InterpretableSAD: Interpretable Anomaly Detection in Sequential Log Data

Xiao Han[1], He Cheng[1], Depeng Xu[2], and Shuhan Yuan[1]

[1]Utah State University

[2]University of Arkansas

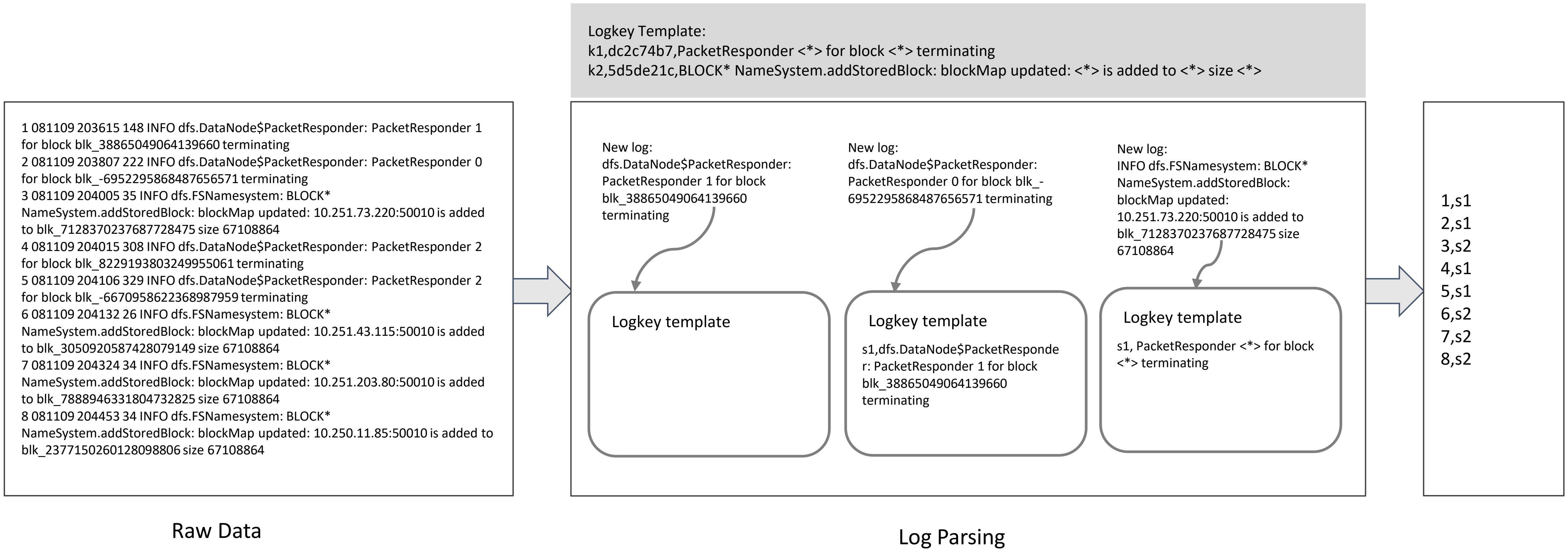
Outline

- **Background**
  - ➢Anomaly Detection
  - ➢System Logs
  - ➢Challenges
- **Preliminary**
- **Problem Statement**
- **Method**
- **Experiment**
- **Conclusion**

What is Anomaly Detection

- Anomaly detection in sequential data aims to identify sequences that deviate from the expected behavior or patterns.
- Anomaly detection receives much attention due to its broad application.
    - E.g., fraud, intrusion, medical, social network, etc.
- Log anomaly detection uses system logs to detect anomalous events or patterns in computer systems.

What are System Logs



Challenges

- Scarcity of anomalous samples
    - ➢Negative sampling algorithm
- Lack of anomalous event interpretability
    - ➢Integrated Gradients (IG)
- No common IG baseline for log data
    - ➢IG baseline generation algorithm

Outline

- Background
- **Preliminary**
  - ➢ Anomaly Detection in Sequential Log Data
  - ➢ Data Augmentation
  - ➢ Interpretable Machine Learning
- Problem Statement
- Method
- Experiment
- Conclusion

Anomaly Detection in Sequential Log Data

- Traditional supervised learning -> Require an enormous number of labeled data
  - ➢Logistic regression, decision tree, and SVM
- Traditional unsupervised learning -> Hard to capture the order information of sequence data
  - ➢PCA, Isolation forest, and OC-SVM
- Deep learning -> No detailed information on the sub-sequence level
  - ➢DeepLog and LogAnomaly

Data Augmentation

- Data augmentation technique is to tackle the scarcity of labeled data issue by artificially expanding the labeled dataset.
- Extensively used in image classification and natural language processing.
    - ➢Rotation and flip for image data, synonym replacement for text data
- Negative sampling is a special data augmentation technique.

Interpretable Machine Learning

- Interpretable machine learning aims at providing a human understandable explanation about the decisions.
- The interpretable anomaly detection models are very limited.
- The attention mechanism provides an attention score that is more about the correlation among events instead of the correlation between events and the label.

Outline

- Background
- Preliminary
- **Problem Statement**
- Method
- Experiment
- Conclusion

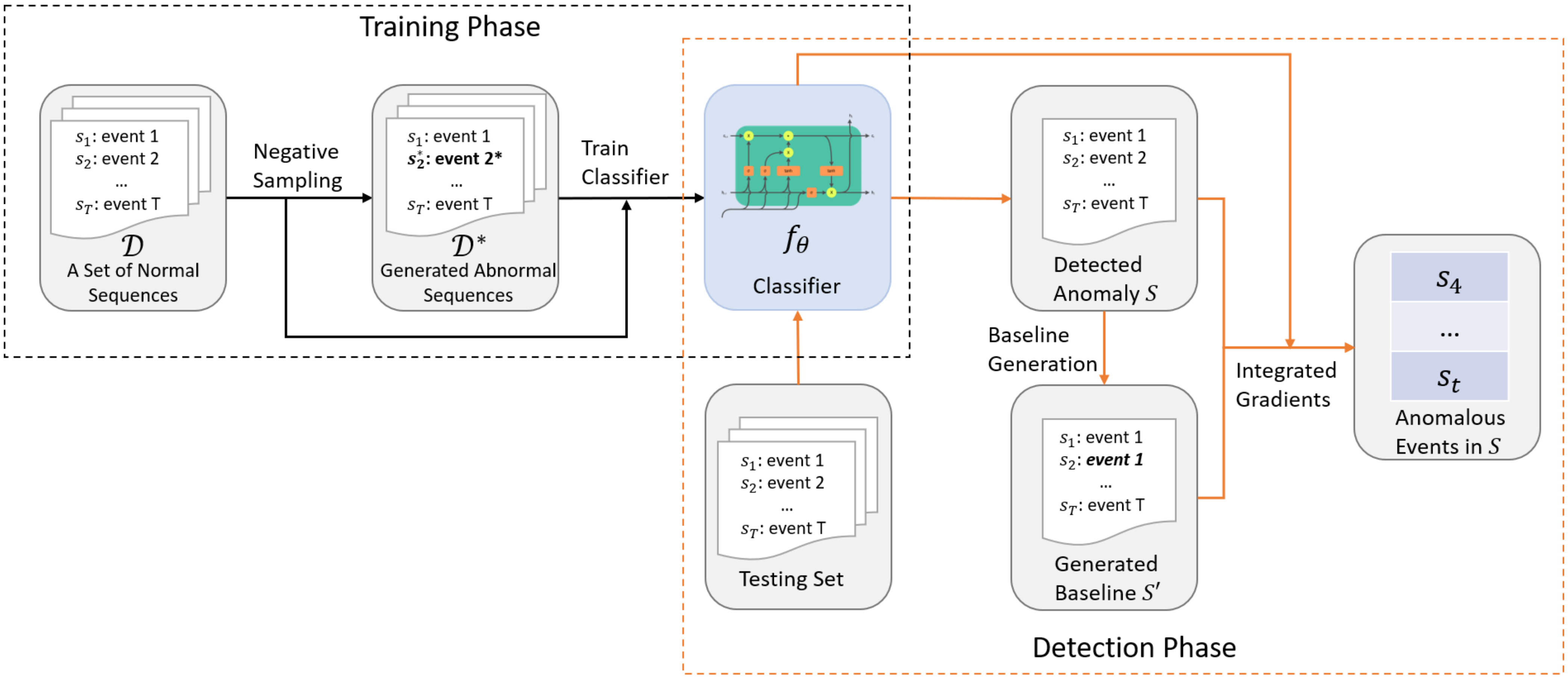
Problem Statement

- Consider a log sequence of discrete events $S = \{s_1, \dots, s_t, \dots, s_T\}$, where $s_t \in \mathcal{E}$ indicates the event at the t-th position, and $\mathcal{E}$ is a set of unique events.
    - ➢Task 1: predicting whether a log sequence S is anomalous based on a training dataset $\mathcal{D} = \{S^i\}_{i=1}^N$ that consists of only normal sequences.
    - ➢Task 2: identifying anomalous events in the sequence

Outline

- **Background**
- **Preliminary**
- **Problem Statement**
- **Method**
  - ➢Data Augmentation via Negative Sampling
  - ➢Anomaly Detection at a Sequence Level
  - ➢Anomalous Event Detection via Integrated Gradients
- **Experiment**
- **Conclusion**

Framework of InterpretableSAD



Data Augmentation via Negative Sampling

- In order to train an accurate binary classifier, we aim to generate a dataset $\mathcal{D}^*$ with sufficient anomalous samples that can cover common anomalous scenarios.
- Two anomalous scenarios for anomalous log sequence generation:
  - ➢Rare events in the sequences
  - ➢Regular events happen in an unusual context

Data Augmentation via Negative Sampling Cont.

Algorithm 1: Negative Sampling

Input : Training set $\mathcal{D}$, Negative sample size M
Output: Negative sample set $\mathcal{D}^*$
Generate a bigram event dictionary $\mathcal{B}$ based on $\mathcal{D}$
for $i = 0$ **to** M **do**
  Randomly select S from $\mathcal{D}$
  $ind \leftarrow$ Randomly select r indices of events from S
  for t in ind **do**
    $(s_t, s^*_{t+1}) \leftarrow$ randomly select or generate a rare or never observed bigram in $\mathcal{B}$
    $(s_t, s_{t+1}) \leftarrow (s_t, s^*_{t+1})$
  $S^* \leftarrow S, \quad \mathcal{D}^* += S^*$
return $\mathcal{D}^*$

Anomaly Detection at a Sequence Level

- After generating a set of anomalous sequences $\mathcal{D}^*$, we use both $\mathcal{D}$ and $\mathcal{D}^*$ to train a binary classification model $f : S \rightarrow [0, 1]$.
- We further adopt the cross-entropy loss to train the neural network:

$$\mathcal{L} = \sum_{j \in \mathcal{D}^* \cup \mathcal{D}} -y_j log \hat{y}_j - \left(1 - y_j\right) \log(1 - \hat{y}_j)$$

Anomalous Event Detection via Integrated Gradients

- Integrated Gradients (IG) is a model interpretable technique that can interpret prediction results by attributing input features.
- Formally, given a neural network $f_\theta : S \rightarrow [0, 1]$, integrated gradients are attributions of the prediction at input S relative to a baseline input S' as a vector $A_{f_\theta}(S, S') = (a_1, \ldots, a_T)$, where a_t is the contribution of s_t to the prediction $f_\theta(S)$.

Anomalous Event Detection via Integrated Gradients Cont.

- Specifically, the integrated gradient for the t-th event for sequence S and the baseline S' is defined as follows:

$$IG_t(S) \equiv (s_t - s_t') \times \int_{\alpha=0}^{1} \frac{\partial f_\theta(S' + \alpha \times (S - S'))}{\partial s_t} d\alpha$$

- Completeness axiom:

$$\sum_{t=1}^{T} A_{f_\theta}(S_t, S_t') = f_\theta(S) - f_\theta(S')$$

IG Baseline Generation

Algorithm 2: Baseline Generation

Input : Neural network f_θ, Anomalous sample S, Training set $\mathcal{D}$, Replacement Threshold τ

Output: Baseline S'

$i = 0$

while $f_\theta(S)$ *is not normal & $i < \tau$* **do**

  $s_t \leftarrow$ Select the event in S with the lowest frequency based on $\mathcal{D}$

  $s_t \leftarrow s_{t-1}$, $i+ = 1$

$S' \leftarrow S$

return S'

Outline

- **Background**
- **Preliminary**
- **Problem Statement**
- **Method**
- **Experiment**
  - ➢Datasets
  - ➢Baselines
  - ➢Experimental Results
- **Conclusion**

Datasets

- Log parser – Drain; Window size – 100; Step size – 20.
- Training dataset consists of 100,000 normal log sequences and 2,000,000 generated anomalous sequences for each log dataset .

TABLE I: Statistics of Test Datasets

Dataset	# of Unique Log Keys	# of Log Sequences		# of Log Keys in Anomalous Sequences	
		Normal	Anomalous	Normal	Anomalous
HDFS	48 (19)	458,223	16,838	N/A	N/A
BGL	396 (318)	19,430	4,190	326,491	7,139
Thunderbird	806 (774)	22,538	76,189	6,866,417	479,883

Baselines for Anomalous Log Sequence Detection

- Traditional machine learning models:
  - ➢Principal Component Analysis (PCA)
  - ➢One-Class SVM (OCSVM)
  - ➢Isolation Forest (iForest)
  - ➢LogCluster
- Deep learning models:
  - ➢DeepLog
  - ➢LogAnomaly

Baselines for Anomalous Event Detection

- Anchors
- Low-Freq
- Integrated Gradients without our IG baseline generation

Results on Anomalous Log Sequence Detection

Method	BGL			Thunderbird			HDFS		
	Precision	Recall	F-1 score	Precision	Recall	F-1 score	Precision	Recall	F-1 score
PCA	67.91	99.79	80.82	94.83	84.43	89.33	97.77	42.12	58.88
iForest	73.13	38.19	50.17	95.06	17.92	30.15	41.59	58.80	48.72
OCSVM	24.60	100	39.49	87.13	100	93.12	6.68	90.58	12.44
LogCluster	8.03	15.97	10.69	86.56	22.94	36.26	98.37	67.45	80.03
DeepLog	42.39	52.08	46.74	82.42	81.36	81.89	56.98	48.37	52.32
LogAnomaly	42.58	53.17	47.29	81.69	82.11	81.90	55.85	48.03	51.65
InterpretableSAD	94.25	88.47	**91.27**	97.31	96.42	**96.86**	92.31	87.04	**89.60**

Results on Anomalous Log Sequence Detection

Method	BGL			Thunderbird			HDFS		
	Precision	Recall	F-1 score	Precision	Recall	F-1 score	Precision	Recall	F-1 score
PCA	67.91	99.79	80.82	94.83	84.43	89.33	97.77	42.12	58.88
iForest	73.13	38.19	50.17	95.06	17.92	30.15	41.59	58.80	48.72
OCSVM	24.60	100	39.49	87.13	100	93.12	6.68	90.58	12.44
LogCluster	8.03	15.97	10.69	86.56	22.94	36.26	98.37	67.45	80.03
DeepLog	42.39	52.08	46.74	82.42	81.36	81.89	56.98	48.37	52.32
LogAnomaly	42.58	53.17	47.29	81.69	82.11	81.90	55.85	48.03	51.65
InterpretableSAD	94.25	88.47	**91.27**	97.31	96.42	**96.86**	92.31	87.04	**89.60**

Results on Anomalous Log Sequence Detection

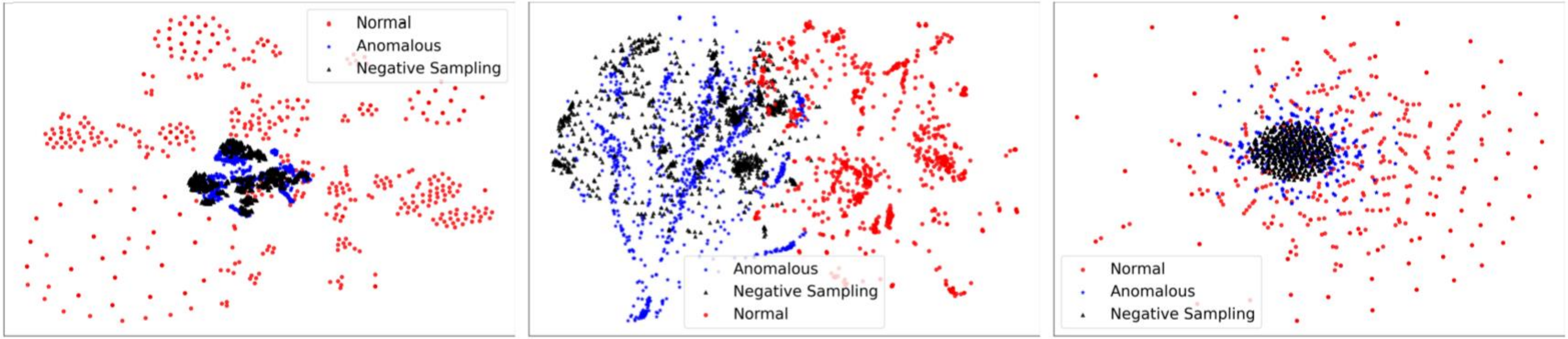
Method	BGL			Thunderbird			HDFS		
	Precision	Recall	F-1 score	Precision	Recall	F-1 score	Precision	Recall	F-1 score
PCA	67.91	99.79	80.82	94.83	84.43	89.33	97.77	42.12	58.88
iForest	73.13	38.19	50.17	95.06	17.92	30.15	41.59	58.80	48.72
OCSVM	24.60	100	39.49	87.13	100	93.12	6.68	90.58	12.44
LogCluster	8.03	15.97	10.69	86.56	22.94	36.26	98.37	67.45	80.03
DeepLog	42.39	52.08	46.74	82.42	81.36	81.89	56.98	48.37	52.32
LogAnomaly	42.58	53.17	47.29	81.69	82.11	81.90	55.85	48.03	51.65
InterpretableSAD	94.25	88.47	**91.27**	97.31	96.42	**96.86**	92.31	87.04	**89.60**

Results on Anomalous Event Detection

We consider two scenarios, with or without a validation set consisting of 10% anomalous sequences in the testing datasets to tune a detection threshold for anomalous event detection.

Method	BGL			Thunderbird		
	Precision	Recall	F-1 score	Precision	Recall	F-1 score
Anchors	0.31	8.56	0.60	4.58	14.62	6.98
Low-Freq	38.76	93.59	54.82	52.61	99.00	68.70
IG w/o val	6.56	90.27	12.23	10.36	85.65	18.49
IG w/ val	42.43	73.83	53.89	20.92	44.48	28.45
InterpretableSAD w/o val	50.87	89.23	64.80	94.98	86.79	90.70
InterpretableSAD w/ val	68.92	82.53	**75.11**	93.84	98.31	**96.02**

Visualization of the normal, anomalous, and generated anomalous sequences



(a) BGL (b) Thunderbird (c) HDFS

Outline

- **Background**
- **Preliminary**
- **Problem Statement**
- **Method**
- **Experiment**
- **Conclusion**

Conclusion

- Leverage the data augmentation strategy to generate anomalous samples by proposing a novel negative sampling algorithm.
- Apply an interpretable machine learning technique, Integrated Gradients (IG), to detect the potential anomalous events.
- Propose a novel feature attribution baseline generation algorithm.
- Experimental results on three log datasets show that our model can achieve state-of-the-art performance on the anomalous sequence and event detection.

Thank You for Your Attention!

This work was supported in part by NSF 1502273 and 2103829.