Complexity-based Analysis for Anomaly Detection in Industrial Control Systems

Saddam Hussain, Ali Tufail, Abul Ghani Haji Naim
School of Digital Science, Universiti Brunei Darussalam
saddamicup1993@gmail.com, ali.tufail@ubd.edu.bn, ghani.naim@ubd.edu.bn

Abstract

Industrial Control Systems (ICS) are important to critical infrastructure and are increasingly vulnerable to cyber threats due to their growing interconnectivity and complexity. The paper provides a complexity-based framework of feature evaluation in ICS cybersecurity based on the Secure Water Treatment (SWaT) datasets. The integrated framework measures the complexity of datasets by incorporating a number of complexity measures (feature-based, neighborhood-based, linearity-based and topological) into a single aggregative complexity score that depicts the complexity of a dataset. The Normalizing method is then used to remove the scale bias to ensure that the measures can be compared adequately. This principled dimensionality methodology also increases the interpretability of systems.

1 Introduction

ICS are a foundation of important infrastructure in the energy, manufacturing, water treatment and transportation industries (1; 2; 3). It include sensors, actuators and Programmable Logic Controllers (PLCs) for controlling real-time processes (4). ICS are now interconnected to external systems by facilitating automation and remote access which broadens the attack surface and opens the ICS to advanced attacks (5).

To overcome this challenge, we have applied the SWaT (2023) testbed (6; 7) that is a six-stage and high fidelity model of a water purification plant containing 51 sensors and actuators. SWaT offers labeled datasets for normal operations and various attacks, thus becoming a top-tier benchmark for the security of ICS. However, its high dimensionality complicates feature selection, explainability, and deployment in real-time. In the paper, we introduce a complexity-based feature evaluation scheme integrating twelve known measures from geometry, neighborhood interactions, linearity, and topology into a normalized composite measure for robust inter-comparisons and bias reduction(8; 9). The following are primarily contributions of our paper.

- **Complexity-Based Framework:** An integrated methodology merging twelve complexity measures (specifying feature-, neighborhood-, linearity- and topology based complexity) into one aggregation score.
- **Normalization for Comparability:** Min-max scaling ensures that all the measures are comparable and do not suffer from scale bias.
- **Key Informative Measures:** In order to identify relevant features, it is best to define the indicators in terms of F1 (Fisher's ratio), N1 (borderline fraction), and L2 (linear model error).

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

2 Materials and Method

In this section, we present the methodology used to calculate the complexity of a given dataset by using a variety of potential measures of data separability and structure. Following the SWaT dataset, we adopt twelve complexity computations categorized into feature-based $(F1 - F4)$, neighbourhood-based $(N1 - N4)$, linearity-based $(L1 - L3)$ and topological $(T1)$ metrics from (10; 11). Each one of the measures of the complexity has been outlined below in detail with mathematical formulations.

2.1 Feature-Based Complexity Metrics:

These matrices measure how effectively individual features or the combination of all of them go in separating one class from another.

F1 *– Fisher's Discriminant Ratio:* The Fisher ratio is a numerical representation of the degree of separation of the feature in terms of the ratio of the inter-class variance to the intra-class variance. For feature i, let $\mu_{i,c}$ be the mean of class c, μ_i the global mean, and n_c the number of samples in class c:

$$F1_i = \frac{\sum_c n_c(\mu_{i,c} - \mu_i)^2}{\sum_c \sum_{x_j \in c}(x_{ji} - \mu_{i,c})^2} \tag{1}$$

F2 – Overlap Volume: Consider two classes whose feature values lie in the intervals $[a_1, b_1]$ and $[a_2, b_2]$. The normalized overlap between them is given by:

$$[\max(a_1, a_2), \min(b_1, b_2)], \quad F2_i = \frac{\max\big(0, \min(b_1, b_2) - \max(a_1, a_2)\big)}{\max(b_1, b_2) - \min(a_1, a_2)}, \quad F2 = \prod_{i=1}^{d} F2_i \tag{2}$$

F3 – Individual Feature Efficiency:

For feature i, let R_i be the overlap between classes c_1 and c_2, $F3_i$ be the rate of instances which lie outside R_i, reflecting a measure of its discriminative power.

$$F3_i = \frac{\#(x_i < \min(R_i) \vee x_i > \max(R_i))}{N}, \quad F3 = \max_i(F3_i) \tag{3}$$

F4 – Collective Feature Efficiency: This is a measure of the joint discriminative power of features. It iteratively select the feature that separates the most number of points outside the overlap interval $R_i = [\min(c_1), \max(c_1)] \cap [\min(c_2), \max(c_2)]$. Selected points are marked as *covered*, and the final score is the proportion of covered points.

$$F4 = \frac{\#\text{Covered Points}}{N} \tag{4}$$

2.2 Neighborhood-Based Complexity Metrics:

These metrics analyze relationships between instances and their neighbors to assess boundary ambiguity and decision complexity.

N1, N2 and N3 Complexity Measures: We consider three neighborhood-based complexity metrics. **N1** (Borderline Fraction in MST) is the proportion of Minimal Spanning Tree (MST) edges connecting samples from different classes. **N2** (Intra–Extra Class Distance Ratio) compares nearest intra- and extra-class distances. **N3** (Leave-One-Out 1NN Error Rate) is the misclassifications rate under leave-one-out cross-validation with the 1-nearest neighbor rule. Their definitions are given by:

$$N1 = \frac{E_b}{N-1}, \quad N2 \quad = \frac{\sum_i d_{\text{extra}}(x_i)}{\sum_i d_{\text{intra}}(x_i)}, \quad N3 = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\Big(y_i \neq \arg\min_{j \neq i} \|x_i - x_j\|\Big). \tag{5}$$

N4 *– Nonlinearity of 1NN Classifier:* To assess the complexity of the decision boundary learnt by the 1NN classifier, we generate interpolated points between randomly sampled pairs from the same class. These interpolations are then classified using 1NN. The misclassification rate is defined as:

$$N4 = \frac{\#(\hat{y}_{\text{interp}} \neq y_{\text{interp}})}{T} \tag{6}$$

2.3 Linearity-Based Complexity Metrics:

These metrics evaluate how well linear models fit the data, indicating whether simple or more complex classifiers are needed.

***L1** – Linear Model Error Sum:* A linear model is trained on the full dataset, and its predictions $\hat{y}_i$ are compared to the true labels y_i. ***L2** – Cross-Validated Linear Model Error:* Using stratified K-fold cross-validation (e.g., $K = 5$). ***L3** – Nonlinearity of Linear:* Similar to $N4$, this metric evaluates how well a linear model preserves class identity under interpolation.

$$L1 = \sum_{i=1}^{N} |y_i - \hat{y}_i|, \quad L2 = \frac{1}{K} \sum_{k=1}^{K} \frac{\#(\hat{y}_k \neq y_k)}{N_k}, \quad L3 = \frac{\#(\hat{y}_{\text{interp}} \neq y_{\text{interp}})}{T} \tag{7}$$

2.4 Topological-Based Complexity Metric:

This measure checks the local structure of the data in terms of the set of hyperspheres centred on each point, by checking how "safe" they are (how consistent in terms of class).

T1 – Safe Hypersphere Ratio: For each point x_i, let r_i be the smallest radius such that the ball $B(x_i, r_i)$ contains at least one sample from a different class. A hypersphere is safe if all neighbors within r_i belong to the same class:

$$\forall j : \|x_j - x_i\| < r_i \;\Rightarrow\; y_j = y_i, \;\; T1 = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\text{hypersphere around } x_i \text{ is safe}) \tag{8}$$

2.5 Normalization:

All measures are scaled to $[0, 1]$ using min–max normalization to ensure comparability across metrics:

$$\text{Normalized} = \frac{\text{Value} - \text{Global Min}}{\text{Global Max} - \text{Global Min}} \tag{9}$$

Here, *Global Min* and *Global Max* denote the minimum and maximum observed values of each measure across all datasets.

2.6 Experimental Setup

The dataset is represented as $\mathbf{X} \in \mathbb{R}^{n \times d}$ with labels $\mathbf{y} \in \{c_1, \ldots, c_k\}^n$. Missing values in $\mathbf{X}$ are handled by mean imputation, replacing each missing entry with the column mean μ_j.

$$x'_{ij} = \begin{cases} x_{ij}, & \text{if } x_{ij} \text{ is not missing} \\ \mu_j = \frac{1}{N_j} \sum_{x_{ij} \text{ not missing}} x_{ij}, & \text{otherwise} \end{cases} \tag{10}$$

3 Results and Discussion

The results of the complexity analysis across ten distinct configurations of the SWaT dataset are summarized in Table 1, which presents the values of twelve complexity measures along with the computed *aggregation score* for each dataset. The overall variation in aggregation scores across the datasets configurations is shown in Figure 1. The aggregation score reflects the degree of consistency among the complexity measures in identifying the difficulty of classification tasks. A *higher aggregation score* indicates that the dataset is *more complex*, as the measures collectively suggest a higher degree of difficulty in distinguishing between normal and attack instances.

Table 1: Complexity measures for datasets with computed Aggregation score

Datasets	F1	F2	F3	F4	N1	N2	N3	N4	L1	L2	L3	T1	Aggregation
Dataset 1	0.0010	0.0000	0.9917	1.0000	0.0011	0.0041	0.0014	0.5000	2.0000	0.2363	0.5000	1.0000	0.252
	0.0098	0.0000	0.9911	0.9994	0.0011	0.0024	0.0014	0.5000	2.0000	0.0989	0.5000	1.0000	
	0.0011	0.0000	0.9803	0.9994	0.0011	0.0050	0.0011	0.5000	2.0000	0.1834	0.5000	1.0000	
	0.0021	0.0000	0.9946	1.0000	0.0018	0.0051	0.0027	0.5000	2.0000	0.0200	0.5000	1.0000	
Dataset 2	0.0005	0.0000	0.9741	1.0000	0.0013	0.0048	0.0013	0.5000	2.0000	0.0240	0.5000	1.0000	0.258
	0.0088	0.0000	0.9836	1.0000	0.0011	0.0054	0.0008	0.5000	2.0000	0.1007	0.5000	1.0000	
Dataset 3	5.7323	0.0000	1.0000	1.0000	0.0003	0.0048	0.0000	0.0500	0.0000	0.1289	0.0000	1.0000	**0.678**
	0.0020	0.0000	0.9837	0.9994	0.0011	0.0047	0.0011	0.5000	2.0000	0.1839	0.5000	1.0000	
	0.0044	0.0000	0.9449	1.0000	0.0027	0.0064	0.0034	0.5000	2.0000	0.2913	0.5000	1.0000	
Dataset 4	0.0041	0.0000	0.9656	0.9987	0.0026	0.0041	0.0026	0.5000	2.0000	0.2664	0.5000	1.0000	0.661
	0.0015	0.0000	0.9538	1.0000	0.0008	0.0044	0.0008	0.5000	2.0000	0.3647	0.5000	1.0000	
	6.2594	0.0000	1.0000	1.0000	0.0003	0.0047	0.0000	0.0500	0.0000	0.2589	0.0000	1.0000	
	0.0013	0.0000	0.9773	1.0000	0.0011	0.0046	0.0014	0.5000	2.0000	0.4241	0.5000	1.0000	
Dataset 5	0.0021	0.0000	0.7619	0.9994	0.0013	0.0058	0.0016	0.5000	2.0000	0.3145	0.5000	1.0000	0.272
	0.0016	0.0000	0.9754	0.9994	0.0011	0.0060	0.0014	0.5000	2.0000	0.3324	0.5000	1.0000	
	0.0010	0.0000	0.9651	1.0000	0.0011	0.0064	0.0014	0.5000	2.0000	0.2418	0.5000	1.0000	
	0.0020	0.0000	0.9914	1.0000	0.0011	0.0052	0.0011	0.5000	2.0000	0.1903	0.5000	1.0000	
Dataset 6	0.0009	0.0000	0.9806	0.9994	0.0011	0.0059	0.0008	0.5000	2.0000	0.1613	0.5000	1.0000	0.265
	0.0010	0.0000	0.9778	1.0000	0.0008	0.0049	0.0011	0.5000	2.0000	0.2388	0.5000	1.0000	
	0.0010	0.0000	0.8233	1.0000	0.0008	0.0058	0.0011	0.5000	2.0000	0.1752	0.5000	1.0000	
Dataset 7	0.0013	0.0000	0.9399	0.9997	0.0011	0.0064	0.0011	0.5000	2.0000	0.2407	0.5000	1.0000	0.281
	0.0014	0.0000	0.9635	1.0000	0.0020	0.0085	0.0020	0.5000	2.0000	0.1504	0.5000	1.0000	
	0.0003	0.0000	0.9712	0.9994	0.0011	0.0063	0.0011	0.5000	2.0000	0.2731	0.5000	1.0000	
	0.0009	0.0000	0.9853	1.0000	0.0011	0.0042	0.0011	0.5000	2.0000	0.0731	0.5000	1.0000	
	0.0011	0.0000	0.9814	0.9994	0.0011	0.0059	0.0014	0.5000	2.0000	0.5213	0.5000	1.0000	
Dataset 8	0.0015	0.0000	0.8977	1.0000	0.0011	0.0037	0.0011	0.5000	2.0000	0.1086	0.5000	1.0000	0.269
	0.0008	0.0000	0.9601	1.0000	0.0011	0.0159	0.0014	0.5000	2.0000	0.0006	0.5000	1.0000	
	0.0014	0.0000	0.9773	1.0000	0.0011	0.0042	0.0008	0.5000	2.0000	0.4198	0.5000	1.0000	
	0.0004	0.0000	0.9227	1.0000	0.0012	0.0056	0.0012	0.5000	2.0000	0.2637	0.5000	1.0000	
Dataset 9	0.0009	0.0000	0.8937	1.0000	0.0011	0.0070	0.0014	0.5000	2.0000	0.2875	0.5000	1.0000	0.312
	0.0010	0.0000	0.8649	1.0000	0.0011	0.0058	0.0008	0.5000	2.0000	0.2838	0.5000	1.0000	
	0.0012	0.0000	0.9055	1.0000	0.0011	0.0066	0.0014	0.5000	2.0000	0.1917	0.5000	1.0000	
	0.0088	0.0000	0.9775	1.0000	0.0053	0.0207	0.0040	0.5000	2.0000	0.0649	0.5000	1.0000	
	0.1848	0.0000	0.9831	1.0000	0.0011	0.0056	0.0014	0.5000	2.0000	0.1188	0.5000	1.0000	
Dataset 10	0.0055	0.0000	0.9868	1.0000	0.0014	0.0019	0.0014	0.5000	2.0000	0.1482	0.5000	1.0000	0.244

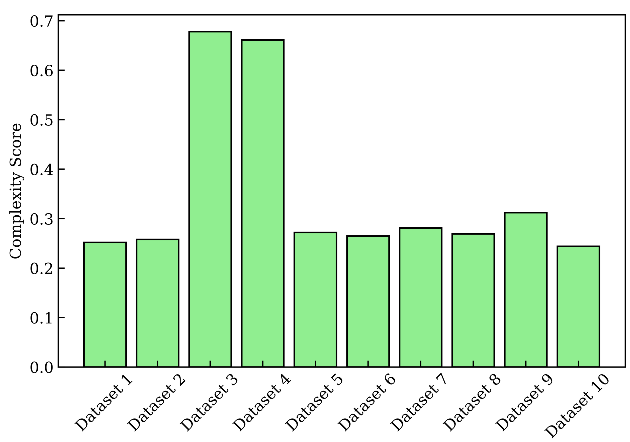

Figure 1: Complexity Aggregation Score of the SWaT datasets

3.1 Key Observations from Complexity Measures

Dataset 3 achieved the highest aggregation score of 0.678, making it the most complex. Notably, this dataset exhibits extreme values in several key measures:

- **F1 = 5.7323**: This high Fisher's Discriminant Ratio suggests that at least one feature in this dataset has strong class-separating power, yet the overall complexity remains high, possibly due to conflicting or overlapping features.
- **F3 = 1.0000 and F4 = 1.0000**: TThese values point to the fact that all instances are perfectly separated with the help of individual and collective features, however, other measures (L1, N4 = 0.0000, 0.0500) imply that the separability of instances by linear and neighborhood are not so good, which implies a contradiction between different types of complexity.

- **L1 = 0.0000 and L3 = 0.0000**: These very low linearity-based errors mean that linear-based models may have difficulty with this dataset, and the separability of instances is apparent.

Dataset 4 achieved the second-highest score (0.661), which has high F1 and F4 (features are highly separable) but high local complexity and boundary uncertainty, with class compactness outweighing the class.

Dataset 9 (0.312) shows high L2 (0.2875) and high N2 and thus indicates noise or overlapping areas. Datasets 7 (0.281) and 5 (0.272) are moderate in complexity with compromised linear separability and intra-class compactness; Dataset 5 has F3 = 0.7619, which confirms the fact that there are overlaps in the feature space and a need for feature engineering.

Dataset 10 has the lowest score (0.244) with low N2 and L2, clear boundaries, and is suitable for simple linear models.

3.2 Progressive Aggregation Analysis

When aggregating the first three feature-based measures (F1–F3), Dataset 3 is most complex due to extreme F1, followed by Dataset 4. Including neighborhood measures (e.g., N1, N2) maintains this ranking. As linearity metrics (L1, L2) are added, Datasets 3 and 4 remain dominant, though the gap with others narrows slightly. This stability across aggregation stages confirms their consistently high complexity from multiple perspectives.

3.3 Analysis of Complexity Measure Sensitivity

F1 effectively identifies discriminative features (high in Datasets 3 and 4). N2 reflects class compactness, lower values (Dataset 10) indicate better separability. L2 estimates linear model error, high values (Dataset 9) signal the need for nonlinear approaches. T1 indicates local class consistency, higher values (Dataset 10) correlate with lower complexity.

3.4 Implications for Feature Selection and Anomaly Detection

Highly complex datasets (3 and 4) may benefit from filter-based feature selection prioritizing high F1 and F3. Moderately complex datasets (5 and 7) may require wrapper or embedded methods considering feature interactions. The less complex Dataset 10 can be effectively handled using linear models, which decreases the overhead time as well as increases the interpretability.

Complexity measures bring actionable knowledge. A high N4 score represents nonlinear decision boundaries, and the use of kernel-based or deep learning models can be more appropriate. Dataset 3 is the most complex, as it has conflicting features and unclear boundaries, and dataset 10 is the simplest, as it has well-separated instances with great linear-separability. These results are directly informative for the feature engineering and model choice for ICS anomaly detection.

4 Conclusion

The paper presents a systematic approach to evaluating the intrinsic complexity of datasets using multiple complexity measures to an aggregated score. The results reveal the high variations of intrinsic difficulty of datasets, indicating that Dataset 3 shows the highest complexity owing to conflicting feature signals and unclear boundaries, whereas Dataset 10 is the easiest with clear separability and powerful linear performance. These findings allow data-driven decisions regarding model selection and feature engineering for ICS anomaly detection. In the future, we will extend the presented framework to dynamically guide model configuration and to incorporate additional measures that capture temporal and structural characteristics of data.

Acknowledgment

The authors thank "iTrust, Centre for Research in Cyber Security, Singapore University of Technology and Design for providing the SWaT2023 datasets".

References

[1] G. B. Gaggero, A. Armellin, G. Portomauro, and M. Marchese, "Industrial control system-anomaly detection dataset (ics-add) for cyber-physical security monitoring in smart industry environments," *IEEE Access*, 2024.

[2] Q. Zhu, Y. Ding, J. Jiang, and S.-H. Yang, "Anomaly detection using invariant rules in industrial control systems," *Control Engineering Practice*, vol. 154, p. 106164, 2025.

[3] B. Kim, M. A. Alawami, E. Kim, S. Oh, J. Park, and H. Kim, "A comparative study of time series anomaly detection models for industrial control systems," *Sensors*, vol. 23, no. 3, p. 1310, 2023.

[4] D.-S. Kim and H. Tran-Dang, "An overview on industrial control networks," *Industrial Sensors and Controls in Communication Networks: From Wired Technologies to Cloud Computing and the Internet of Things*, pp. 3–16, 2018.

[5] M. Nankya, R. Chataut, and R. Akl, "Securing industrial control systems: Components, cyber threats, and machine learning-driven defense strategies," *Sensors*, vol. 23, no. 21, p. 8840, 2023.

[6] iTrust, Centre for Research in Cyber Security, Singapore University of Technology and Design, "Secure Water Treatment (SWaT) Dataset, Dec 2023 Release," 2023. Available from iTrust, SUTD upon request.

[7] J. Goh, S. Adepu, K. N. Junejo, and A. Mathur, "A dataset to support research in the design of secure water treatment systems," in *International conference on critical information infrastructures security*, pp. 88–99, Springer, 2016.

[8] M. Z. Ali, A. Abdullah, A. M. Zaki, F. H. Rizk, M. M. Eid, and E. M. El-Kenway, "Advances and challenges in feature selection methods: a comprehensive review," *J. Artif. Intell. Metaheuristics*, vol. 7, no. 1, pp. 67–77, 2024.

[9] A. C. Lorena, L. P. Garcia, J. Lehmann, M. C. Souto, and T. K. Ho, "How complex is your classification problem? a survey on measuring classification complexity," *ACM Computing Surveys (CSUR)*, vol. 52, no. 5, pp. 1–34, 2019.

[10] T. K. Ho and M. Basu, "Complexity measures of supervised classification problems," *IEEE transactions on pattern analysis and machine intelligence*, vol. 24, no. 3, pp. 289–300, 2002.

[11] X. Wan, Z. Zheng, F. Qin, and X. Lu, "Data complexity: a new perspective for analyzing the difficulty of defect prediction tasks," *ACM Transactions on Software Engineering and Methodology*, vol. 33, no. 6, pp. 1–45, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.

- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).

- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [No]

Justification: In our approach, we build upon existing work by utilizing the BERT model, incorporating the contextual information of both the head and tail entities. This allows us to leverage richer contextual cues from both ends of the relationship, enhancing the model's ability to understand and capture the nuances in the data.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.