A Comparative Study of Semi-supervised Deep Anomaly Detection and LLMs for Monitoring Patients with Severe Health Status Undergoing Radiotherapy

Yang Yan
Southern Illinois University
yang.yan@siu.edu

Xinglei Shen
University of Kansas Medical Center
xshen@kumc.edu

Ronald Chen
University of Kansas Medical Center
rchen2@kumc.edu

Hao Gao
UT Southwestern Medical Center
hao.gao.2012@gmail.com

Chen Zhao
Baylor University
chen_zhao@baylor.edu

Guihong Wan
Harvard Medical School
guihong_wan@hsph.harvard.edu

Yi He
William & Mary
yihe@wm.edu

Zhong Chen
Southern Illinois University
zhong.chen@siu.edu

Abstract

This study assesses semi-supervised deep anomaly detection (AD) methods and novel zero-shot LLM prompts for identifying prostate cancer patients at risk of severe radiotherapy-induced symptoms via patient-reported outcomes (PROs). While LLMs underperformed compared to semi-supervised AD models in key metrics (e.g., precision, recall, and F1-score), they provided valuable decision explanations and required no training data. This highlights their potential for straightforward clinical deployment without the need for extensive model development.

1 Introduction

The integration of patient-reported outcomes (PROs) [15, 21] in radiation therapy addresses the critical need to capture the subjective patient experience, including symptom burden and quality of life [6], which traditional clinician-reported metrics often miss. This shift toward patient-centered care, driven by regulatory mandates and validated instruments, is essential for evaluating treatment efficacy, guiding personalized decisions, and improving supportive care throughout treatment and survivorship.

To analyze PROs, machine learning [12, 20, 18, 19] is increasingly employed. While traditional models like ensemble methods and deep learning architectures [4, 7, 17] predict high-risk symptoms, they face limitations including data scarcity. This work introduces zero-shot Large Language Model (LLM) prompts under anomaly detection settings. This approach leverages pre-trained clinical reasoning to interpret PRO data without task-specific training, enabling explainable, adaptable, and real-time identification of high-risk patients for proactive intervention.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: 2nd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences.

2 Related Work

While PRO analysis methods specific to RT remain underdeveloped, ML/DL approaches [20] have shown promising applications in PRO research across related domains. For example, Grover et al. [5] compared patient-reported toxicities across prostate cancer treatments, revealing that surgery was associated with higher rates of urinary and sexual side effects compared to RT. Confirmatory factor analysis was employed to identify clinically relevant symptom clusters and their relationships with radiation dose distributions [8]. Brocas et al. [2] further investigated functional outcomes and adverse effects across treatment modalities, including radical prostatectomy, RT, and active surveillance, highlighting differences in patient experiences. Exploratory work by Arbab et al. [1] identified distinct subgroups of cancer patients undergoing RT with divergent mental health and well-being trajectories, alongside their socio-demographic and clinical predictors. Beyond oncology, Staartjes et al. [13] developed deep neural networks and logistic regression models to predict PROs following lumbar discectomy, demonstrating the versatility of ML in surgical outcomes research. Similarly, Tschuggnall et al. [14] applied ML to a multi-modal dataset of over 1,000 patients to predict rehabilitation success, underscoring its utility in post-treatment care. In breast cancer research, Pfob et al. [10] validated ML algorithms to predict clinically significant differences in long-term breast satisfaction using BREAST-Q scores, offering tools for personalized outcome forecasting. Finally, Yang et al. [23] combined ML with traditional statistical modeling to explore associations between radiation treatment parameters and post-therapy gastro-urinary function, bridging technical and patient-centered perspectives.

3 Proposed Method

3.1 Problem Statement

Predicting patients with severe health status undergoing radiotherapy can be framed as an anomaly detection problem (binary classification) by treating the minority of high-risk patients as rare deviations from the majority of patients who respond typically to treatment. In this context, "normal" data corresponds to patients with mild symptoms, while "anomalies" represent those at risk of acute adverse effects, unexpected deterioration or life-threatening reactions. Anomaly detection models can be trained on non-severe cases to learn baseline patterns in features like treatment parameters, biomarkers, or patient history. Deviations from these patterns signal potential high-risk patients, addressing class imbalance inherent in severe outcomes. Challenges include ensuring model sensitivity to clinically meaningful anomalies, minimizing false positives, and integrating temporal data to enable early intervention and timely clinical adjustments.

3.2 Semi-supervised Deep Anomaly Detection Approaches

Deep Semi-supervised Anomaly Detection (DeepSAD) [11] extends the unsupervised deep SVDD framework by incorporating labeled normal and anomalous samples alongside unlabeled data. It learns a latent representation where normal data is tightly clustered around a predefined center (minimizing entropy), while labeled anomalies are pushed away (maximizing entropy), avoiding the restrictive cluster assumption for anomalies; (2) Pairwise Relation prediction-based ordinal regression Network (PReNet) [9] is a deep weakly-supervised anomaly detection method designed to detect both seen and unseen anomalies. It learns discriminative patterns by predicting pairwise relations (anomaly-anomaly, anomaly-unlabeled, unlabeled-unlabeled) between instances, leveraging a small set of labeled anomalies and a large unlabeled dataset (mostly normal). By training a neural network to assign ordinal scores to these pairs, PReNet captures various abnormality and normality interactions, avoiding overfitting to seen anomalies. During inference, anomaly scores for test instances are computed by aggregating predictions from multiple pairs formed with training data; (3) Extreme Gradient Boosting Outlier Detection (XGBOD) [24] is a semi-supervised ensemble framework designed to enhance anomaly detection by integrating unsupervised representation learning with supervised classification. It operates in three phases: first, it applies diverse unsupervised outlier detection methods to generate transformed outlier scores (TOS), which serve as enriched data representations capturing outlier characteristics; second, it selects the most useful TOS via strategies like accuracy-based or diversity-aware selection to balance computational efficiency and predictive power; third, it combines these selected TOS with original features and trains an XGBoost classifier on the augmented feature space.

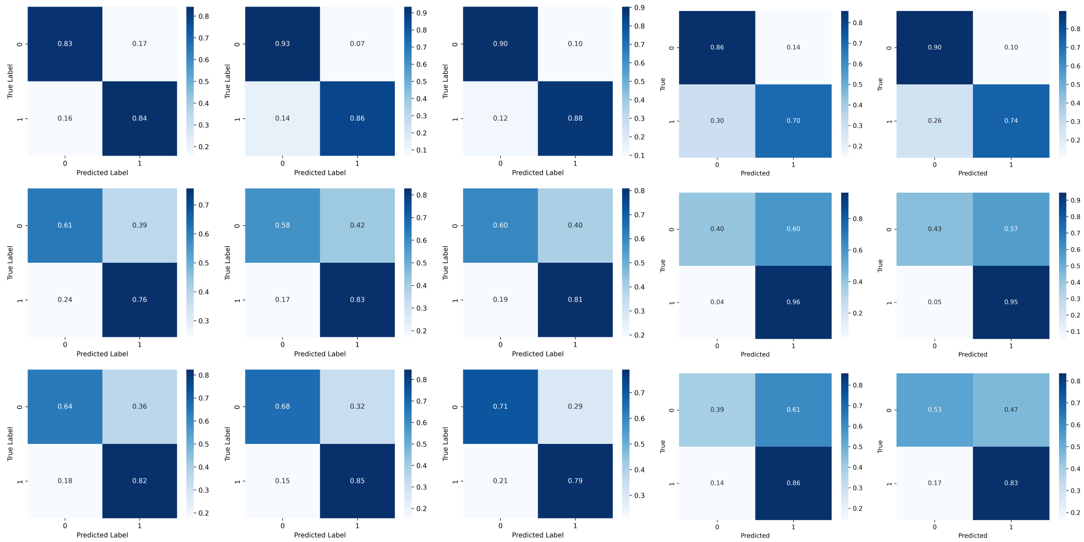

Figure 1: Confusion matrix of three different semi-supervised AD approaches (First column: DeepSAD, Second column: PReNet, Third column: XGBOD) and two zero-shot LLM-AD AD approaches (Fourth and Fifth columns) for prostate cancer patients (First row: bowel pain; Second row: sexual function; Third row: depression feeling).

A Reasoning Example in Llama 3.1-8B

Text sample: “Overall, how much of a problem has your urinary function been for you? : Big problem | Increased frequency of your bowel movements: Big problem | ... ... | Sleeping- How many times did you most typically get up to urinate from the time you went to bed at night until the time you got up in the morning?: Two Times (2) | Total International Prostate Symptom Score = 1-7 Mild Symptoms 8-19 Moderate Symptoms 20-35 Severe Symptoms Regardless of the score, if your symptoms are bothersome you should notify your doctor: 30.0 | If you were to spend the rest of your life with your urinary condition just the way it is now, how would you feel about that?: Unhappy”

Response: {“The patient reports a ‘Big problem’ with urinary function, pain or burning with urination, weak urine stream/incomplete bladder emptying, need to urinate frequently, increased frequency of bowel movements, and overall problems with bowel habits. These symptoms suggest a significant impact on the patient’s quality of life, indicating a clinically significant rectal pain or urgency of bowel movements issue.”, “anomaly_score”: 0.9}
Note: The patient’s responses indicate a high level of symptom severity, particularly with regard to urinary function and bowel habits, which suggests a clinically significant rectal pain issue. The anomaly score of 0.9 reflects a high confidence in this assessment.

3.3 LLM-AD Prompts

The design of the prompts plays an important role for the zero-shot LLM-based anomaly detection as the performance of LLM is highly affected by the clarity and correctness of the prompts. We implement two settings in AD-LLM [22] using Meta Llama 3.1-8B. The prompt template setting for “Zero-shot LLM-based using patients with mild and severe symptoms” is summarized in Table 2 (Appendix). The prompt template setting for “Zero-shot LLM-based using patients with mild symptoms solely” is similar to Table 2. The main difference between these two prompt settings is that the anomaly categories in the zero-shot LLM are defined as mild and severe symptoms, which are highlighted in red in Table 2 (Appendix).

Task Information [3] and Chain-of-Thought (CoT) [16] techniques are used in prompt design. Firstly, we give a clear definition of the PRO task. We collaborate with a radiation oncologist to carefully design scenarios of severe symptoms to ensure that the LLMs can understand anomaly definition in text-based PRO reports. Secondly, a strict scoring mechanism description is given to LLM for determining the anomaly score. Thirdly, step-by-step reasoning with CoT encourages LLMs to sequentially produce the reason and anomaly score in a structural and logical way. Lastly, a clear response format ensures the LLM can answer in a correct JSON format.

Text Box 3.2 shows the reasoning process of the zero-shot Llama 3.1-8B model. For the input text sample, each PRO questionnaire and corresponding response are organized in "key: value" format, separated by pip symbols (|). By passing the **Text Sample** to the Prompt, the PROs query will embedded as "{text}" in Table 2 (Appendix). The Llama 3.1-8B follows the prompt and interprets structured PRO data then provides the reason and anomaly score. For example, in Box 3.2, the LLM reports the patient's responses to indicate a severe problem with urinary function and increased frequency of bowel movements. As a result, these symptoms suggest a clinically significant rectal or urgency of bowel movements issue. Zero-shot Llama 3.1-8B gives a high confidence in this assessment with an anomaly score of 0.9.

4 Experimental Results

4.1 PRO Data

The prostate PRO dataset comprises 22 instruments measuring urinary, bowel, sexual, and vitality/hormonal health, featuring three target measurements: (1) rectal pain/urgency of bowel (assessed via EPIC-CP/IPSS for HRQoL), (2) sexual function, and (3) depression (both capturing prostate-specific symptoms). Responses were classified as mild symptoms ("No problem", "Very small problem", "Small problem") or severe symptoms ("Moderate problem", "Big problem"). Following exclusion of records with missing PRO data, the final cohort includes: 10,356 mild vs. 586 severe cases for rectal pain/urgency; 8,444 mild vs. 1,998 severe cases for sexual function; and 10,266 mild vs. 665 severe cases for depression.

4.2 Baseline Methods

In our experiments, the baselines include three semi-supervised AD-PRO approaches (i.e., DeepSAD [11], PReNet [9], and XGBOD [24]), and our two zero-shot LLM-AD methods. One is the zero-shot LLM-based prompts using patients with mild symptoms solely (Meta Llama-3.1-8B) for PRO severe symptoms prediction. Another one is the zero-shot LLM-based Prompts using patients with mild and severe symptoms (Meta Llama-3.1-8B) for PRO severe symptoms prediction. We perform the experiments on a Windows workstation with a 3.10 GHz Intel Xeon w5-2445 processor and 32GB 4800 MHz memory. Llama 3.1-8B runs on dual NVLink NVIDIA RTX A4500 GPU with 40 GB RAM.

4.3 Overall Comparison

Table 1: Overall Performance Comparison of different AD Models (including AUC, AUCPR, Precision, Recall, and F1-score).

Category	Model	Bowel Pain					Depressed Feeling					Sexual Function				
		AUC	AUCPR	Precision	Recall	F1	AUC	AUCPR	Precision	Recall	F1	AUC	AUCPR	Precision	Recall	F1
Semi-supervised AD	DeepSAD	0.9200	0.8618	0.7416	0.8426	0.7888	0.7872	0.6982	0.5982	0.8233	0.6919	0.7522	0.3771	0.3177	0.7568	0.4475
	PReNet	0.9565	0.9254	0.8826	0.8638	0.8731	0.8240	0.7312	0.6338	0.8459	0.7246	0.7691	0.4033	0.3198	0.8283	0.4616
	XGBOD	0.9554	0.9123	0.8381	0.8809	0.8589	0.7890	0.6929	0.6344	0.7890	0.7035	0.7745	0.4182	0.3263	0.8088	0.4650
Zero-shot LLM-AD	Llama-3.1-8B Mild	0.8352	0.7173	0.7348	0.6997	0.7168	0.6555	0.4993	0.4774	0.8571	0.6132	0.7078	0.2901	0.2769	0.9650	0.4303
	Llama-3.1-8B Mild + Severe	0.8843	0.8038	0.8011	0.7423	0.7706	0.7274	0.5799	0.5362	0.8346	0.6529	0.6938	0.2802	0.2836	0.9459	0.4364

Table 1 presents the overall experimental results comparing all approaches across AUC, AUCPR, Precision, Recall, and F1 score. Overall, the semi-supervised model PReNet demonstrates the strongest and most consistent performance. It achieves the highest scores in the Bowel Pain category across all metrics (AUC: 0.9565, AUCPR: 0.9254, Precision: 0.8826, Recall: 0.8638, F1: 0.8731) and also leads in the Depressed Feeling category. The XGBOD model is a close competitor, particularly in Bowel Pain, where its performance is nearly on par with PReNet. In contrast, the DeepSAD model shows solid performance in Bowel Pain but falls behind in the other two categories.

The zero-shot LLM-based models, Llama-3.1-8B Mild and Llama-3.1-8B Mild+Severe, exhibit a significant performance trade-off, particularly in the Sexual Function category. While their Recall

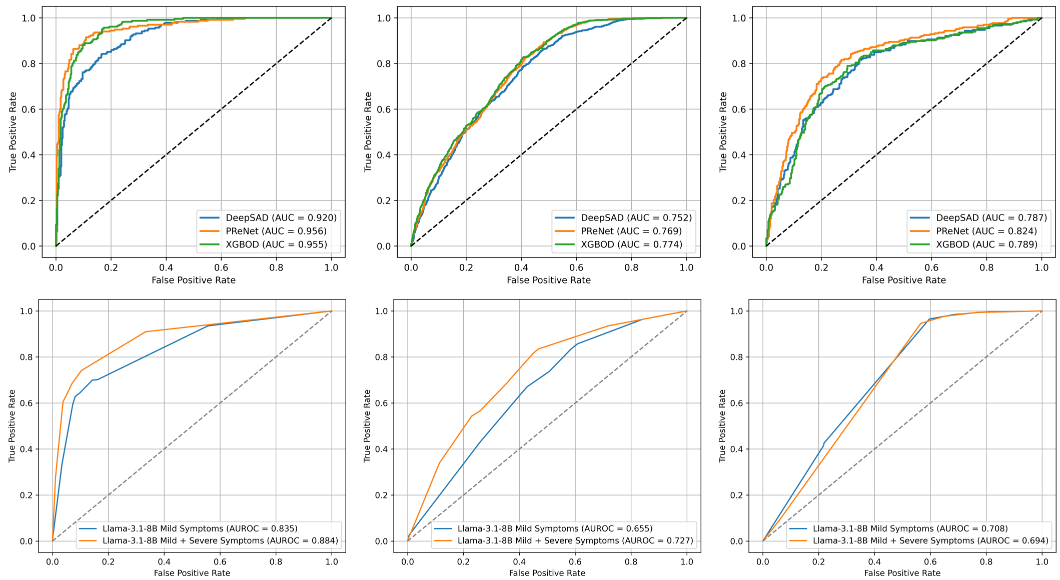

Figure 2: ROC curves with AUC values of different semi-supervised AD approaches (top panel) and zero-shot LLM-AD AD approaches (bottom panel) for prostate cancer patients (First column: bowel pain; Second column: sexual function; Third column: depression feeling).

scores are exceptionally high (up to 0.9650), indicating a strong ability to identify most true severe patients, this comes at the cost of very low Precision (as low as 0.2769). This results in a high number of false positives and consequently poor F1-scores (~ 0.43) and AUCPR values (~ 0.28), the lowest among all models. This pattern suggests that without fine-tuning, these LLMs tend to be overly cautious, flagging too many instances as anomalous in the more challenging Sexual Function task.

4.4 Confusion Matrix Comparison

The PReNet and XGBOD methods have demonstrated significantly better confusion matrices, achieving high diagonal values for both Class 0 and Class 1. This improvement is especially pronounced for patients with severe bowel pain. The zero-shot Llama 3.1-8B model (Figure 1) also shows this capability. Furthermore, for severe sexual function problems and depression feelings, zero-shot Llama 3.1-8B demonstrates promising performance, achieving very high true positive rates (0.96/0.86 and 0.95/0.83, respectively) – significantly exceeding DeepSAD and XGBOD on these specific symptoms.

4.5 ROC Curves and AUC Comparison

From Figure 2, PReNet and XGBOD (using 50% labels) follow closely with highly competitive performance. This indicates PReNet and XGBOD excel at classifying both majority (mild/intermediate symptom) and minority (severe symptom) classes. While zero-shot Llama 3.1-8B achieves relatively high AUC/AUCPR, it misidentifies nearly half of the majority class (mild/intermediate symptoms) compared to PReNet and XGBOD. Among semi-supervised methods, PReNet and XGBOD surpass DeepSAD in AUC/AUCPR, though all three demonstrate competitive ROC/PR curve performance.

5 Conclusion

In this study, we investigate a set of semi-supervised and zero-shot LLM-based anomaly detection approaches to classify patients' health status with mild/intermediate and severe outcomes. According to the confusion matrix and ROC/PR curve evaluation, PReNet has showcased a good potential by considering patients with mild/intermediate and severe health status with partial label information. Without any label information, zero-shot Llama 3.1-8B is a very good fit for patients who suffer severe sexual function problems and depression feelings as the true positive rates are relatively high.

References

[1] Mona Arbab, Yu-Hui Chen, Roy B Tishler, Lauren Gunasti, Jason Glass, Jo Ann Fugazzotto, Joseph H Killoran, Rosh Sethi, Eleni Rettig, Donald Annino, et al. Association between radiation dose to organs at risk and acute patient reported outcome during radiation treatment for head and neck cancers. *Head & Neck*, 44(6):1442–1452, 2022.

[2] Daniel A Barocas, JoAnn Alvarez, Matthew J Resnick, Tatsuki Koyama, Karen E Hoffman, Mark D Tyson, Ralph Conwill, Dan McCollum, Matthew R Cooperberg, Michael Goodman, et al. Association between radiation therapy, surgery, or observation for localized prostate cancer and patient-reported outcomes after 3 years. *Jama*, 317(11):1126–1140, 2017.

[3] Yunkang Cao, Xiaohao Xu, Chen Sun, Xiaonan Huang, and Weiming Shen. Towards generic anomaly detection and understanding: Large-scale visual-linguistic model (gpt-4v) takes the lead. *arXiv preprint arXiv:2311.02782*, 2023.

[4] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *KDD*, pages 785–794, 2016.

[5] Surbhi Grover, James M Metz, Carolyn Vachani, Margaret K Hampshire, Gloria A DiLullo, and Christine Hill-Kayser. Patient-reported outcomes after prostate cancer treatment. *Journal of Clinical Urology*, 7(4):286–294, 2014.

[6] Melisa Guelhan Inci, Rolf Richter, Kathrin Heise, Ricarda Dukatz, Hannah Woopen, and Jalid Sehouli. Patient-reported outcomes (pros) and health-related quality of life (hr-qol) in patients with ovarian cancer: What is different compared to healthy women? *Cancers*, 13(4):631, 2021.

[7] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *NeurIPS*, 30, 2017.

[8] Jung Hun Oh, Maria Thor, Caroline Olsson, Viktor Skokic, Rebecka Jörnsten, David Alsadius, Niclas Pettersson, Gunnar Steineck, and Joseph O Deasy. A factor analysis approach for clustering patient reported outcomes. *Methods of Information in Medicine*, 55(05):431–439, 2016.

[9] Guansong Pang, Chunhua Shen, Huidong Jin, and Anton van den Hengel. Deep weakly-supervised anomaly detection. In *KDD*, pages 1795–1807, 2023.

[10] André Pfob, Babak J Mehrara, Jonas A Nelson, Edwin G Wilkins, Andrea L Pusic, and Chris Sidey-Gibbons. Machine learning to predict individual patient-reported outcomes at 2-year follow-up for women undergoing cancer-related mastectomy and breast reconstruction (inspired-001). *The Breast*, 60:111–122, 2021.

[11] Lukas Ruff, Robert A Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. *arXiv preprint arXiv:1906.02694*, 2019.

[12] Mohammad Shehab, Laith Abualigah, Qusai Shambour, Muhannad A Abu-Hashem, Mohd Khaled Yousef Shambour, Ahmed Izzat Alsalibi, and Amir H Gandomi. Machine learning in medical applications: A review of state-of-the-art methods. *Computers in Biology and Medicine*, 145:105458, 2022.

[13] Victor E Staartjes, Marlies P de Wispelaere, William Peter Vandertop, and Marc L Schröder. Deep learning-based preoperative predictive analytics for patient-reported outcomes following lumbar discectomy: feasibility of center-specific modeling. *The Spine Journal*, 19(5):853–861, 2019.

[14] Michael Tschuggnall, Vincent Grote, Michael Pirchl, Bernhard Holzner, Gerhard Rumpold, and Michael J Fischer. Machine learning approaches to predict rehabilitation success based on clinical and patient-reported outcome measures. *Informatics in Medicine Unlocked*, 24:100598, 2021.

[15] Vivek Verma, Charles B Simone, and Mark V Mishra. Quality of life and patient-reported outcomes following proton radiation therapy: a systematic review. *JNCI: Journal of the National Cancer Institute*, 110(4):341–353, 2018.

[16] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35:24824–24837, 2022.

[17] Yi-Xuan Xu, Ming Pang, Ji Feng, Kai Ming Ting, Yuan Jiang, and Zhi-Hua Zhou. Reconstruction-based anomaly detection with completely random forest. In *SDM*, pages 127–135. SIAM, 2021.

[18] Yang Yan, Zhong Chen, Xinglei Shen, Ronald C Chen, and Hao Gao. Short-and long-term weekly patient-reported outcomes prediction undergoing radiotherapy: single-patient time series model vs. transformer-based multi-patient time series model. *BioData Mining*, 18(1):53, 2025.

[19] Yang Yan, Zhong Chen, Cai Xu, Xinglei Shen, Jay Shiao, John Einck, Ronald C Chen, and Hao Gao. An oversampling-enhanced multi-class imbalanced classification framework for patient health status prediction using patient-reported outcomes. *IEEE Access*, 2025.

[20] Yang Yan, Christopher Lominska, Gregory N Gan, Hao Gao, and Zhong Chen. Accounting for cancer patients with severe outcomes: An anomaly detection perspective. In *IEEE BigData*, pages 8253–8255. IEEE, 2024.

[21] Yang Yan, Christopher Lominska, Gregory N Gan, Hao Gao, and Zhong Chen. Advanced anomaly detection framework for enhancing prediction of severe health outcomes in cancer patients undergoing radiotherapy. In *2025 IEEE 11th International Conference on Big Data Computing Service and Machine Learning Applications (BigDataService)*, pages 73–80. IEEE, 2025.

[22] Tiankai Yang, Yi Nian, Shawn Li, Ruiyao Xu, Yuangang Li, Jiaqi Li, Zhuo Xiao, Xiyang Hu, Ryan Rossi, Kaize Ding, et al. Ad-llm: Benchmarking large language models for anomaly detection. *arXiv preprint arXiv:2412.11142*, 2024.

[23] Zhijian Yang, Daniel Olszewski, Chujun He, Giulia Pintea, Jun Lian, Tom Chou, Ronald C Chen, and Blerta Shtylla. Machine learning and statistical prediction of patient quality-of-life after prostate radiation therapy. *Computers in Biology and Medicine*, 129:104127, 2021.

[24] Yue Zhao and Maciej K Hryniewicki. Xgbod: improving supervised outlier detection with unsupervised representation learning. In *IJCNN*, pages 1–8. IEEE, 2018.

Appendix: Zero-shot LLM Prompts

Table 2: LLM prompt template used for zero-shot LLM-AD with mild and severe symptoms on pain/urgency of bowel dataset.

{normal_categories} refers to the ["No problem", "Very small problem", "Small problem"] on rectal pain/urgency of bowel symptoms. {anomaly_categories} refers to ["Moderate problem", "Big problem"] on rectal pain/urgency of bowel symptoms.

```
You will analyze structured clinical survey responses (EPIC-CP / IPSS format) to determine whether
the prostate cancer patient's symptom profile indicates a clinically significant Rectal pain or
urgency of bowel movements issues.


## Categories:
Normal Category(ies):
{normal_categories}
Anomaly Category(ies):
{anomaly_categories}


## Guidelines:
1. Anomaly Definition:
  - The patient's report is considered anomalous if symptoms are consistently severe and suggest
advanced-stage concerns regarding bowel function.

2. Scoring:
  - Assign an anomaly confidence score between 0 and 1.
  - Use higher scores when you are highly confident in your decision.
  - Use lower scores when you believe the response is not indicative of significant issues.

3. Step-by-step Reasoning (Chain of Thought):
  - Step 1. Read the entire text (The patient's report) sample carefully and understand it
thoroughly.
  - Step 2. Analyze the text sample by comparing its content to {changable_part}.
  - Step 3. Determine which category (normal or anomaly) the text sample most closely aligns with.
    -If it most closely aligns with any of the anomaly category(ies) on rectal pain or urgency of
bowel movements, it is an anomaly.
    - If it most closely aligns with any of the normal category(ies) on rectal pain or urgency of
bowel movements instead, it is not an anomaly.
  - Step 4. Assign an anomaly confidence score based on how confident you are that the text sample
is an anomaly.

4. Response Format:
  - Provide responses in a strict JSON format with the keys "reason" and "anomaly_score".
    - "reason": Your brief explanation of the reasoning in one to three sentences logically.
    - "anomaly_score": Your anomaly confidence score between 0 and 1.
  - Ensure the JSON output is correctly formatted, including correct placement of commas between
key-value pairs.
  - Add a backslash (\) before any double quotation marks (") within the values of JSON output for
proper parsing (i.e., from " to \"), and ensure that single quotation marks (') are preserved without
escaping.


Text sample:
"{text}"


Response in JSON format:
```