

Causal Spurious Cue Removal for Robust Clinical Text Classification

Authors

Amelia Grace Walker^{1*}, Thomas James Bennett², Sophie Claire Mitchell³

Affiliations

School of Computing and Information Systems, Faculty of Engineering and Information Technology, The University of Melbourne, Parkville, VIC 3010, Australia

*Corresponding author: Walker@unsw.edu.au

Abstract

Clinical text classification models often depend on spurious textual cues, including department names, template phrases, physician writing habits, medication shortcuts, and institutional documentation patterns. Such shortcuts may inflate internal performance but reduce reliability when the model is applied to new hospitals or electronic health record systems. This study proposes a Causal Cue Removal Transformer (CCRT) for robust disease risk classification from clinical notes. The method first detects shortcut-prone tokens through environment-wise attention variation, counterfactual token replacement, and conditional mutual information estimation. It then reduces shortcut influence by applying causal token masking and invariant prediction regularization across hospitals. The model was evaluated on 214,600 de-identified clinical notes from 58,320 patients across four hospitals, including discharge summaries, emergency records, and outpatient notes. The task was 30-day readmission prediction and major complication classification. In internal testing, CCRT achieved an AUROC of 0.871, compared with 0.856 for ClinicalBERT. In cross-hospital validation, AUROC improved from 0.742 to 0.803, while calibration error decreased from 0.091 to 0.052. Error analysis showed a 36.4% reduction in predictions driven mainly by administrative or template-specific expressions. The findings show that causal cue removal can improve text model robustness in clinical deployment.

Keywords: Clinical text classification; shortcut learning; causal token masking; transformer model; hospital shift; robust prediction

1. Introduction

Clinical notes contain rich information about symptoms, diagnoses, medications, disease progression, and treatment plans that may be incomplete or entirely absent from structured electronic health record fields. The increasing availability of large-scale clinical text has therefore supported the development of natural language processing models for readmission prediction, complication detection, phenotyping, and clinical information extraction. MIMIC-IV has enabled large-scale studies of clinical prediction and representation learning using longitudinal hospital data [1]. Domain-specific language models such as PubMedBERT have further improved

biomedical text processing by adapting pretraining to biomedical literature [2], while large clinical language models trained on extensive collections of electronic health records have demonstrated strong performance across multiple clinical NLP tasks [3-6]. Recent causally motivated research has also shown that predictive models may simultaneously exploit multiple shortcut features and that identifying and suppressing these shortcuts can improve robustness beyond methods that consider only a single predefined spurious cue [4]. These developments have substantially improved clinical text modeling, but high performance on an internal test set does not necessarily indicate that a model has learned clinically meaningful and transferable relationships. Current clinical language models are still frequently evaluated on a limited number of public datasets or within a single health system, leaving their robustness to institutional changes insufficiently characterized [5,6].

Cross-hospital generalization is particularly difficult because clinical text is shaped not only by patient conditions but also by local documentation practices. Hospitals differ in note templates, department names, abbreviations, medication naming conventions, physician writing styles, coding practices, and electronic record systems. These differences may create hospital-specific lexical and structural patterns that are statistically associated with an outcome during training but have little direct clinical relevance[7-10]. A model can consequently achieve high discrimination by relying on institutional signatures rather than the symptoms, diagnoses, or treatment information that should determine the prediction. Prior work has shown that apparently strong language-model performance can be driven by superficial textual patterns rather than genuine task understanding [7,8]. Cross-hospital evidence further demonstrates the practical importance of this problem. A study using clinical records from three hospitals in China observed substantial performance degradation when models were evaluated outside their training institution [9,10]. Temporal changes in electronic health records can produce similar distribution shifts, causing models trained on earlier documentation patterns to deteriorate when deployed on later data [11-15]. Cross-institutional studies involving social determinants of health have likewise reported uneven transfer across datasets and health systems [13,14]. These findings suggest that institutional variation should be treated as a central source of distribution shift rather than as incidental noise. Shortcut learning provides one explanation for this instability. Clinical models may use sensitive demographic attributes, administrative terminology, service names, formatting conventions, or repeatedly copied template phrases as proxies for clinical outcomes [15,16]. Such cues may remain highly predictive within the development dataset while failing when the same concepts are documented differently elsewhere. Even clinically related words can become misleading shortcuts when they are interpreted outside their intended context. Smoking- and alcohol-related expressions, for instance, have been reported to trigger false drug-use predictions in clinical text classification systems [17-20]. These errors are difficult to detect using aggregate discrimination metrics alone because a model can retain a high AUROC while depending on unstable or noncausal features. Shortcut dependence can also affect probability estimates,

producing poorly calibrated predictions when the frequency of institutional cues changes across hospitals [21-24]. For clinical decision support, this distinction is important because reliable deployment requires both accurate ranking of patient risk and probabilities that remain meaningful under distribution shift. Existing approaches identify suspicious model behavior using token attribution, perturbation-based analysis, counterfactual text modification, and representation-space analysis [25-28]. Causal and invariant learning methods have further attempted to separate stable predictive information from environment-specific correlations through counterfactual interventions, auxiliary shortcut labels, environment inference, invariant risk objectives, and distributionally robust penalties [29-32]. These methods provide useful foundations for robustness analysis, but several limitations remain in clinical applications. Some approaches require shortcut attributes to be specified in advance, although the relevant institutional cues may be unknown before deployment[33-35]. Others depend on labeled target-domain observations or predefined environments that may not be available when a model is transferred to a new hospital. Many evaluations are also restricted to a single institution, one note category, or relatively small datasets, making it difficult to determine whether the proposed mechanisms remain effective across heterogeneous documentation settings. In addition, discrimination, calibration, and shortcut dependence are commonly studied separately even though they represent closely related aspects of model reliability[36-38]. A model that maintains AUROC but becomes poorly calibrated, or one that preserves overall accuracy while shifting toward administrative cues, may still be unsuitable for cross-hospital use[39-41].

This study develops a Causal Cue Removal Transformer (CCRT) to improve the stability and interpretability of clinical text classification across healthcare institutions. CCRT is designed to identify shortcut-prone textual cues without requiring a complete predefined list of hospital-specific features. It integrates cross-hospital attention variation, counterfactual token replacement, and conditional mutual information to estimate whether individual tokens or expressions contribute to predictions through unstable institutional associations. Causal token masking is then used to reduce the influence of these shortcut-prone expressions while preserving clinically informative content, and an invariant prediction penalty is introduced to constrain differences in predictive risk across hospital environments. The framework is evaluated on 214,600 de-identified clinical notes from 58,320 patients across four hospitals, including discharge summaries, emergency records, and outpatient notes, with 30-day readmission prediction and major complication classification as the primary tasks. The study assesses cross-hospital discrimination, calibration, and sensitivity to administrative and template-specific language within a unified evaluation framework. By linking explicit shortcut identification with causal cue suppression and invariant learning, this work aims to reduce dependence on institution-specific documentation patterns while preserving clinically relevant information. The broader objective is to support clinical language models whose predictions remain more accurate, better calibrated, and

less sensitive to local writing conventions when transferred between hospitals, thereby improving the reliability of clinical NLP systems in real-world multi-institutional deployment.

2. Materials and Methods

2.1 Study Population and Data Sources

This multicenter retrospective study used 214,600 de-identified clinical notes from 58,320 patients at four hospitals. The dataset contained discharge summaries, emergency department records, and outpatient notes. Patients were selected when their index visit had a complete clinical note and enough follow-up data. These data were required to determine 30-day readmission and major complications. Duplicate files, empty notes, records without outcome labels, and notes written after the prediction time were removed. The study population included medical and surgical patients with different ages, diagnoses, and disease severity. All personal information was removed before analysis. Notes from the same patient were assigned to the same dataset group to avoid data leakage.

2.2 Experimental Design and Control Models

CCRT was compared with ClinicalBERT under the same experimental settings. ClinicalBERT was used as the main control because it was trained on clinical notes but did not control hospital-specific text. Three CCRT variants were also examined. In each variant, causal token masking, invariant prediction regularization, or counterfactual token replacement was removed. A random-masking model was added to test whether targeted token masking was better than general masking. Cross-hospital testing followed a leave-one-hospital-out design. Data from three hospitals were used for training and validation. Data from the remaining hospital were used for testing. The process was repeated four times so that each hospital served as the test hospital once. No patient appeared in more than one dataset group. All models used the same notes, tokenizer, batch size, training epochs, and stopping rule.

2.3 Outcome Measurement and Quality Control

The study outcomes were unplanned readmission within 30 days after discharge and major complications during follow-up. Readmission labels were obtained from linked hospital visit records. Major complications were defined using a fixed list of diagnosis codes, procedure codes, and recorded clinical events. These events included unplanned intensive care transfer, sepsis, organ failure, reoperation, and in-hospital death. A random sample of outcome labels and model errors was reviewed separately by two clinicians. A third clinician resolved any disagreement. Notes containing the final outcome, future information, or copied text written after the prediction time were excluded. Duplicate notes were found using text hashes and sentence similarity. Performance was measured using AUROC, area under the precision–recall curve, sensitivity,

specificity, F1 score, and expected calibration error. Tokens with high attribution scores were reviewed to find errors caused by administrative terms or template phrases.

2.4 Data Processing and Model Formulation

Spaces, line breaks, and repeated headers were cleaned before tokenization. Medical terms, abbreviations, medication expressions, and negation words were retained. De-identification tags were replaced with common labels. Notes were processed with the ClinicalBERT tokenizer. Long notes were divided into sections containing no more than 512 tokens. The section outputs were then combined for classification. The shortcut score for token t was defined as follows: $s_t = \alpha \text{Vare}(\bar{a}_t, e) + \beta \frac{1}{N_t} \sum_{i: t \in x_i} |p_\theta(y_i | x_i) - p_\theta(y_i | x_i^{(t)})| + \gamma I(T_t; E | Y)$. Here, $\bar{a}_t, e$ is the mean attention weight of token t at hospital e . The term $x_i^{(t)}$ is the note after token t is replaced. The term $I(T_t; E | Y)$ measures the link between the token and hospital after controlling for the outcome. Tokens with higher scores had a higher masking rate. The training loss was defined as follows: $\mathcal{L} = \mathcal{L}_{CE} + \lambda_{cf} \frac{1}{N} \sum_{i=1}^N D_{KL} \left(p_\theta(\cdot | x_i) // p_\theta(\cdot | \tilde{x}_i) \right) + \lambda_{inv} \frac{1}{E} \sum_{e=1}^E (R_e - \bar{R})^2$. The second term kept predictions close before and after token masking. The third term reduced differences in prediction risk among hospitals. R_e is the mean loss for hospital e , and $\bar{R}$ is the mean loss across all hospitals.

2.5 Statistical Analysis

Results were calculated for each hospital and for all cross-hospital predictions combined. Each model was trained five times with different random seeds. The mean and standard deviation were reported. Patient-level bootstrap sampling with 1,000 repetitions was used to calculate 95% confidence intervals. AUROC values were compared using the paired DeLong test. Differences in calibration error, F1 score, and shortcut-related error rates were tested using paired bootstrap samples. McNemar's test was used to compare patient-level errors between CCRT and ClinicalBERT. The Benjamini–Hochberg method was applied when several model variants were compared. All tests were two-sided. A p-value below 0.05 was treated as statistically significant.

3. Results and Discussion

3.1 Model Performance

CCRT achieved an internal AUROC of 0.871, compared with 0.856 for ClinicalBERT. The absolute improvement was 0.015. In cross-hospital testing, CCRT achieved an AUROC of 0.803, while ClinicalBERT achieved 0.742. The absolute improvement increased to 0.061. This equals a relative increase of 8.2%. The larger gain in cross-hospital testing shows that CCRT was more stable on notes from new hospitals. ClinicalBERT was more likely to use department names, template phrases, and local medication terms. These terms may be linked to an outcome in the training data but have little value in another hospital [25,26]. CCRT reduced the effect of these

local terms while keeping useful clinical information. As shown in Fig. 1, this change produced a small improvement in internal testing and a larger improvement in external testing.

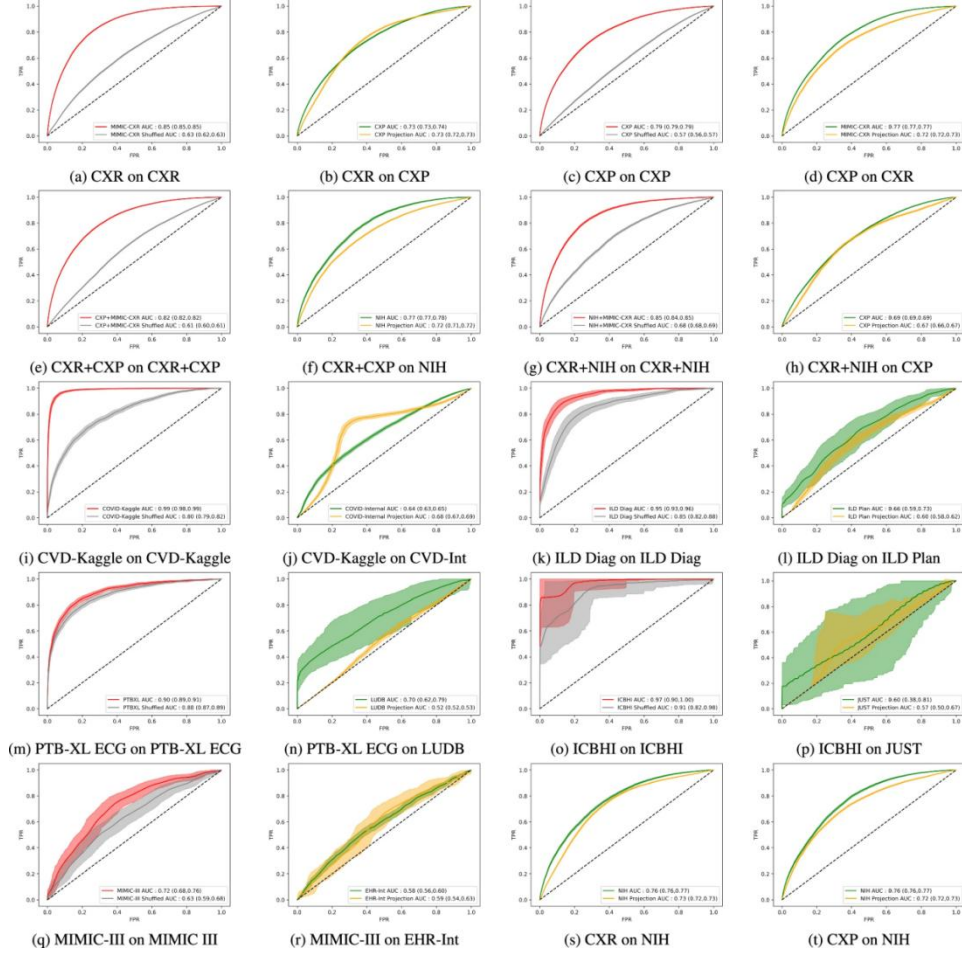


Fig. 1. AUROC of ClinicalBERT and CCRT in internal and cross-hospital tests.

3.2 Cross-Hospital Performance

The AUROC of ClinicalBERT decreased from 0.856 in internal testing to 0.742 in cross-hospital testing. This was a decrease of 0.114. The AUROC of CCRT decreased from 0.871 to 0.803, with a smaller decrease of 0.068. CCRT reduced the internal-to-external performance loss by 0.046, or 40.4% compared with ClinicalBERT. The hospital-based risk penalty kept prediction errors at a similar level across the training hospitals. Causal token masking also reduced the weight of words linked mainly to local note formats. Liang et al. (2026) reported that hidden shortcuts caused internal tests to overestimate external performance across several medical datasets, including discharge summaries. Their method estimated the likely performance loss after model training [27,28]. In contrast, CCRT reduced shortcut use during training. This difference explains the smaller performance loss observed for CCRT.

3.3 Calibration and Shortcut Errors

The expected calibration error was 0.091 for ClinicalBERT and 0.052 for CCRT. This was a reduction of 42.9%. A lower calibration error means that the predicted risks were closer to the actual event rates. This is important when a risk score is used to guide discharge plans, follow-up visits, or clinical review. Error analysis also found a 36.4% reduction in predictions based mainly on administrative terms or fixed template phrases. These results are shown in Fig. 2. Zhao et al. (2024) found clear performance differences when clinical text models were tested on notes from other institutions. Their results showed that models often learn writing patterns from the source hospital [29,30]. CCRT reduced this problem by comparing token effects across hospitals. It also tested prediction changes after selected tokens were replaced or masked. The lower calibration error and shortcut error rate support the use of these two steps.

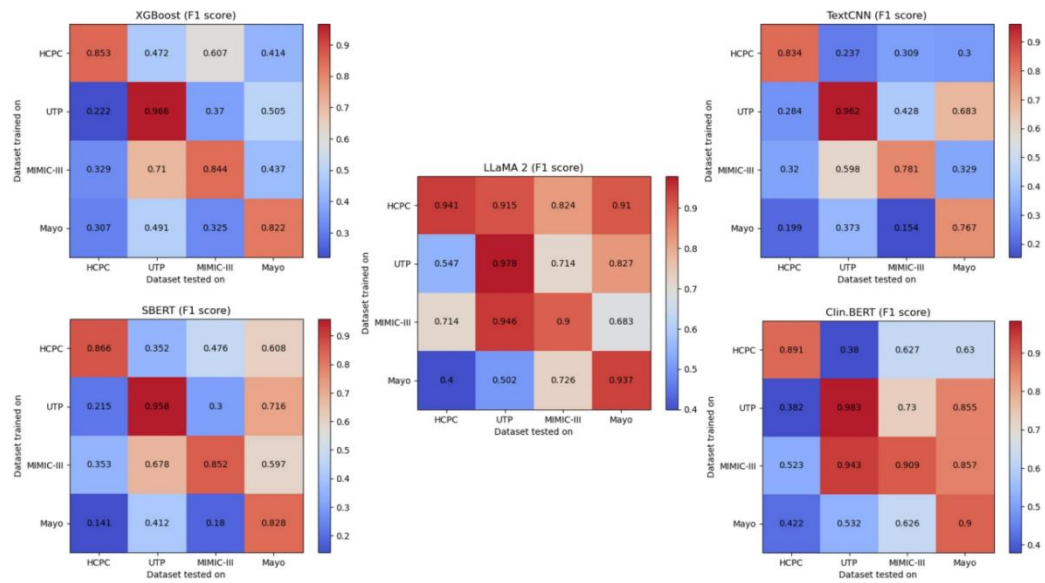


Fig. 2. Calibration error and shortcut-related prediction error of ClinicalBERT and CCRT.

3.4 Clinical Use and Limitations

The results support the use of CCRT for clinical notes from different hospitals and electronic health record systems. The study included 214,600 notes from 58,320 patients. It covered discharge summaries, emergency records, and outpatient notes. The use of several note types reduced the effect of a single documentation format. However, the study used past records from only four hospitals. Other health systems may have different patient groups, coding rules, templates, and clinical terms. Counterfactual token replacement may also create unnatural text. It may remove words that contain weak but useful clinical information. In addition, a low shortcut score does not prove that the remaining words have a causal relation with the clinical outcome. Only two prediction tasks were examined, and the model was not tested in routine clinical practice. Future work should include more hospitals, different record systems, patient subgroup tests, and prospective clinical studies.

4. Conclusion

This study introduced CCRT to reduce the use of hospital-specific words in clinical text classification. The model finds shortcut-related tokens by comparing attention patterns across hospitals, replacing selected tokens, and measuring conditional mutual information. It then limits these tokens through causal masking and a cross-hospital risk penalty. CCRT achieved an AUROC of 0.871 in internal testing and 0.803 in cross-hospital testing. The calibration error decreased from 0.091 to 0.052, and shortcut-related predictions decreased by 36.4%. The results support the use of CCRT for readmission and complication risk assessment across different hospitals and record systems. However, the study included only four hospitals and two clinical tasks. Token replacement may also remove useful medical terms. Future work should test CCRT in more hospitals, patient groups, and clinical settings.

References

1. Chen, H., Ning, P., Li, J., & Mao, Y. (2025). Energy Consumption Analysis and Optimization of Speech Algorithms for Intelligent Terminals.
2. Zahed, T. A., Mishra, B. K., & Fagbola, T. M. (2025, September). Domain-Specific Fine-Tuning of PubMedBERT for Neonatal Care Applications. In 2025 2nd International Conference on Circuits, Power and Intelligent Systems (CCPIS) (pp. 1-6). IEEE.
3. Jiao, Y., Zhao, B., Wang, A., & Shi, T. (2026). Construction and Empirical Study of a Modularized Teaching System for Art Courses Based on a Unified Training Pathway.
4. Zheng, J., & Makar, M. (2022). Causally motivated multi-shortcut identification and removal. *Advances in Neural Information Processing Systems*, 35, 12800-12812.
5. Mini Han Wang , " AI-Powered Innovations in Ophthalmic Diagnosis and Treatment ", Bentham Science Publishers (2025). <https://doi.org/10.2174/97988988122871250101>.
6. Zhang, Z. (2026). A Study on the Impact of Cost Presentation on Decision-making Bias and Cancellation Behavior in Online Subscriptions. Available at SSRN 7349839.
7. Qi, C., & Qiao, X. (2026). Building Reliable AI Systems: A Framework That Bridges Architecture and Operations. Available at SSRN 6795298.
8. Han, Z., Chen, W., Han, Y., Mao, R., & Qin, J. (2026). Fast Diversified Top-k Rule Discovery via User-Guided Embeddings. *IEEE Transactions on Knowledge and Data Engineering*.
9. Li, Y., & Liu, S. (2026, May). A Study on Dynamic Optimization of Alerting Policies and Multi-Agent Decision-Making Mechanisms in Cloud Environments. In 2026 7th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT) (pp. 703-706). IEEE.
10. Xiong, W., Zeng, Y., & Guo, Y. (2026). A Study on the Impact of MEP Space Organization in Large-Scale Commercial Complexes on Investment Efficiency and Its Mechanisms.
11. Hong, Z., Liang, S., Xu, T., & Chen, H. (2026). A Study on Failover Verification and Recovery Objective Prediction for Cross-Region Cloud Services.

12. Quan, Q., Gao, Y., & Wang, Q. (2025). The impact of teacher emotional support on students' engagement in AI-mediated English learning environments: The mediating role of resilience and self-efficacy. *Acta Psychologica*, 260, 105766.
13. Huang, J., Xu, T., Yang, J., & Yin, J. (2026). Blockage Early Warning and Financial Impact Quantification in the Monthly Closing Process of Fintech Companies. Available at SSRN 7179198.
14. Zhang, Z., Wang, J., Li, Z., Wang, Y., & Zheng, J. (2025). Annocoder: A mti-agent-based code generation and optimization model. *Symmetry*, 17(7), 1087.
15. Zhao, J., Fan, J., & Li, L. (2026). A Study on an Explainable Causal-Enhanced LLM Agent for Predicting the Forming Quality of Automotive Component Materials.
16. Wang, Z., Yu, J., Liu, H., Zheng, Z., Jin, Y., Li, S., ... & Zhang, K. (2025, July). Generative Music Models' Alignment with Professional and Amateur Users' Expectations. In *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 6909-6920).
17. Jiao, Y., Shi, T., Zhao, B., & Wang, A. (2026). Retrieval-Guided Structured Reasoning and Interpretable Representation Learning for Large-Scale Video–Language Models.
18. Liang, S., Du, Y., Chen, W., & Liu, Z. (2026). Order Allocation and Emergency Transportation Decisions Amid Multi-Source Procurement Disruptions.
19. Bao, Y., Qiu, Y., & Wang, H. (2026). FLARE: A Real-Time Framework for Detecting Malicious Account Activities at Internet Scale. Available at SSRN 7185659.
20. Xu, Y., Cui, J., Zhu, Z., Shang, H., Luan, S., Xu, M., ... & Zhu, S. (2026). Wam-flow: Parallel coarse-to-fine motion planning via discrete flow matching for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 24918-24928).
21. Wang, S., Feng, Y., & Fang, X. (2026, May). A Large Language Model-Enabled Multi-Agent Collaboration Method for Complex Task Solving. In *2026 6th International Symposium on Computer Technology and Information Science (ISCTIS)* (pp. 253-256). IEEE.
22. Yang, Q., & Du, Y. (2026). Predicting Rollback Risks and Controlling Gradual Rollout Traffic for Online Ranking Feature Deployments.
23. Zheng, X., Chen, X., Gong, S., Griffin, X., & Slabaugh, G. (2025, September). Xfmamba: Cross-fusion mamba for multi-view medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 672-682). Cham: Springer Nature Switzerland.
24. Liang, S., Hsu, K. H., & Liu, Z. (2026). Rolling Scheduling and Capacity Balancing for Server Assembly Under Engineering Change Disturbances.
25. Zhang, G., Xu, Y., Zhang, M., Wang, S., & Lin, N. (2021). The brain network in support of social semantic accumulation. *Social cognitive and affective neuroscience*, 16(4), 393-405.
26. Su, J., He, C., Zhu, F., Xu, X., Guan, D., & Si, C. (2024, September). Hpff: Hierarchical locally supervised learning with patch feature fusion. In *European Conference on Computer Vision* (pp. 293-309). Cham: Springer Nature Switzerland.

27. Yuan, Y., Huang, J., Yin, J., & Xu, T. (2026). Confidence Calibration and Semantic Error Analysis for Ambiguous Coreference Resolution in User-generated Social Media Text. Available at SSRN 7393238.
28. Zhao, Z., & Welsch, R. E. (2024). Hierarchical reinforced trader (hrt): A bi-level approach for optimizing stock selection and execution. arXiv preprint arXiv:2410.14927.
29. Lyu, Y. (2025). Enhancing Approximate Conformance Checking Accuracy with Hierarchical Clustering Model Behaviour Sampling. *International Journal of Advanced Computer Science & Applications*, 16(8).
30. Yang, Z., Alexandrova, A. N., & Sautet, P. (2026). Modeling CO₂ Hydrogenation to Methanol on an Ensemble of Inverse ZrO₂ on Cu Catalytic Sites: Mechanism, Reactivity, and Deactivation. *Angewandte Chemie*, e5448247.
31. Xu, Z., Zhang, X., Zhou, X., & Zhang, J. (2025). AvatarShield: Visual Reinforcement Learning for Human-Centric Video Forgery Detection. arXiv preprint arXiv:2505.15173.
32. Xi, X., Chen, P., Qiu, Y., Luo, R., Tong, P., & Liang, J. Video-BCI: Bayesian Cognitive Integration of Self-Prior Hypotheses for Video Understanding. In *Forty-third International Conference on Machine Learning*.
33. Liang, J., Xiao, Y., Yang, H., Yu, Z., Liu, J. N., & Yang, K. (2026, April). C-HyPOD: Causal Hyperbolic Representation Learning with Prototype Orthogonal Disentanglement for Graph Out-of-Distribution Recommendation. In *Proceedings of the ACM Web Conference 2026* (pp. 6541-6550).
34. Xu, Z., Zhang, X., Li, R., Tang, Z., Huang, Q., & Zhang, J. (2024). Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. arXiv preprint arXiv:2410.02761.
35. Huang, Q., Xu, Z., Zhang, X., & Zhang, J. (2025). UniShield: An Adaptive Multi-Agent Framework for Unified Forgery Image Detection and Localization. arXiv preprint arXiv:2510.03161.
36. Jiang, C. (2026). Information doesn't work? Testing experiential drivers of VR application in heritage tourism by affective-cognitive pathways. *Journal of Hospitality and Tourism Technology*, 1-28.
37. Wang, T., & Xia, Z. (2025). Stability of In-Context Learning: A Spectral Coverage Perspective. arXiv preprint arXiv:2509.20677.
38. Liu, T., He, J., Zhou, R., Song, J., Liu, H., & Liang, Y. (2025). Revolutionizing clinical decision making through deep learning and topic modeling for pathway optimization. *Scientific Reports*, 15(1), Article 28787. <https://doi.org/10.1038/s41598-025-12679-z>
39. Wu, X., Xing, X., Yao, J., Qian, Q., & Song, J. (2025). Scnet: spectral convolutional networks for multivariate time series classification. *Applied Intelligence*, 55(6), Article 456. <https://doi.org/10.1007/s10489-025-06352-1>

40. Ma, C., Song, J., Xu, Y., Fan, H., Wu, X., & Sun, T. (2024). Vehicle-Based Machine Vision Approaches in Intelligent Connected System. IEEE Transactions on Intelligent Transportation Systems, 25(3), 2827-2836. <https://doi.org/10.1109/TITS.2023.3276325>
41. Chaoyong Ma, Nan Si, Xuewei Song, Chen Liang and Haoming Luo, A Bearing Fault Diagnosis Method Based on Concise Empirical Wavelet Transform, International Journal of Comprehensive Engineering, 2023, 12(1), 1-6