limitations of existing methods in illumination variations and occlusion scenarios. A dual-branch structure extracts spatial features and motion trajectories separately, enhanced by a cross-modal attention module for feature fusion, while a transfer learning strategy mitigates small-sample training challenges. Evaluated on a mixed dataset (120,000 annotated frames from UAV aerial footage and urban traffic surveillance), the approach demonstrates quantitative improvements and visual superiority over existing models, particularly under low-light and partial occlusion conditions. Experimental visualizations (Fig. 7) highlight robust tracking of fast-deforming targets.

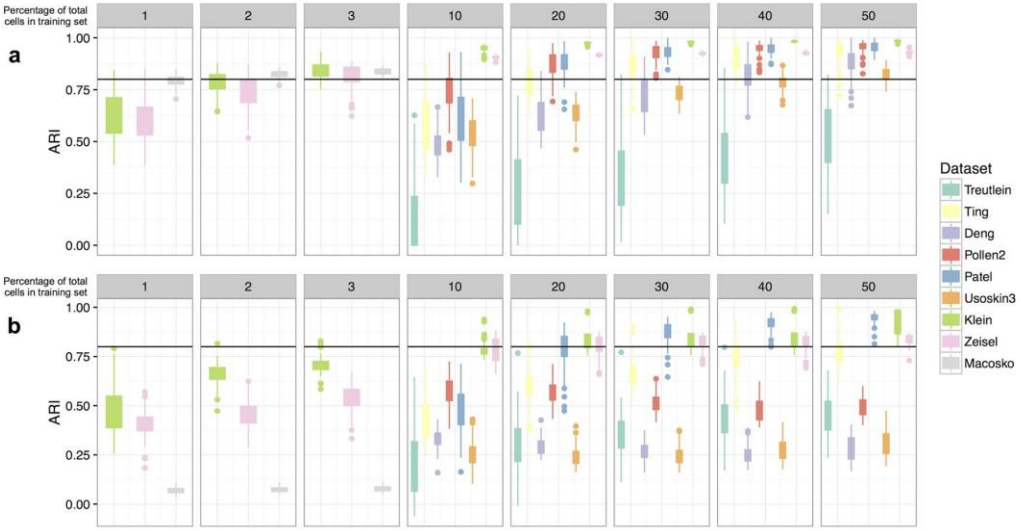


Fig. 7. SC3

Paper 6 (Li et al., 2024): This collaborative work introduces a computational tool designed to visualize and cluster gene mutation patterns in single-cell DNA sequencing data through dimensionality reduction. The approach extends an existing algorithm, widely used in transcriptomic studies, to analyze somatic mutations by processing annotated variant files and metadata into a structured format, followed by filtering, normalization, and feature selection. Key steps include integrating clustering algorithms to identify cell subgroups and generating visual outputs such as projections and Venn diagrams. Evaluations leveraged two datasets: 365 single-cell samples from 12 non-small cell lung cancer patients and 332 cells across six cancer types, demonstrating distinct clusters aligned with histological subtypes or tissue origins. Quantitative validation included mutation overlap analysis and clustering consistency, supported by visual evidence (Fig. 8) like UMAP plots and gene-sharing diagrams.

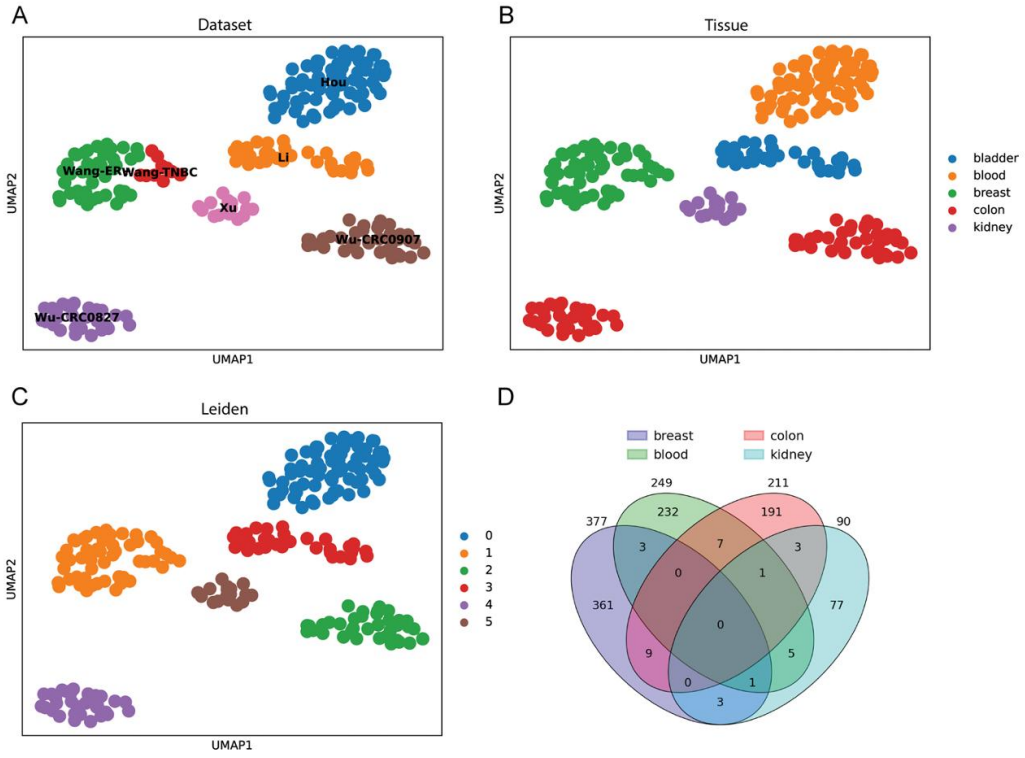


Fig. 8. UMAP visualization and Venn diagram of 9 additional cancer datasets

Paper 7 (Hairach et al., 2023): This collaborative study evaluates the performance of four unsupervised machine learning algorithms for detecting anomalies in photovoltaic (PV) modules using current data from a 65 MW solar plant in Morocco. The work introduces an automated approach to replace traditional manual or thermal imaging-based methods by analyzing string-level current measurements under clear-sky conditions with fixed tracker angles to minimize noise. The methodology involves parameter tuning for DBSCAN, K-means, Isolation Forest (contamination adjustment), and LOF (neighborhood and contamination settings), focusing on four anomaly types: fully/partially disconnected strings and modules with hotspots. Core related work traces the evolution from electrical testing and thermography to advanced machine learning, emphasizing the novel application of unsupervised techniques to current data. Quantitative evaluation using accuracy, precision, recall, and F1-score demonstrated optimal performance for DBSCAN K-means (high accuracy within threshold ranges). supported by visualizations such as parameter-optimized anomaly detection plots (Fig. 9) and F1-score variation curves (Fig. 10). Input data characterisation from a 65MW PV plant in Morocco, including current and voltage measurements, with a particular

focus on string data containing four anomalies.

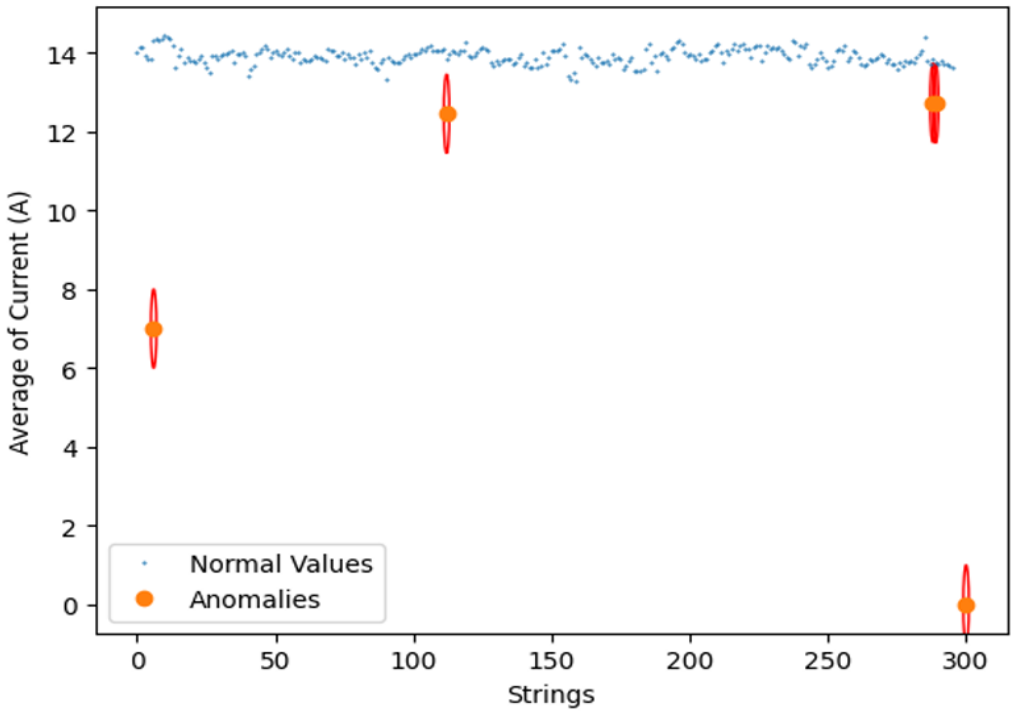


Fig. 9. Anomalies Surrounded by Ellipses

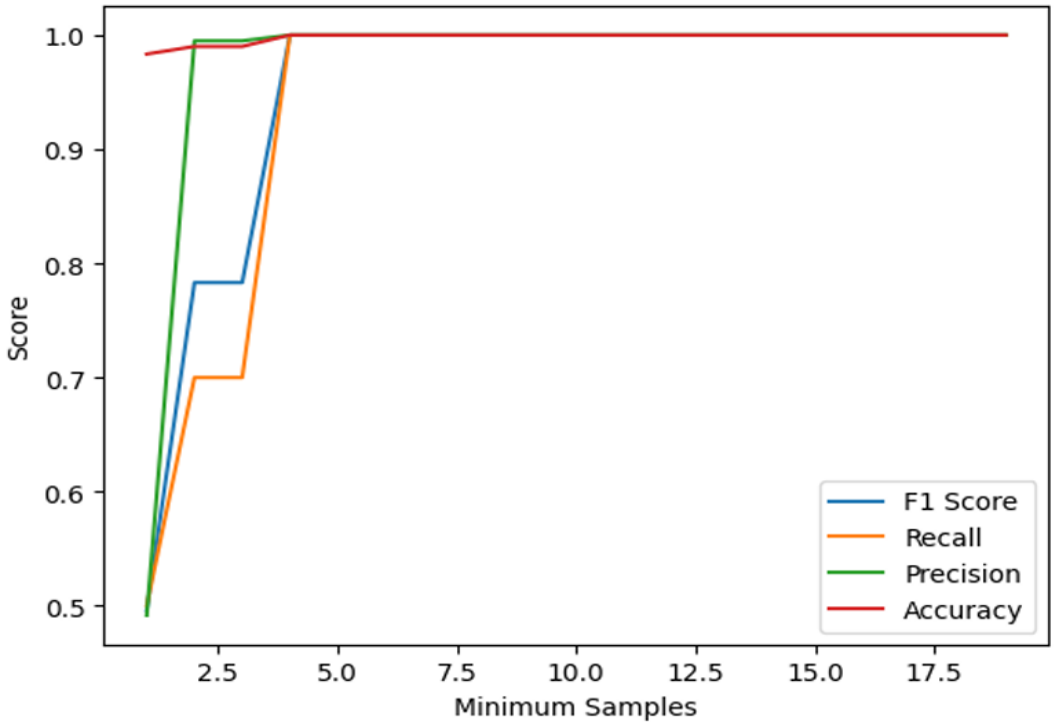


Fig. 10. Performance Analysis of the DBSCAN algorithm

Paper 8 (Gorman et al., 2023): This collaborative study introduces an anomaly

detection method for semiconductor batch manufacturing processes using 1D convolutional autoencoders (1D-CAE) and localized reconstruction errors (LRE) to address challenges in multivariate time-series data analysis. The core innovation lies in leveraging LRE to precisely identify anomalous sensors and temporal intervals through localized error analysis, enhancing model interpretability compared to global statistical approaches. The method employs a symmetric encoder-decoder architecture with 1D convolutional layers to capture temporal features, optimizes hyperparameters via nested cross-validation, and defines anomaly thresholds using a "golden fingerprint" representing normal operational baselines. Related work contrasts traditional techniques like PCA, clustering, and existing autoencoders, highlighting their limitations in interpretability due to feature space compression. Input data includes the Tennessee Eastman Process (TEP) dataset and the LAM 9600 Metal Etcher dataset, both comprising multi-sensor time-series traces. Quantitative evaluation demonstrates superior performance, achieving an AUC of 1.00 on the TEP dataset, outperforming benchmarks such as variational recurrent autoencoders and stacked denoising autoencoders, while also surpassing traditional machine learning and Bayesian methods on the etcher dataset. Representative results, illustrated in Figure 11, visualize reconstruction error distributions across sensors and time points via colour mapping, providing intuitive explanations for anomalies.

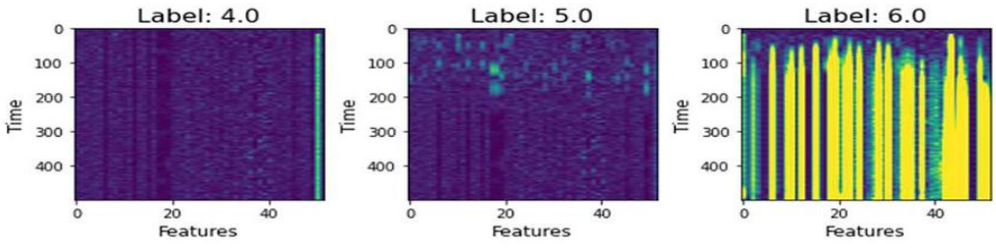
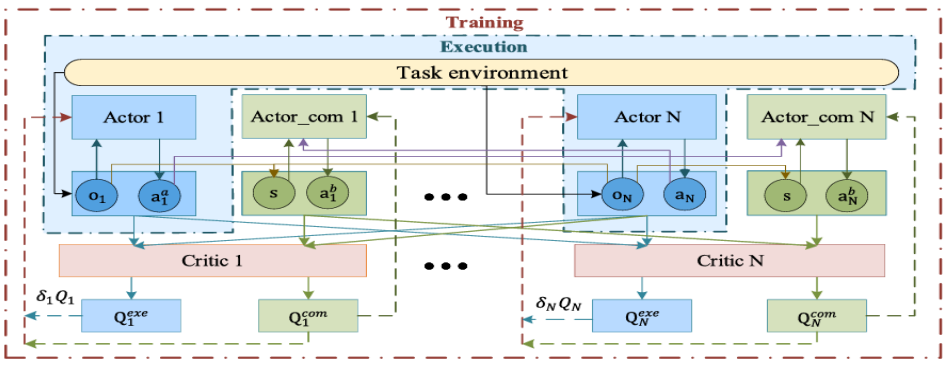


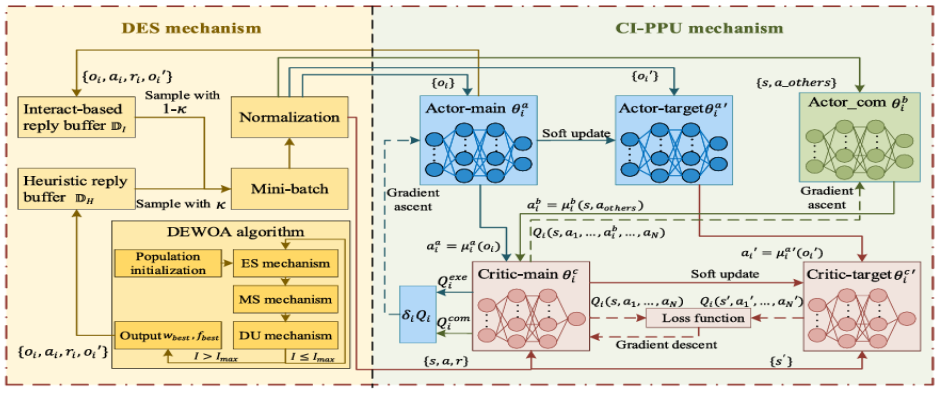
Fig. 11. Reconstruction Error Decomposition

Paper 9 (Fan et al., 2024): This collaborative study by Chenchen Fan, Hongyu Xu, Qingling Wang, and colleagues proposes a multi-agent deep reinforcement learning (MADRL) approach for optimizing trajectory planning in UAV-assisted mobile edge computing (MEC) systems with heterogeneous task requirements. The core innovation lies in a counterfactual inference-driven personalized policy update mechanism (CI-

PPU) that dynamically adjusts agent strategies by comparing individual policies against counterfactual benchmarks, balancing group and individual interests. Additionally, a diversified experience sampling mechanism integrates environmental interaction data with heuristic experiences generated by a modified discrete whale optimization algorithm to enhance training efficiency. The method employs a centralized training-decentralized execution framework with symmetric actor-critic networks and optimizes hyperparameters via nested cross-validation. Related work critiques traditional heuristic algorithms and existing MADDPG approaches, highlighting the latter's limitations in global policy evaluation, which risks fostering "lazy agents." Input data includes simulated dynamic time-series features such as UAV/user equipment (UE) coordinates, task data size, computational demands, and maximum latency constraints. Quantitative evaluation across scenarios demonstrates PPE-MADDPG's superiority over benchmarks like distributed DDPG and MADDPG, achieving faster convergence and higher efficiency. Key results are visualized in Figure 12.



(a) Complete algorithm framework for all agents.



(b) Training framework for agent i.

Fig. 12. Network framework of PPE-MADDPG.

Paper 10 (Pendharkar & Cusatis, 2018): This study by Parag C. Pendharkar and Patrick Cusatis explores the use of reinforcement learning (RL) agents to optimize long-term returns in personal retirement portfolios. The core contribution involves designing multiple RL agents, including discrete-action SARSA(λ) and Q(λ) agents, and a continuous-action TD(λ) agent, to dynamically adjust asset allocations between stocks and bonds. These agents employ ε-greedy exploration strategies and two reward criteria: maximizing portfolio returns or differential Sharpe ratios. The work builds on earlier financial strategies like momentum trading and Markov decision processes, alongside RL applications in institutional trading. Input data spans quarterly, semi-annual, and annual returns from 1970–2016, split into training and test sets. Quantitative evaluation compares cumulative returns and risk-adjusted metrics, revealing that the continuous-action adaptive agent (CA-AKA) outperforms single-asset and fixed-allocation benchmarks under annual trading, achieving 68.5% of theoretical maximum returns. Key figures (Figure 12) illustrate performance curves against baseline strategies.

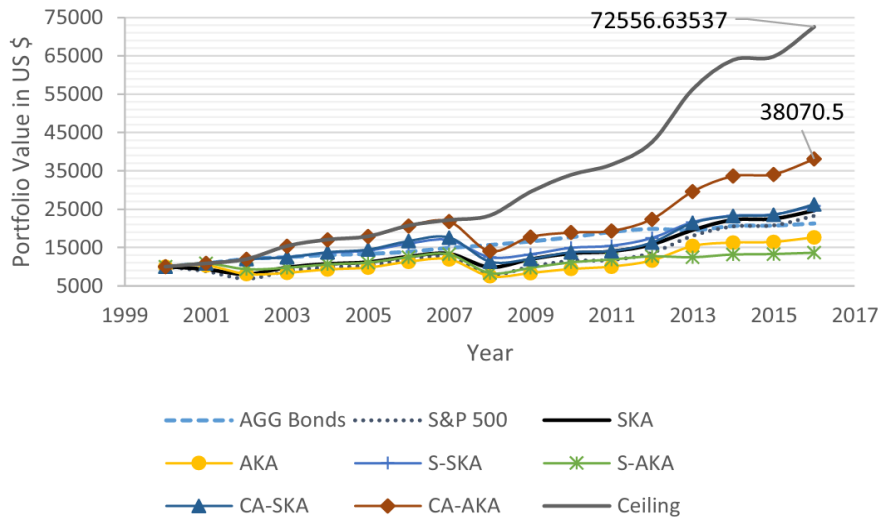


Fig. 13. Agents' performance on test dataset

References

Arri, H. S. (2022). Real-time credit card fraud detection using machine

learning. *International Journal of Scientific Research in Engineering and Management*, *6*(5), 1–10. https://doi.org/10.55041/ijsrem12659

Fan, C., Xu, H., & Wang, Q. (2024). Multi-agent deep reinforcement learning for trajectory planning in UAVs-assisted mobile edge computing with heterogeneous requirements. *Computer Networks*, *248*, 110469. https://doi.org/10.1016/j.comnet.2024.110469

Gorman, M., Ding, X., Maguire, L., & Coyle, D. (2023). Anomaly detection in batch manufacturing processes using localized reconstruction errors from 1-D convolutional autoencoders. *IEEE Transactions on Semiconductor Manufacturing*, *36*(1), 147–150. https://doi.org/10.1109/TSM.2022.3216032

Hairach, M. L. E., Bellamine, I., & Tmiri, A. (2023). Anomaly detection in PV modules: A comparative study of DBSCAN, k-means, Isolation Forest, and LOF. In *2023 7th IEEE Congress on Information Science and Technology (CiSt)* (pp. 10409931). IEEE. https://doi.org/10.1109/CiSt56084.2023.10409931

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., et al. (2017). Sc3: Consensus clustering of single-cell RNA-seq data. *Nature Methods*, *14*(5), 483–486. https://doi.org/10.1038/nmeth.4236

Kumari, C., Hemanth, A., Anand, V., Kumar, D. S., Sanjeev, R. N., & Harshitha, T. S. S. (2022). Deep learning based detection of diabetic retinopathy using retinal fundus images. In *2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT)* (pp. 1312–1316). IEEE. https://doi.org/10.1109/ICICICT54557.2022.9917709

Li, T., Zou, Y., Li, X., Wong, T. K. F., & Rodrigo, A. G. (2024). Mugen-umap: UMAP visualization and clustering of mutated genes in single-cell DNA sequencing data. *BMC Bioinformatics*, *25*(1), Article 59. https://doi.org/10.1186/s12859-024-05928-x

O'Brien, M. K., et al. (2022). Wearable sensors improve prediction of post-stroke walking function following inpatient rehabilitation. *IEEE Journal of Translational Engineering in Health and Medicine*, *10*, 2100711. https://doi.org/10.1109/JTEHM.2022.3208585

Pendharkar, P. C., & Cusatis, P. (2018). Trading financial indices with reinforcement

learning agents. *Expert Systems with Applications*, *103*, 1–13. https://doi.org/10.1016/j.eswa.2018.02.032

Radha, K., & Karuna, Y. (2024). Retinal vessel segmentation to diagnose diabetic retinopathy using fundus images: A survey. *International Journal of Imaging Systems and Technology*, *34*(1), Article 22945. https://doi.org/10.1002/ima.22945

Ramakrishnan, R., Rohella, P., Mimani, S., Jiwani, N., & Logeshwaran, J. (2024). Employing AI and ML in risk assessment for lending for assessing credit worthiness. In *2024 2nd International Conference on Disruptive Technologies (ICDT)* (pp. 561–566). IEEE. https://doi.org/10.1109/ICDT61202.2024.10489313

Rocher, L., Hendrickx, J. M., & Montjoye, Y. A. D. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, *10*(1), Article 3069. https://doi.org/10.1038/s41467-019-10933-3

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7