








On the Influence of Data Resampling for Deep Learning-Based Log Anomaly Detection: Insights and Recommendations

Xiaoxue Ma , Member, IEEE, Huiqi Zou , Pinjia He , Member, IEEE, Jacky Keung , Senior Member, IEEE, Yishu Li , Member, IEEE, Xiao Yu , and Federica Sarro , Member, IEEE

Abstract—Numerous Deep Learning (DL)-based approaches have gained attention in software Log Anomaly Detection (LAD), yet class imbalance in training data remains a challenge, with anomalies often comprising less than 1% of datasets like Thunderbird. Existing DLLAD methods may underperform in severely imbalanced datasets. Although data resampling has proven effective in other software engineering tasks, it has not been explored in LAD. This study aims to fill this gap by providing an in-depth analysis of the impact of diverse data resampling methods on existing DLLAD approaches from two distinct perspectives. Firstly, we assess the performance of these DLLAD approaches across four datasets with different levels of class imbalance, and we explore the impact of resampling ratios of normal to abnormal data on DLLAD approaches. Secondly, we evaluate the effectiveness of the data resampling methods when utilizing optimal resampling ratios of normal to abnormal data. Our findings indicate that oversampling methods generally outperform undersampling and hybrid sampling methods. Data resampling on raw data yields superior results compared to data resampling in the feature space. These improvements are attributed to the increased attention given to important tokens. By exploring the resampling ratio of normal to abnormal data, we suggest generating more data for minority classes through oversampling while removing less data from majority classes through undersampling. In conclusion, our study provides valuable insights into the intricate relationship between data resampling methods and DLLAD. By addressing the challenge of class imbalance, researchers and practitioners can enhance DLLAD performance.

Index Terms—Deep learning-based log anomaly detection, data resampling methods, class imbalance, empirical analysis.

I. INTRODUCTION

SOFTWARE-INTENSIVE systems, which cater to a wide user base [1], are susceptible to minor issues that can lead to adverse consequences such as data corruption and performance degradation [2]. In this context, logs play a crucial role in system maintenance [3], [4], [5], [6], as they capture essential runtime information required for troubleshooting and performance monitoring [7]. Consequently, there is a considerable interest in utilizing logs for anomaly detection. Recently, many Deep Learning-based Log Anomaly Detection (DLLAD) approaches [2], [8], [9], [10], [11], [12], [13] have been proposed to automatically identify system anomalies, showing promising results.

In real-world scenarios in DLLAD, the proportion of normal data greatly outweighs that of abnormal data. For instance, consider the Thunderbird dataset in Table I, one of the commonly used public datasets where logs are grouped into log sequences (with 20, 50, or 100 logs constituting a sequence) for data analysis. In this dataset, anomalies only account for 0.16%–0.35% of the total, highlighting the serious imbalance in the data distribution. Le et al. [1] have revealed that DLLAD models trained on highly imbalanced datasets exhibit low precision or recall values. Low recall leads to missed anomalies, leaving potential threats undetected, while low precision generates numerous false alarms, causing alert fatigue and resource wastage on normal logs [1], [12], [13].

Despite its significance, class imbalance in DLLAD has been largely overlooked. Data resampling offers a potential solution by either generating abnormal data or removing normal data, thereby enabling the model to learn from a more balanced representation of both classes. Previous surveys [14], [15] indicate that no single data resampling method consistently excels across different domains. Given that log data has a unique format, often involving the grouping of log events into log sequences, which distinguishes it from other types of data, this study aims to evaluate the impact of class imbalance on DLLAD performance, identify the most effective resampling methods for DLLAD, and determine the optimal resampling ratio of

Received 1 May 2024; revised 21 October 2024; accepted 29 November 2024. Date of publication 9 December 2024; date of current version 10 January 2025. Recommended for acceptance by N. Tsantalidis. (Corresponding author: Xiao Yu.)

Xiaoxue Ma and Yishu Li are with the Department of Electronic Engineering and Computer Science, Hong Kong Metropolitan University, Hong Kong 999077, China (e-mail: kxma@hkmu.edu.hk; sliy@hkmu.edu.hk).

Huiqi Zou is with the Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: hzou11@jh.edu).

Pinjia He is with the School of Data Science, The Chinese University of Hong Kong, Shenzhen 518172, China (e-mail: hepinjia@cuhk.edu.cn).

Jacky Keung is with the Department of Computer Science, City University of Hong Kong, Hong Kong 999077, China (e-mail: jacky.keung@cityu.edu.hk).

Xiao Yu is with the State Key Laboratory of Blockchain and Data Security, Zhejiang University, Hangzhou 310058, China (e-mail: xiaoyu_cs@hotmail.com).

Federica Sarro is with the Department of Computer Science, University College London, WC1E 6BT London, U.K. (e-mail: f.sarro@ucl.ac.uk).

Digital Object Identifier 10.1109/TSE.2024.3513413

TABLE I

THE STATISTICS OF THE FOUR PUBLIC DATASETS. TB, *ws*, *Msg*, *Seq*, AND *A* ARE THE ABBREVIATIONS OF THE THUNDERBIRD DATASET, WINDOW SIZES, MESSAGES, SEQUENCES, AND ANOMALIES, RESPECTIVELY

Dataset	# of <i>Msg</i>	<i>ws</i>	Training Data		Testing Data	
			# of <i>Seq</i>	# of <i>A</i>	# of <i>Seq</i>	# of <i>A</i>
BGL	4,713,493	20	188,540	17,252	47,134	3,006
		50	75,416	7,415	18,853	1,383
		100	37,708	4,009	9,425	817
TB	5,000,000	20	200,000	328	50,000	37
		50	79,999	195	19,999	29
		100	39,999	138	9,999	23
Spirit	5,000,000	20	200,000	8,817	50,000	290
		50	79,999	4,275	19,999	270
		100	39,999	2,577	9,999	250
Huawei	1,048,575	20	41,943	84	10,486	21
		50	16,777	60	4,195	16
		100	8,388	47	2,098	12

normal to abnormal data. To this end, we conduct an extensive empirical study by employing three oversampling methods (*Random OverSampling* (*ROS*), *SMOTE* [16], and *ADASYN* [17]), three undersampling methods (*Random UnderSampling* (*RUS*), *NearMiss* [18], and *InstanceHardnessThreshold* (*IHT*) [19]), and two hybrid sampling methods (*SMOTEENN* [20] and *SMOTETomek* [20]) on DLLAD approaches (CNN [9], LogRobust [10], NeuralLog [2]) across four publicly available datasets using six evaluation metrics. The results are compared with those obtained without any resampling (*NoSampling*). Furthermore, the data resampling methods can also be categorized into resampling on raw data and resampling in the feature space. It is important to note that many data resampling methods are designed for application only within the feature space, as they rely on distance computations. Simpler methods, like *ROS* and *RUS*, can be applied to both raw data (by duplicating/removing log sequences with identical texts) and feature space (by duplicating/removing sequences with the same embedding vectors). We structure our study with the research questions:

RQ1: Do the existing DLLAD approaches perform well enough with varying degrees of class imbalance? We evaluate the performance of existing DLLAD approaches across datasets with different levels of class imbalance. In addition, we systematically examine how class imbalance impacts these approaches while maintaining data variety. **Findings:** The performance of DLLAD approaches is quite influenced by the degree of class imbalance, with their effectiveness notably decreasing in the presence of more severe data imbalance.

RQ2: How does the resampling ratio of normal to abnormal data affect the performance of DLLAD approaches? We explore how different resampling ratios of normal to abnormal data impact the DLLAD performance by using quarter-based multipliers of the original ratio of normal to abnormal log sequences. **Findings:** The effectiveness of oversampling methods in DLLAD approaches improves when more abnormal log sequences are generated. Conversely, undersampling methods are more effective when fewer normal log sequences are removed. For hybrid sampling methods, no specific resampling ratio consistently improves DLLAD performance.

RQ3: Does data resampling improve the effectiveness of existing DLLAD approaches? We assess the effectiveness of data resampling on DLLAD approaches utilizing an optimal resampling ratio (obtained from RQ2) of normal to abnormal data. **Findings:** Overall, oversampling methods demonstrate superior performance compared to undersampling and hybrid sampling methods. Remarkably, the straightforward methods applied directly to raw data outperform methods applied within the feature space. Surprisingly, in many scenarios, certain undersampling methods (i.e., *NearMiss* and *IHT*), and even a hybrid sampling method *SMOTEENN* aimed at mitigating data imbalance, fail to effectively enhance the performance of DLLAD approaches.

Our study makes the following two main contributions.

- 1) To the best of our knowledge, we undertake the first extensive study aimed at systematically assessing the impact of data resampling methods on model performance in DLLAD. Our study encompasses a total of 5,580 experiments, wherein we employ ten data resampling methods to existing DLLAD approaches and provide a comprehensive evaluation and statistical analysis across four benchmark datasets.
- 2) We present findings and provide implications for researchers and practitioners in log anomaly detection. For example, we recommend utilizing *ROS* on raw data for DLLAD approaches, particularly in datasets with severe class imbalance.

II. BACKGROUND

A. Overview of DLLAD Models

The typical workflow of DLLAD approaches (shown in Fig. 1) consists of four phases: 1) log parsing, 2) log grouping, 3) log embedding, and 4) model training and prediction. To effectively extract valuable information for analysis, previous studies [2], [8], [10], [11], [21] convert the unstructured log messages generated during system operation into structured log events. Each log message comprises a header and content, where the header includes information like timestamps, typically omitted from analysis [2]. The log content is then segmented into constant and variable sections [2]. By replacing variable elements with a special symbol, the original log messages are converted into log events as illustrated in Fig. 1. To group log events into log sequences, we adopt the fixed window strategy used in prior studies [8], [11], [12]. If an abnormal log event is part of a log sequence, the log sequence is labeled as abnormal. Conversely, if the log sequence comprises solely normal log events, it is labeled as normal. These log sequences are subsequently transformed into embedding vectors and used as input for training a classification model to predict whether a log sequence is abnormal.

B. Existing DLLAD Approaches

Recent DL approaches for anomaly detection can be categorized into three main groups (as discussed in Section VI): Convolutional Neural Network (CNN)-based models, Long Short-Term Memory-based models, and Transformer-based models. To select the most suitable DLLAD approaches, we

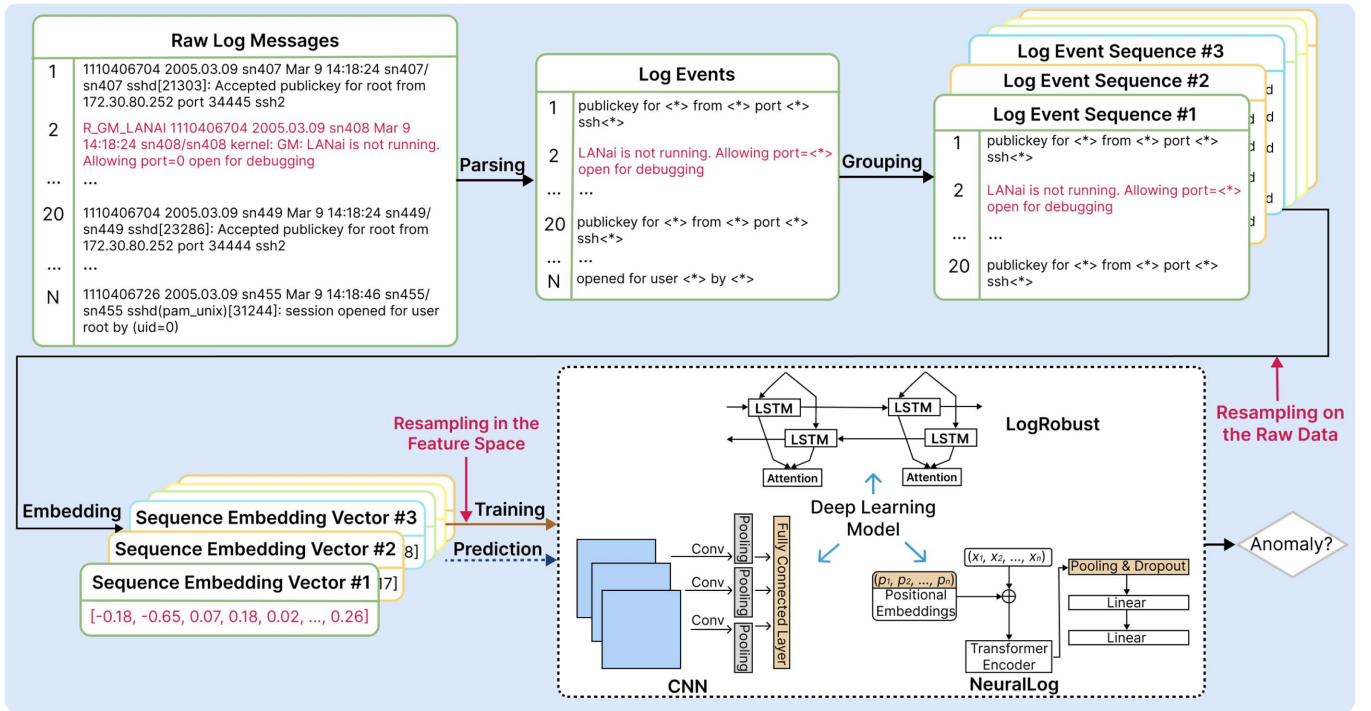


Fig. 1. The common workflow of DLLAD approaches.

take two factors into consideration. Firstly, we look for models that are representative of the DL models described in Section VI. Secondly, we aim to include models that have been recently proposed. Therefore, we choose the following models:

CNN. Lu et al. [9] adopted a CNN-based model to automatically detect log anomalies. Logs were parsed based on log keys, which were then encoded using logkey2vec. These embeddings were structured into a trainable matrix, simplifying neural network training. The model architecture comprised three convolutional layers, a dropout layer, and max-pooling layers.

LogRobust. Zhang et al. [10] employed Drain [22] for log parsing and integrated the FastText [23], a pre-trained Word2vec model, with TF-IDF weights [24] to represent log events as semantic vectors. Subsequently, these vectors are utilized as input to an attention-based Bi-directional LSTM (Bi-LSTM) model for detecting anomalies.

NeuralLog. Le et al. [2] preprocessed log messages without log parsing and encoded them into vector representations via a pre-trained Transformer-based language model BERT [25]. Then, they apply a transformer encoder to classify log sequences, with the primary objective of capturing semantic information comprehensively.

C. Data Resampling

We make use of commonly adopted data resampling methods in the software engineering domain [26], [27], [28], [29]. These methods are classified into three primary categories: oversampling, undersampling, and hybrid sampling. As depicted by

the two red arrows in Fig. 1, data resampling methods can be applied either to raw data (before embedding) or to the feature space (after embedding). Methods like random oversampling and undersampling [27] are applicable to both contexts, while other methods are specific to the feature space. Data resampling methods applied to raw data modify log sequence distribution directly in the training set to address class imbalance, which affects the number of log sequences belonging to each class and the generated embeddings. For example, with a set of ten log sequences (including one abnormal sequence), applying random undersampling to the raw data removes some normal sequences, leaving five normal and one abnormal sequence. Embeddings are then generated from these six resampled sequences. In contrast, data resampling methods applied within the feature space first generate embeddings from the original log sequences and then perform resampling based on these embeddings. For instance, if random undersampling is applied in the feature space, embeddings are created from all ten log sequences before the resampling process occurs. In the following, we provide an overview of each of the data resampling methods compared in our study.

(1) **Random OverSampling (ROS)** randomly replicates log sequences of the minority class without generating new ones. These replicated abnormal sequences are then added to the original dataset. This method, when applied to raw data and feature space, is denoted as ROS_R and ROS_F , respectively.

(2) **Random UnderSampling (RUS)** randomly selects log sequences from the majority class and subsequently removes them from the original dataset. This method, when applied to

raw data and feature space, is denoted as RUS_R and RUS_F , respectively.

(3) Synthetic Minority Oversampling Technique (SMOTE) [16] is an oversampling method applied to the feature space. It augments the minority class by generating synthetic log sequences instead of duplications. This method first randomly selects log sequences from the minority class. For each selected abnormal log sequence M_A , one of its k nearest neighbors M_B are randomly chosen. The embedding vector of the synthetic log sequence M_S is calculated with the formula $x_s = x_A + \text{Random}(0, 1)(x_B - x_A)$, where x_A and x_B represent the embedding vectors of M_A and M_B , separately. These newly generated synthetic abnormal log sequences are subsequently added to the original dataset.

(4) Adaptive Synthetic Sampling Approach [17] (ADASYN) serves as an extension of SMOTE. Unlike SMOTE, ADASYN generates new synthetic abnormal log sequences near the class boundary instead of within the abnormal log sequences themselves.

(5) NearMiss [18] operates as an undersampling method. It calculates the distance between two classes and randomly removes normal log sequences based on the distance. In our evaluation, we adopt NearMiss-3, which has demonstrated superior performance compared to NearMiss-1 and NearMiss-2. Specifically, NearMiss-3 selects a number of the nearest normal log sequences for each abnormal log sequence and removes them from the dataset.

(6) Instance Hardness Threshold (IHT) [19] involves the application of a classifier to the dataset, followed by the removal of log sequences that are hard to classify. The Random Forest (RF) [30] algorithm serves as the default estimator for estimating the Instance Hardness (IH) [19] of individual log sequences.

(7) SMOTEENN [20] is a hybrid sampling method that combines the oversampling method SMOTE and the undersampling method Edited Nearest Neighbour (ENN) [31]. SMOTE generates abnormal log sequences that can sometimes overlap with the majority class, making classification challenging. ENN, acting as a data cleaning method, helps address this issue. It removes log sequences when any or most of its closest neighbors are from a different class.

(8) SMOTETomek [20] is a hybrid sampling method that shares similarities with SMOTEENN. It incorporates Tomek links [32] for data cleaning, defined by the distances between log sequences M_i and M_j from two classes. A pair (M_i, M_j) forms a Tomek link if no log sequence M exists with $d(M_i, M) < d(M_i, M_j)$ or $d(M_j, M) < d(M_i, M_j)$. After oversampling by SMOTE, the log sequences that form Tomek links are then removed to help reduce potential noise or borderline log sequences that may affect classification performance.

III. STUDY DESIGN

A. Datasets

To assess the performance of DLLAD approaches with the ten data resampling methods compared in our study, we use considered four widely used publicly available datasets (namely

HDFS, BGL, Thunderbird, and Spirit) as well as a recently released integrated industrial dataset [1], [2]. After careful analysis, we decided to exclude the HDFS dataset [33] from our evaluation because most existing approaches, like CNN, LogRobust, and NeuralLog, have already achieved near-optimal results on it, with F1 scores over 0.98 as reported in previous studies. The BGL dataset [34] comprises log data from supercomputing system at Lawrence Livermore National Labs. Thunderbird and Spirit datasets [34] are acquired from two real-world supercomputers at Sandia National Labs. In addition, we utilize a combined annotated dataset from recent research [35], which we refer to as Huawei. This dataset primarily consists of data from two distinct industrial cloud services within Huawei Cloud. All these datasets consist of both normal and abnormal log messages, which have been manually identified.

To group log events into a log sequence, a fixed window grouping strategy is commonly used [8], [11], [12], where each sequence contains a fixed number X of log events, with X representing the window size (ws). However, choosing an appropriate ws is challenging. A small ws makes it difficult for log anomaly detection models to capture anomalies that span multiple log sequences [1]. Additionally, employing smaller ws results in more log sequences containing fewer log events, ultimately leading to slower training speed. On the other hand, if ws is large, log sequences may include multiple anomalies and confuse the detection scheme [1], [36]. In the majority of prior research studies [2], [8], [11], [12], [37], a single window size is typically employed to evaluate the proposed approaches, with $ws = 20$ being the most common choice. A few studies [1], [13] have investigated multiple window sizes including 20, 100, and 200. In most cases, the F1 performance is found to be better at $ws = 20$ and 100 compared to $ws = 200$. However, there is no consistent indication of whether a window size of 20 or 100 performs better. He et al. [38] suggest that window size settings can impact the performance of supervised LAD approaches. Our experiment results (shown in Table IV) also emphasize the absence of a universally optimal window size across all DLLAD datasets. For instance, LogRobust exhibits superior F1 and MCC performance on the Thunderbird dataset at $ws = 100$, while achieving better performance on other datasets at $ws = 20$. As a result, in our experiments, we consider both $ws = 20$ and $ws = 100$ as window sizes to account for potential variations in performance. Additionally, we introduce a ws of 50 to provide a balanced perspective between the shorter and longer sequences analyzed. By including this intermediate window size, we aim to uncover a more nuanced understanding of how log sequence length impacts DLLAD performance, and whether the effects of data resampling across datasets with different window sizes are robust.

In Table I, we detail the number of log sequences (# of Sequences) for each dataset across various window sizes, as well as the count of abnormal log sequences (# of Anomalies) within both training and test sets. To assess log data variety, Table II presents the quantitative statistics on the variety of log events and log sequences. The BGL and Spirit datasets contain thousands of unique abnormal log sequences (i.e., log sequences that appear only once in the dataset), whereas the

TABLE II

LOG VARIETY IN THE DATASETS: E_n , AND E_a DENOTE THE NUMBER OF UNIQUE NORMAL AND ABNORMAL LOG EVENTS; $S_{n/a}(20)$, $S_{n/a}(50)$, AND $S_{n/a}(100)$ DENOTE THE NUMBER OF UNIQUE NORMAL/ABNORMAL LOG SEQUENCES GROUPED BY VARYING WINDOW SIZES

Dataset	E_n	E_a	$S_n(20)$	$S_a(20)$	$S_n(50)$	$S_a(50)$	$S_n(100)$	$S_a(100)$
BGL	1,552	53	36,262	3,534	20,074	2,605	10,889	1,901
TB	3,153	8	188,722	241	92,668	194	47,937	155
Spirit	95,673	20	109,817	8,957	79,103	4,495	45,384	2,804
Huawei	90	28	23,635	105	20,354	76	10,424	59

Thunderbird and Huawei datasets exhibit only hundreds. This analysis enables a comprehensive evaluation of DLLAD approaches across datasets with varying levels of abnormal log sequence variety.

In Table III, we report the anomaly proportions (i.e., the proportions of abnormal sequences among all sequences) before and after employing data resampling methods. The anomaly proportions are adjusted according to the resampling ratios of normal to abnormal log sequences, which are obtained by multiplying the original ratio by quarter-based constants. For example, starting with an original ratio of normal to abnormal log sequences being 20:1, we apply quarter-based constants (1/4, 1/2, and 3/4) to derive new ratios of 5:1, 10:1, and 15:1, respectively. Consequently, the desired anomaly proportions are 1/6, 1/11, and 1/16, respectively. The datasets, as detailed in Table III, exhibit very low original anomaly proportions, ranging from 0.16% to 10.63%. Moreover, enlarging the window size has minimal impact on the level of class imbalance across each dataset. For example, in BGL, the anomaly proportion is 9.15% with $ws = 20$ and 10.63% with $ws = 100$. After applying the specified resampling ratios, the anomaly proportions have shown substantial increases, such as from 9.15% to 28.72% (BGL dataset with $ws = 20$), 9.83% to 30.37% (BGL dataset with $ws = 50$), and 10.63% to 33.03% (BGL dataset with $ws = 100$). The hybrid sampling methods *SMOTEENN* and *SMOTETomek* combine oversampling and undersampling to achieve a desired anomaly proportion. *SMOTE* generates synthetic samples for the minority class, while ENN or Tomek Links removes noisy or borderline samples, which may result in slight discrepancies between the final and desired anomaly proportions. The last three columns of Table III show the average anomaly proportions resulting from the application of *SMOTEENN* and *SMOTETomek*.

B. Evaluation

We use four commonly used evaluation metrics Recall, Precision, Specificity, and F1-score in previous DLLAD studies [1], [2], [12], [35]. Given that Matthews Correlation Coefficient (MCC) and Area Under the Curve (AUC) are recommended for evaluating software engineering tasks with class imbalance [39], [40], [41], [42], [43], we include both MCC and AUC in our evaluation to provide a comprehensive assessment of DLLAD model performance. The commonly used four metrics originate from the confusion matrix, which describes four types of instances: TP (True Positives) represents the number of abnormal log sequences correctly predicted as anomalies,

TN (True Negatives) represents the number of normal log sequences correctly predicted as normal, FP (False Positives) represents the number of normal log sequences incorrectly predicted as anomalies, and FN (False Negatives) represents the number of abnormal log sequences incorrectly predicted as normal. The definitions of these metrics are as follows:

(1) $Recall = \frac{TP}{TP+FN}$ represents the proportion of actual anomalies that are correctly predicted by DLLAD models out of all actual anomalies present in the testing dataset. It indicates DLLAD models' ability to capture all abnormal log sequences correctly.

(2) $Precision = \frac{TP}{TP+FP}$ measures the proportion of predicted anomalies by DLLAD models that are actual anomalies out of all anomalies predicted by the models. It indicates the accuracy of the DLLAD models in identifying actual anomalies without falsely labeling normal log sequences as anomalies.

(3) $Specificity = \frac{TN}{TN+FP}$ represents the proportion of actual normal log sequences that are correctly predicted as normal by DLLAD models out of all actual normal log sequences. It indicates the ability of the DLLAD models to correctly identify normal log sequences as normal.

(4) $F1\text{-score} = \frac{2 \times (Recall \times Precision)}{Recall + Precision}$ calculates the harmonic mean of Recall and Precision. It provides a balanced measure between Precision and Recall, giving equal weight to false positives and false negatives.

(5) $MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$ is a fully symmetric metric that takes into account all four values (TP, TN, FP, and FN) in the confusion matrix when calculating the correlation between ground truth and predicted values.

(6) *AUC* is a threshold-independent measure that can be calculated by assessing the area under the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate (Sensitivity) against the false positive rate (1 - Specificity) at various threshold settings. Unlike other metrics such as Precision, Recall, F1-score, and MCC, which depend on the choice of a threshold, AUC evaluates the classifier's performance across all possible threshold values, however this is not applicable in practice [41].

To determine the statistical significance of the observed performance differences among these data resampling methods, we employ the Scott-Knott Effect Size Difference (ESD) test [44] based on the assumptions of Analysis Of Variance (ANOVA). The Scott-Knott ESD test is a multiple comparison approach that leverages hierarchical clustering to partition these data resampling methods into distinct groups, exhibiting statistically significant differences at the predetermined significance level of 0.05 ($\alpha = 0.05$). There are no statistically significant differences between data resampling methods within the same group, but significant differences are observed between data resampling methods located in different groups.

C. Research Questions

RQ1. Do the existing DLLAD approaches perform well enough with varying degrees of class imbalance? In this

TABLE III

THE ANOMALY PROPORTIONS BEFORE AND AFTER OVER-/UNDER-/HYBRID SAMPLING. r DENOTES THE ORIGINAL RATIO OF NORMAL TO ABNORMAL LOG SEQUENCES IN THE TRAINING DATASET, AND QUARTER-BASED CONSTANTS ARE REPRESENTED AS 1/4, 1/2, AND 3/4. THE LAST NINE COLUMNS CORRESPOND TO THE ANOMALY PROPORTIONS AFTER DATA RESAMPLING

Dataset	ws	Original Anomaly Proportion	Anomaly Proportion After Oversampling			Anomaly Proportion After Undersampling			Anomaly Proportion After Hybrid Sampling		
			$r^*1/4$	$r^*1/2$	$r^*3/4$	$r^*1/4$	$r^*1/2$	$r^*3/4$	$r^*1/4$	$r^*1/2$	$r^*3/4$
BGL	20	9.15%	28.72%	16.76%	11.84%	28.72%	16.77%	11.84%	27.97%	16.22%	11.42%
	50	9.83%	30.37%	17.90%	12.69%	30.37%	17.90%	12.69%	30.02%	17.66%	12.52%
	100	10.63%	32.25%	19.22%	13.69%	32.24%	19.22%	13.69%	33.03%	18.02%	11.92%
TB	20	0.16%	0.66%	0.33%	0.22%	0.65%	0.33%	0.22%	0.58%	0.29%	0.20%
	50	0.24%	0.98%	0.49%	0.33%	0.97%	0.49%	0.32%	0.85%	0.43%	0.32%
	100	0.35%	1.38%	0.69%	0.46%	1.37%	0.69%	0.46%	1.36%	0.67%	0.46%
Spirit	20	4.41%	15.58%	8.44%	5.79%	15.57%	8.44%	5.79%	15.61%	8.43%	5.81%
	50	5.34%	18.42%	10.15%	7.00%	18.42%	10.15%	7.00%	17.70%	9.50%	6.81%
	100	6.44%	21.60%	12.11%	8.41%	21.60%	12.11%	8.41%	18.33%	10.04%	6.89%
Huawei	20	0.20%	0.80%	0.40%	0.27%	0.80%	0.40%	0.27%	0.75%	0.40%	0.28%
	50	0.36%	1.42%	0.71%	0.48%	1.42%	0.71%	0.48%	1.44%	0.66%	0.45%
	100	0.56%	2.20%	1.11%	0.75%	2.20%	1.11%	0.75%	2.15%	1.10%	0.73%

RQ, we aim to assess the impact of class imbalance on the effectiveness of existing DLLAD approaches. First, we analyze the performance of DLLAD approaches across four datasets with varying window sizes, each exhibiting different levels of class imbalance. Next, we focus on the most balanced dataset, BGL (as detailed in Table III). To systematically investigate the impact of class imbalance, we progressively remove duplicate abnormal log sequences until only unique abnormal sequences remain. This method preserves the variety of the log data, enabling us to assess how varying levels of class imbalance influence DLLAD performance.

RQ2. How does the resampling ratio of normal to abnormal data affect the performance of DLLAD approaches? Due to the varying levels of class imbalance in the different datasets, it is challenging to establish a fixed resampling ratio of normal to abnormal data, such as maintaining a consistent 10:1 ratio of normal to abnormal log sequences across all datasets. Conducting an exhaustive exploration of countless potential ratios to identify an optimal resampling ratio for each dataset is not practically feasible. Considering the substantial data imbalance, we avoid pursuing a 1:1 ratio of normal to abnormal data during the data resampling process. This is particularly evident in the Thunderbird dataset, where the training set has a mere 0.16% anomaly rate with a window size of 20. Applying a 1:1 ratio in such cases would result in excessive data removal during undersampling, leading to a considerable loss of information. Thus, this resampling ratio is not considered. Instead, we adopt the quarter as a foundational unit for our empirical investigations in a flexible and adaptive manner, as shown in Table III.

RQ3. Does data resampling improve the effectiveness of existing DLLAD approaches? In this RQ, we use the recommended resampling ratios of normal to abnormal data, as identified in RQ2, for the different categories of data resampling methods—oversampling, undersampling, and hybrid sampling. Subsequently, we assess the effectiveness of these resampling methods when applied to existing DLLAD approaches.

D. Implementations

We implement the existing approaches introduced in Section II-B using their respective GitHub repositories or reproduced codebases. Following previous works [1], [2], in our dataset setup, the training set comprises the first 80% of raw logs, while the remaining 20% is allocated for testing. Data resampling is applied exclusively to the training data. To reduce computational complexity and memory demands during data resampling operations, for NeuralLog [2], we reduce the embedding dimension from 768 to 256. For CNN [9] and LogRobust [10], we utilize the implementations [1], adhering to the instructions provided by the authors. The implementation of data resampling methods is carried out using the Python toolbox [45] Imbalanced-learn¹. Our experiments encompass 36 distinct instances, resulting from the combination of 3 DLLAD approaches, 4 datasets, and 3 window sizes. For each experimental instance, we investigate 10 data resampling methods with 3 different resampling ratios of normal to abnormal data and *NoSampling*. As a result, we have a total of $3 \times 4 \times 3 \times (10 \times 3 + 1)$ combinations, summing up to 1,116 unique scenarios. To mitigate the variations in performance across different runs, we perform five runs for each data resampling method (including *NoSampling*), culminating in a total of 5,580 experiments conducted in this study. These five-run results are utilized for statistical significance analysis using the Scott-Knott ESD test, which calculates the group ranking of each data resampling method across different datasets. Furthermore, the averages of the five-run results are provided in Tables IV–VII in Section IV. We run our experiments on a Linux server with an Intel Xeon Silver 4210 CPU and four Nvidia GeForce RTX 3090-Ti GPUs.

IV. RESULTS AND ANALYSIS

A. RQ1. Do the Existing DLLAD Approaches Perform Well Enough With Varying Degrees of Class Imbalance?

Table IV summarizes the performance of CNN, LogRobust, and NeuralLog across the BGL, Thunderbird, Spirit, and

¹ <https://imbalanced-learn.org/stable/references/index.html>

TABLE IV
THE RECALL, PRECISION, SPECIFICITY, F1-SCORE, MCC, AND AUC VALUES OF THE THREE DLLAD APPROACHES ACROSS THE FOUR DATASETS, EACH WITH THREE DIFFERENT WINDOW SIZES (VARYING ANOMALY PROPORTIONS). BOLD FONT HIGHLIGHTS THE BEST PERFORMANCE AMONG THE THREE WINDOW SIZES

Model	Metric	BGL			TB			Spirit			Huawei		
		$ws = 20$ (9.15%)	$ws = 50$ (9.83%)	$ws = 100$ (10.63%)	$ws = 20$ (0.16%)	$ws = 50$ (0.24%)	$ws = 100$ (0.35%)	$ws = 20$ (4.41%)	$ws = 50$ (5.34%)	$ws = 100$ (6.44%)	$ws = 20$ (0.20%)	$ws = 50$ (0.36%)	$ws = 100$ (0.56%)
CNN	R	0.948	0.907	0.916	0.495	0.460	0.304	0.815	0.893	0.909	0.476	0.688	0.333
	P	0.985	0.708	0.362	0.443	0.169	0.213	0.797	0.582	0.726	1.000	1.000	0.800
	S	0.999	0.965	0.827	0.999	0.997	0.997	0.999	0.991	0.991	1.000	1.000	1.000
	F1	0.966	0.787	0.510	0.441	0.247	0.241	0.806	0.704	0.807	0.645	0.815	0.471
	MCC	0.964	0.779	0.509	0.454	0.277	0.247	0.805	0.716	0.807	0.690	0.829	0.515
	AUC	0.977	0.919	0.903	0.866	0.812	0.786	0.967	0.996	0.995	0.988	0.997	0.928
LogRobust	R	0.949	0.911	0.903	0.766	0.609	0.449	0.892	0.910	0.918	0.667	0.625	0.250
	P	0.860	0.709	0.793	0.084	0.317	0.277	0.973	0.805	0.863	1.000	1.000	0.750
	S	0.989	0.968	0.978	0.994	0.998	0.997	1.000	0.997	0.994	1.000	1.000	1.000
	F1	0.903	0.792	0.844	0.151	0.415	0.340	0.930	0.832	0.889	0.800	0.769	0.375
	MCC	0.897	0.783	0.831	0.252	0.437	0.350	0.931	0.854	0.885	0.816	0.790	0.431
	AUC	0.970	0.960	0.954	0.915	0.861	0.826	0.991	0.990	0.997	0.995	0.994	0.890
NeuralLog	R	0.896	0.627	0.598	0.772	0.730	0.756	0.899	0.931	0.938	0.238	0.095	0.150
	P	0.852	0.872	0.671	0.758	0.469	0.470	0.899	0.864	0.800	1.000	1.000	1.000
	S	0.989	0.991	0.878	1.000	0.999	0.999	0.999	0.999	0.999	1.000	1.000	1.000
	F1	0.872	0.721	0.496	0.760	0.571	0.579	0.895	0.896	0.862	0.385	0.174	0.261
	MCC	0.864	0.718	0.511	0.762	0.585	0.596	0.897	0.896	0.865	0.488	0.308	0.387
	AUC	0.943	0.809	0.738	0.856	0.865	0.828	0.949	0.965	0.964	0.619	0.548	0.575

Huawei datasets using three window sizes. When analyzing the impact of window size, we observe that a window size of 20 typically yields the best results, as highlighted in bold: 8/12 cases for recall, 10/12 for precision, 10/12 for specificity, 8/12 for F1, 9/12 for MCC, and 7/12 for AUC. Conversely, a window size of 100 achieves the best performance in only 1-3 out of 12 cases for these metrics. This suggests that smaller window sizes generally lead to better outcomes, likely because they enable DLLAD models to focus on a more concise and relevant set of log events, capturing anomaly features more effectively. In comparison, larger window sizes may introduce too many log events, making it more difficult to identify key features, which hinders the models' ability to distinguish between abnormal and normal behaviors in log sequences.

Finding 1: DLLAD approaches generally perform better when log sequences have fewer log events, i.e., smaller window sizes.

Regarding the severity of class imbalance (Table III), it follows the order: Thunderbird > Huawei >> Spirit > BGL. In terms of the variety of abnormal sequences (Table II), the order is: Spirit > BGL >> Thunderbird > Huawei. In other words, BGL and Spirit datasets are more balanced (anomaly proportion > 4.4%) and exhibit greater anomaly variety (thousands of unique abnormal log sequences), while Thunderbird and Huawei datasets suffer from severe class imbalance (anomaly proportion < 0.6%) and have less anomaly variety (up to hundreds of unique abnormal log sequences). On the Thunderbird dataset, all three DLLAD approaches, particularly CNN and LogRobust, show poor performance, with F1 and MCC values below 0.5. Although NeuralLog achieves a better F1 score of 0.76 at $ws = 20$, its overall performance remains inadequate, with F1 and MCC values below 0.6 at $ws = 50$ and 100. A similar unsatisfactory performance of DLLAD approaches is observed in the Huawei dataset. Despite achieving nearly 100% precision, all approaches display

low recall, indicating that while most detected anomalies are true positives, many remain undetected. Notably, CNN and LogRobust outperform the more complex NeuralLog model across different window sizes. A potential reason for this is that NeuralLog's Transformer-based architecture is optimized for handling large, diverse datasets and capturing complex dependencies. However, in the Huawei dataset, which has the smallest size and least log data variety, Transformers struggle to generalize effectively, resulting in lower performance. In contrast, CNN and LogRobust are better suited for smaller datasets with less log data variety [46]. For the BGL and Spirit datasets, all three DLLAD approaches demonstrate strong performance, with F1, MCC, and AUC scores exceeding 0.80 at $ws = 20$. This suggests that DLLAD approaches generally achieve better results when applied to more balanced datasets with greater anomaly variety.

To examine whether class imbalance impacts DLLAD performance while maintaining log data variety, we systematically remove varying proportions of duplicate abnormal log sequences from the most balanced BGL dataset with $ws = 50$ and 100. We examine proportions of 0, 1/4, 1/2, 3/4, and all duplicates (retaining only unique sequences). For example, if there are three distinct abnormal log sequences, each with five identical copies (totaling fifteen sequences), removing all duplicates results in three unique abnormal sequences. This method allows us to analyze how different levels of class imbalance affect DLLAD performance while maintaining log data variety. Since NeuralLog does not perform log parsing, we focus on evaluating the other two approaches: CNN and LogRobust. With adjustments to class imbalance through the removal of abnormal sequences, the anomaly proportions in the BGL dataset decrease from 9.2% to 1.3% with $ws = 50$, and from 9.8% to 2.7% with $ws = 100$. Fig. 2 illustrates the performance trends for each evaluation metric corresponding to these class imbalance changes. For CNN, as we remove increasing proportions of duplicate abnormal sequences on BGL with $ws = 20$, the performance consistently declines, with decreases

TABLE V
THE RECALL, PRECISION, SPECIFICITY, F1-SCORE, MCC, AND AUC VALUES OF CNN WHEN EMPLOYING VARIOUS DATA RESAMPLING METHODS (I.E., NOSAMPLING (NS), SMOTE (SMO), ADASYN (ADA), NEARMISS (NM), INSTANCEHARDNESSTHRESHOLD (IHT), SMOTEENN (SE), SMOTETOMEK (ST), RANDOMOVERSAMPLING IN THE FEATURE SPACE (ROS_F), RANDOMUNDER_SAMPLING IN THE FEATURE SPACE (RUS_F), RANDOMOVERSAMPLING APPLIED TO RAW DATA (ROS_R), AND RANDOMUNDER_SAMPLING APPLIED TO RAW DATA (RUS_R)) TO THE FOUR DATASETS, EACH WITH THREE DIFFERENT WINDOW SIZES. DARKER CELLS SIGNIFY SUPERIOR PERFORMANCE, WHILE VARIOUS COLORS DENOTE STATISTICAL SIGNIFICANCE AMONG DATA RESAMPLING METHODS FOR EACH EVALUATION METRIC (DETERMINED BY THE SCOTT-KNOTT TEST WITH A p -VALUE<0.05), AS OBSERVED IN THE SUBSEQUENT TABLES

Dataset	ws	Metric	NS	SMO	ADA	NM	IHT	SE	ST	ROS _F	RUS _F	ROS _R	RUS _R
BGL	20	R	0.948	0.963	0.962	0.956	0.967	0.961	0.962	0.962	0.960	0.961	0.961
		P	0.985	0.978	0.982	0.989	0.945	0.995	0.989	0.984	0.986	0.984	0.986
		S	0.999	0.999	0.999	0.999	0.996	1.000	0.999	0.999	0.999	0.992	0.999
		F1	0.966	0.970	0.972	0.972	0.956	0.977	0.975	0.973	0.973	0.972	0.974
		MCC	0.964	0.968	0.970	0.970	0.953	0.976	0.973	0.971	0.971	0.972	0.972
		AUC	0.977	0.982	0.978	0.978	0.974	0.977	0.979	0.984	0.974	0.984	0.983
	50	R	0.907	0.910	0.912	0.898	0.944	0.896	0.915	0.931	0.915	0.918	0.899
		P	0.708	0.853	0.877	0.895	0.192	0.882	0.761	0.538	0.205	0.885	0.885
		S	0.965	0.987	0.990	0.991	0.684	0.990	0.971	0.916	0.717	0.990	0.991
		F1	0.787	0.880	0.894	0.896	0.319	0.888	0.820	0.665	0.335	0.901	0.892
		MCC	0.779	0.871	0.886	0.888	0.341	0.880	0.814	0.665	0.352	0.893	0.883
		AUC	0.919	0.941	0.959	0.952	0.824	0.937	0.943	0.929	0.857	0.944	0.951
	100	R	0.916	0.930	0.900	0.911	0.937	0.894	0.902	0.927	0.901	0.907	0.898
		P	0.362	0.857	0.832	0.923	0.218	0.885	0.665	0.905	0.798	0.928	0.613
		S	0.827	0.996	0.983	0.993	0.681	0.989	0.920	0.991	0.977	0.993	0.945
		F1	0.510	0.891	0.865	0.917	0.353	0.888	0.729	0.916	0.844	0.917	0.727
		MCC	0.509	0.889	0.852	0.909	0.359	0.878	0.722	0.908	0.831	0.910	0.713
		AUC	0.903	0.962	0.928	0.958	0.837	0.953	0.912	0.954	0.929	0.953	0.907
TB	20	R	0.495	0.495	0.532	0.532	0.550	0.000	0.495	0.568	0.541	0.550	0.495
		P	0.443	0.821	0.426	0.123	0.185	1.000	0.821	0.154	0.230	0.825	0.817
		S	0.999	1.000	0.999	0.997	0.998	1.000	1.000	0.998	0.998	1.000	1.000
		F1	0.441	0.618	0.443	0.198	0.274	0.000	0.618	0.241	0.315	0.659	0.613
		MCC	0.454	0.637	0.460	0.253	0.316	0.000	0.638	0.293	0.347	0.673	0.634
		AUC	0.866	0.840	0.873	0.880	0.894	0.545	0.839	0.894	0.865	0.881	0.843
	50	R	0.460	0.379	0.414	0.000	0.391	0.000	0.345	0.402	0.000	0.345	0.000
		P	0.169	0.800	0.315	0.000	0.255	0.000	0.537	0.481	0.000	0.923	0.000
		S	0.997	1.000	0.999	1.000	0.998	1.000	0.999	0.999	1.000	1.000	1.000
		F1	0.247	0.490	0.355	0.000	0.299	0.000	0.383	0.376	0.000	0.501	0.000
		MCC	0.277	0.536	0.359	0.000	0.309	0.000	0.410	0.406	0.000	0.563	0.000
		AUC	0.812	0.799	0.826	0.713	0.816	0.638	0.793	0.810	0.725	0.841	0.738
	100	R	0.304	0.290	0.348	0.043	0.261	0.000	0.319	0.348	0.000	0.348	0.000
		P	0.213	0.326	0.447	0.333	0.092	0.000	0.465	0.463	0.000	0.600	0.000
		S	0.997	0.998	0.999	1.000	0.996	1.000	0.998	0.999	1.000	0.999	1.000
		F1	0.241	0.276	0.391	0.077	0.136	0.000	0.334	0.397	0.000	0.420	0.000
		MCC	0.247	0.290	0.393	0.120	0.153	0.000	0.360	0.400	0.000	0.444	0.000
		AUC	0.786	0.788	0.776	0.736	0.786	0.656	0.788	0.801	0.733	0.802	0.730
Spirit	20	R	0.815	0.822	0.814	0.822	0.851	0.614	0.823	0.826	0.808	0.823	0.800
		P	0.797	0.922	0.864	0.797	0.482	0.845	0.934	0.975	0.975	0.991	0.946
		S	0.999	1.000	0.999	0.999	0.995	0.999	1.000	1.000	1.000	1.000	1.000
		F1	0.806	0.868	0.837	0.806	0.615	0.646	0.875	0.894	0.884	0.899	0.866
		MCC	0.805	0.869	0.837	0.807	0.638	0.681	0.876	0.897	0.887	0.902	0.869
		AUC	0.967	0.969	0.968	0.968	0.974	0.963	0.969	0.971	0.973	0.972	0.971
	50	R	0.893	0.907	0.888	0.908	0.963	0.065	0.893	0.935	0.903	0.908	0.889
		P	0.582	0.677	0.769	0.415	0.243	0.080	0.592	0.873	0.510	0.786	0.618
		S	0.991	0.994	0.996	0.983	0.959	0.990	0.992	0.998	0.988	0.997	0.992
		F1	0.704	0.774	0.823	0.570	0.388	0.071	0.712	0.903	0.652	0.842	0.728
		MCC	0.716	0.780	0.823	0.607	0.472	0.060	0.723	0.902	0.673	0.842	0.737
		AUC	0.996	0.997	0.997	0.993	0.978	0.921	0.996	0.998	0.995	0.998	0.997
	100	R	0.909	0.919	0.919	0.899	0.961	0.888	0.908	0.931	0.907	0.925	0.924
		P	0.726	0.853	0.771	0.510	0.311	0.623	0.657	0.904	0.784	0.880	0.803
		S	0.991	0.996	0.993	0.977	0.945	0.986	0.988	0.997	0.994	0.997	0.994
		F1	0.807	0.882	0.838	0.649	0.470	0.730	0.762	0.917	0.841	0.902	0.858
		MCC	0.807	0.881	0.837	0.666	0.530	0.735	0.765	0.915	0.839	0.899	0.857
		AUC	0.995	0.998	0.998	0.992	0.976	0.994	0.997	0.999	0.988	0.998	0.988
Huawei	20	R	0.476	0.571	0.619	0.524	0.524	0.048	0.571	0.524	0.524	0.571	0.619
		P	1.000	0.857	0.867	0.917	0.611	0.333	0.750	1.000	0.917	1.000	0.867
		S	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000	1.000
		F1	0.645	0.686	0.722	0.667	0.564	0.083	0.649	0.688	0.667	0.727	0.722
		MCC	0.690	0.699	0.732	0.693	0.565	0.125	0.654	0.723	0.693	0.756	0.732
		AUC	0.988	0.990	0.995	0.989	0.977	0.888	0.989	0.983	0.983	0.989	0.995
	50	R	0.688	0.688	0.750	0.750	0.875	0.000	0.625	0.813	0.563	0.813	0.813
		P	1.000	1.000	1.000	1.000	0.286	0.000	1.000	1.000	1.000	1.000	1.000
		S	1.000	1.000	1.000	1.000	0.992	1.000	1.000	1.000	1.000	1.000	1.000
		F1	0.815	0.815	0.857	0.857	0.431	0.000	0.769	0.897	0.720	0.897	0.897
		MCC	0.829	0.829	0.866	0.866	0.497	0.000	0.790	0.901	0.749	0.901	0.901
		AUC	0.997	0.997	0.999	0.999	0.995	0.931	0.998	0.997	0.998	0.999	0.998
	100	R	0.333	0.333	0.417	0.417	0.250	0.000	0.333	0.417	0.500	0.417	0.417
		P	0.800	1.000	0.833	0.714	0.750	0.000	0.667	0.625	0.667	1.000	0.625
		S	1.000	1.000	1.000	0.999	1.000	1.000	0.999	0.999	0.999	1.000	0.999
		F1	0.471	0.500	0.556	0.526	0.375	0.000	0.444	0.500	0.571	0.588	0.500
		MCC	0.515	0.576	0.588	0.544	0.431	0.000	0.469	0.508	0.575	0.644	0.508
		AUC	0.928	0.923	0.952	0.953	0.950	0.847	0.940	0.870	0.922	0.944	0.933

TABLE VI
THE RECALL, PRECISION, SPECIFICITY, F1-SCORE, MCC, AND AUC VALUES OF **LOGROBUST** WHEN EMPLOYING VARIOUS DATA RESAMPLING METHODS TO THE FOUR DATASETS, EACH WITH THREE DIFFERENT WINDOW SIZES

Dataset	ws	Metric	NS	SMO	ADA	NM	IHT	SE	ST	ROS_F	RUS_F	ROS_R	RUS_R
BGL	20	R	0.949	0.958	0.955	0.951	0.952	0.952	0.956	0.938	0.942	0.951	0.948
		P	0.860	0.885	0.889	0.882	0.858	0.887	0.889	0.892	0.892	0.912	0.904
		S	0.989	0.992	0.992	0.991	0.989	0.992	0.992	0.992	0.992	0.994	0.993
		F1	0.903	0.920	0.921	0.915	0.903	0.918	0.921	0.914	0.916	0.931	0.925
		MCC	0.897	0.915	0.916	0.910	0.897	0.913	0.916	0.909	0.911	0.926	0.920
		AUC	0.970	0.985	0.988	0.980	0.974	0.983	0.982	0.974	0.975	0.980	0.982
	50	R	0.911	0.901	0.920	0.871	0.940	0.944	0.904	0.919	0.901	0.901	0.898
		P	0.709	0.895	0.879	0.648	0.565	0.821	0.823	0.891	0.872	0.902	0.896
		S	0.968	0.992	0.990	0.953	0.942	0.983	0.983	0.991	0.989	0.992	0.992
		F1	0.792	0.898	0.899	0.726	0.706	0.877	0.859	0.905	0.886	0.901	0.897
		MCC	0.783	0.890	0.891	0.718	0.703	0.869	0.849	0.897	0.877	0.893	0.889
		AUC	0.960	0.966	0.966	0.964	0.962	0.960	0.942	0.966	0.965	0.963	0.966
	100	R	0.903	0.928	0.878	0.917	0.945	0.887	0.908	0.889	0.900	0.936	0.887
		P	0.793	0.892	0.908	0.870	0.803	0.896	0.957	0.914	0.900	0.894	0.896
		S	0.978	0.989	0.992	0.987	0.978	0.990	0.996	0.992	0.990	0.989	0.990
		F1	0.844	0.909	0.892	0.893	0.868	0.892	0.932	0.901	0.900	0.915	0.892
		MCC	0.831	0.901	0.882	0.883	0.858	0.882	0.926	0.892	0.890	0.907	0.882
		AUC	0.954	0.963	0.961	0.959	0.961	0.961	0.971	0.967	0.964	0.966	0.961
TB	20	R	0.766	0.459	0.641	0.775	0.811	0.000	0.775	0.802	0.703	0.652	0.529
		P	0.084	0.708	0.218	0.108	0.127	0.000	0.111	0.141	0.313	0.285	0.352
		S	0.994	1.000	0.998	0.995	0.996	1.000	0.995	0.996	0.997	0.998	0.998
		F1	0.151	0.532	0.313	0.189	0.220	0.000	0.193	0.239	0.368	0.376	0.371
		MCC	0.252	0.557	0.364	0.286	0.319	0.000	0.290	0.335	0.425	0.415	0.403
		AUC	0.915	0.849	0.903	0.920	0.933	0.544	0.916	0.922	0.905	0.685	0.870
	50	R	0.609	0.529	0.598	0.347	0.644	0.000	0.540	0.552	0.425	0.598	0.425
		P	0.317	0.804	0.433	0.647	0.236	0.000	0.271	0.404	0.636	0.852	0.805
		S	0.998	1.000	0.999	1.000	0.996	1.000	0.998	0.998	1.000	1.000	1.000
		F1	0.415	0.637	0.481	0.450	0.333	0.000	0.355	0.444	0.497	0.703	0.554
		MCC	0.437	0.651	0.497	0.473	0.379	0.000	0.378	0.459	0.513	0.713	0.583
		AUC	0.861	0.881	0.864	0.847	0.885	0.524	0.872	0.854	0.860	0.870	0.865
	100	R	0.449	0.464	0.435	0.536	0.565	0.029	0.485	0.435	0.507	0.464	0.420
		P	0.277	0.436	0.510	0.216	0.146	0.667	0.374	0.597	0.469	0.655	0.488
		S	0.997	0.998	0.999	0.995	0.992	1.000	0.998	0.999	0.999	0.999	0.999
		F1	0.340	0.429	0.457	0.297	0.232	0.056	0.418	0.501	0.484	0.526	0.450
		MCC	0.350	0.438	0.463	0.331	0.284	0.139	0.422	0.507	0.485	0.541	0.451
		AUC	0.826	0.868	0.846	0.851	0.848	0.727	0.859	0.865	0.864	0.831	0.837
Spirit	20	R	0.892	0.910	0.914	0.910	0.937	0.791	0.913	0.913	0.915	0.918	0.915
		P	0.973	0.991	0.990	0.836	0.510	0.547	0.975	0.993	0.996	0.990	0.989
		S	1.000	1.000	1.000	0.999	0.995	0.992	1.000	1.000	1.000	1.000	1.000
		F1	0.930	0.949	0.950	0.862	0.660	0.542	0.943	0.951	0.954	0.953	0.950
		MCC	0.931	0.950	0.951	0.867	0.689	0.599	0.943	0.951	0.954	0.953	0.951
		AUC	0.991	0.990	0.993	0.988	0.991	0.980	0.992	0.993	0.994	0.993	0.992
	50	R	0.910	0.903	0.914	0.896	0.974	0.899	0.903	0.907	0.910	0.925	0.907
		P	0.805	0.949	0.763	0.736	0.218	0.717	0.893	0.957	0.884	0.905	0.917
		S	0.997	0.999	0.996	0.996	0.952	0.995	0.999	0.999	0.998	0.999	0.999
		F1	0.832	0.925	0.832	0.808	0.356	0.798	0.898	0.931	0.897	0.915	0.912
		MCC	0.854	0.925	0.833	0.809	0.449	0.800	0.897	0.930	0.896	0.914	0.911
		AUC	0.990	0.990	0.990	0.989	0.967	0.989	0.990	0.990	0.990	0.991	0.990
	100	R	0.918	0.923	0.920	0.925	0.981	0.765	0.913	0.939	0.923	0.909	0.916
		P	0.863	0.945	0.904	0.609	0.330	0.725	0.644	0.972	0.895	0.983	0.950
		S	0.994	0.999	0.996	0.984	0.949	0.991	0.987	0.999	0.997	1.000	0.999
		F1	0.889	0.933	0.911	0.733	0.493	0.717	0.753	0.955	0.908	0.945	0.932
		MCC	0.885	0.932	0.908	0.742	0.553	0.725	0.759	0.954	0.906	0.944	0.931
		AUC	0.996	0.999	0.997	0.993	0.982	0.994	0.995	0.998	0.997	0.997	0.996
Huawei	20	R	0.667	0.714	0.762	0.619	0.619	0.190	0.762	0.762	0.762	0.714	0.619
		P	1.000	1.000	1.000	0.813	0.684	1.000	0.941	1.000	0.941	1.000	0.929
		S	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000	1.000
		F1	0.800	0.833	0.865	0.703	0.650	0.320	0.842	0.865	0.842	0.833	0.743
		MCC	0.816	0.845	0.873	0.709	0.650	0.436	0.847	0.873	0.847	0.845	0.758
		AUC	0.995	0.990	0.991	0.984	0.973	0.986	0.995	0.995	0.995	0.995	0.989
	50	R	0.625	0.875	0.688	0.625	0.813	0.000	0.813	0.688	0.563	0.750	0.563
		P	1.000	1.000	1.000	0.769	0.176	0.000	0.929	1.000	1.000	1.000	1.000
		S	1.000	1.000	1.000	0.999	0.985	1.000	1.000	1.000	1.000	1.000	1.000
		F1	0.769	0.933	0.815	0.690	0.289	0.000	0.867	0.815	0.720	0.857	0.720
		MCC	0.790	0.935	0.829	0.692	0.374	0.000	0.868	0.829	0.749	0.866	0.749
		AUC	0.994	1.000	0.996	0.996	0.992	0.949	0.998	0.996	0.995	0.998	0.995
	100	R	0.250	0.417	0.417	0.083	0.750	0.000	0.250	0.500	0.250	0.417	0.250
		P	0.750	1.000	0.833	1.000	0.113	0.000	0.545	0.375	1.000	1.000	0.750
		S	1.000	1.000	1.000	1.000	0.966	1.000	1.000	0.998	0.998	1.000	1.000
		F1	0.375	0.588	0.556	0.154	0.196	0.000	0.400	0.522	0.300	0.588	0.375
		MCC	0.431	0.644	0.588	0.288	0.282	0.000	0.499	0.520	0.303	0.644	0.431
		AUC	0.890	0.909	0.964	0.955	0.915	0.632	0.929	0.942	0.877	0.974	0.929

of 13.9%, 1.7%, 0.1%, 8.3%, 8.5%, and 0.2% in terms of recall, precision, specificity, F1, MCC, and AUC, respectively. Similarly, on BGL with $ws = 50$, the reductions are more pronounced: 6.3%, 72.1%, 24.9%, 59.3%, 58.5%, and 3.9% for the same metrics. For LogRobust, the decreases are 2.1%,

5.1%, 0.4%, 3.7%, 3.9%, and 1.8% for $ws = 20$, and 8.7%, 8.6%, 0.7%, 8.6%, 9.7%, and 3.2% for $ws = 50$. For CNN, increasing the removal of duplicate abnormal sequences from the BGL dataset results in a decline across performance metrics. With $ws = 20$, recall decreases by 13.9%, precision by 1.7%,

TABLE VII
THE RECALL, PRECISION, SPECIFICITY, F1-SCORE, MCC, AND AUC VALUES OF **NEURALLOG** WHEN EMPLOYING VARIOUS DATA RESAMPLING METHODS TO THE FOUR DATASETS, EACH WITH THREE DIFFERENT WINDOW SIZES

Dataset	ws	Metric	NS	SMO	ADA	NM	IHT	SE	ST	ROS_F	RUS_F	ROS_R	RUS_R
BGL	20	R	0.896	0.951	0.929	0.842	0.766	0.867	0.875	0.939	0.897	0.937	0.925
		P	0.852	0.932	0.908	0.166	0.818	0.908	0.925	0.894	0.958	0.923	0.910
		S	0.989	0.995	0.994	0.691	0.987	0.994	0.995	0.992	0.997	0.995	0.994
		F1	0.872	0.941	0.918	0.275	0.788	0.886	0.899	0.915	0.926	0.930	0.917
		MCC	0.864	0.937	0.913	0.282	0.776	0.880	0.893	0.910	0.922	0.925	0.912
		AUC	0.943	0.971	0.956	0.767	0.877	0.930	0.935	0.966	0.947	0.966	0.955
	50	R	0.627	0.753	0.891	0.755	0.732	0.645	0.675	0.899	0.680	0.880	0.807
		P	0.872	0.825	0.840	0.284	0.740	0.900	0.888	0.844	0.956	0.897	0.903
		S	0.991	0.985	0.984	0.621	0.979	0.992	0.990	0.986	0.998	0.992	0.992
		F1	0.721	0.770	0.854	0.325	0.735	0.739	0.747	0.869	0.794	0.888	0.837
		MCC	0.718	0.763	0.849	0.297	0.715	0.740	0.749	0.860	0.794	0.879	0.836
		AUC	0.809	0.869	0.937	0.688	0.856	0.819	0.833	0.943	0.839	0.936	0.899
	100	R	0.598	0.892	0.828	0.640	0.874	0.755	0.713	0.842	0.520	0.878	0.656
		P	0.671	0.715	0.831	0.323	0.190	0.799	0.887	0.883	0.808	0.879	0.904
		S	0.878	0.963	0.982	0.601	0.650	0.980	0.991	0.989	0.985	0.989	0.993
		F1	0.496	0.787	0.818	0.336	0.312	0.767	0.789	0.859	0.603	0.878	0.759
		MCC	0.511	0.774	0.807	0.259	0.300	0.751	0.778	0.848	0.608	0.866	0.752
		AUC	0.738	0.919	0.916	0.620	0.763	0.890	0.852	0.905	0.753	0.927	0.829
TB	20	R	0.772	0.862	0.561	0.821	0.683	0.496	0.691	0.772	0.683	0.756	0.740
		P	0.758	0.800	0.852	0.717	0.825	0.886	0.869	0.842	0.886	0.942	0.883
		S	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		F1	0.760	0.827	0.676	0.765	0.738	0.635	0.770	0.805	0.771	0.838	0.805
		MCC	0.762	0.829	0.691	0.766	0.745	0.662	0.775	0.806	0.778	0.843	0.808
		AUC	0.856	0.931	0.780	0.910	0.841	0.748	0.845	0.886	0.841	0.886	0.878
	50	R	0.730	0.854	0.805	0.854	0.748	0.537	0.748	0.813	0.813	0.879	0.715
		P	0.469	0.788	0.797	0.700	0.803	0.880	0.671	0.808	0.603	0.986	0.935
		S	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000
		F1	0.571	0.812	0.797	0.769	0.770	0.667	0.686	0.808	0.683	0.928	0.811
		MCC	0.585	0.816	0.799	0.773	0.773	0.687	0.697	0.809	0.695	0.930	0.818
		AUC	0.865	0.927	0.902	0.927	0.874	0.768	0.874	0.906	0.906	0.939	0.899
	100	R	0.756	0.756	0.911	0.683	0.732	0.512	0.618	0.805	0.780	0.813	0.732
		P	0.470	0.697	0.497	0.583	0.698	0.778	0.790	0.750	0.681	0.828	0.882
		S	0.999	1.000	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		F1	0.579	0.718	0.641	0.629	0.714	0.618	0.642	0.776	0.727	0.820	0.800
		MCC	0.596	0.722	0.671	0.631	0.714	0.631	0.672	0.777	0.729	0.820	0.803
		AUC	0.828	0.878	0.955	0.841	0.866	0.756	0.832	0.902	0.890	0.906	0.866
Spirit	20	R	0.899	0.919	0.929	0.950	0.940	0.881	0.925	0.939	0.920	0.919	0.924
		P	0.899	0.969	0.902	0.380	0.510	0.976	0.945	0.928	0.800	0.976	0.889
		S	0.999	1.000	0.999	0.990	0.995	1.000	1.000	1.000	0.999	1.000	0.999
		F1	0.895	0.943	0.915	0.534	0.661	0.925	0.935	0.933	0.855	0.946	0.905
		MCC	0.897	0.943	0.915	0.591	0.690	0.927	0.935	0.933	0.857	0.947	0.905
		AUC	0.949	0.959	0.964	0.970	0.968	0.940	0.962	0.969	0.960	0.960	0.962
	50	R	0.931	0.913	0.937	0.944	0.946	0.889	0.919	0.932	0.933	0.923	0.919
		P	0.864	0.975	0.921	0.446	0.588	0.969	0.957	0.947	0.921	0.997	0.997
		S	0.999	1.000	1.000	0.993	0.996	1.000	1.000	1.000	1.000	1.000	1.000
		F1	0.896	0.943	0.929	0.603	0.724	0.926	0.937	0.939	0.927	0.959	0.957
		MCC	0.896	0.943	0.928	0.644	0.743	0.927	0.937	0.939	0.926	0.959	0.957
		AUC	0.965	0.962	0.968	0.968	0.971	0.944	0.960	0.966	0.966	0.962	0.961
	100	R	0.938	0.930	0.937	0.829	0.951	0.853	0.937	0.923	0.926	0.937	0.937
		P	0.800	0.985	0.989	0.810	0.556	0.938	0.927	0.989	0.974	0.996	0.947
		S	0.999	1.000	1.000	1.000	0.996	1.000	1.000	1.000	1.000	1.000	1.000
		F1	0.862	0.957	0.962	0.819	0.702	0.893	0.932	0.955	0.950	0.966	0.942
		MCC	0.865	0.957	0.962	0.819	0.726	0.894	0.932	0.955	0.950	0.966	0.941
		AUC	0.964	0.965	0.968	0.915	0.973	0.926	0.968	0.961	0.963	0.968	0.964
Huawei	20	R	0.238	0.571	0.524	0.200	0.000	0.048	0.381	0.429	0.238	0.476	0.095
		P	1.000	0.923	0.733	1.000	0.000	1.000	0.800	1.000	0.833	1.000	1.000
		S	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		F1	0.385	0.706	0.611	0.333	0.000	0.091	0.516	0.600	0.370	0.645	0.174
		MCC	0.488	0.726	0.619	0.447	0.000	0.218	0.551	0.654	0.445	0.690	0.308
		AUC	0.619	0.786	0.762	0.600	0.500	0.524	0.690	0.714	0.619	0.738	0.548
	50	R	0.095	0.429	0.476	0.100	0.238	0.238	0.095	0.571	0.190	0.667	0.143
		P	1.000	0.600	0.500	0.286	0.625	1.000	1.000	1.000	1.000	0.667	1.000
		S	1.000	0.999	0.999	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000
		F1	0.174	0.500	0.488	0.148	0.345	0.385	0.174	0.727	0.320	0.667	0.250
		MCC	0.308	0.506	0.487	0.168	0.385	0.488	0.308	0.756	0.436	0.666	0.378
		AUC	0.548	0.714	0.738	0.550	0.619	0.619	0.548	0.786	0.595	0.833	0.571
	100	R	0.150	0.476	0.619	0.048	0.190	0.000	0.619	0.476	0.286	0.476	0.250
		P	1.000	0.833	0.650	1.000	1.000	0.000	0.867	1.000	1.000	0.833	1.000
		S	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		F1	0.261	0.606	0.634	0.091	0.320	0.000	0.722	0.645	0.444	0.606	0.400
		MCC	0.387	0.629	0.634	0.218	0.436	0.000	0.732	0.690	0.534	0.629	0.500
		AUC	0.575	0.738	0.809	0.524	0.595	0.500	0.809	0.738	0.643	0.738	0.625

specificity by 0.1%, F1 by 8.3%, MCC by 8.5%, and AUC by 0.2%. When $ws = 50$, these reductions become more pronounced: precision drops by 72.1%, recall by 6.3%, specificity by 24.9%, F1 by 59.3%, MCC by 58.5%, and AUC by 3.9%. LogRobust shows more moderate declines. For $ws = 20$, recall

falls by 2.1%, precision by 5.1%, specificity by 0.4%, F1 by 3.7%, MCC by 3.9%, and AUC by 1.8%. For $ws = 50$, the decreases are 8.7% in recall, 8.6% in precision, 0.7% in specificity, 8.6% in F1, 9.7% in MCC, and 3.2% in AUC. Overall, these results highlight the negative impact of class imbalance on

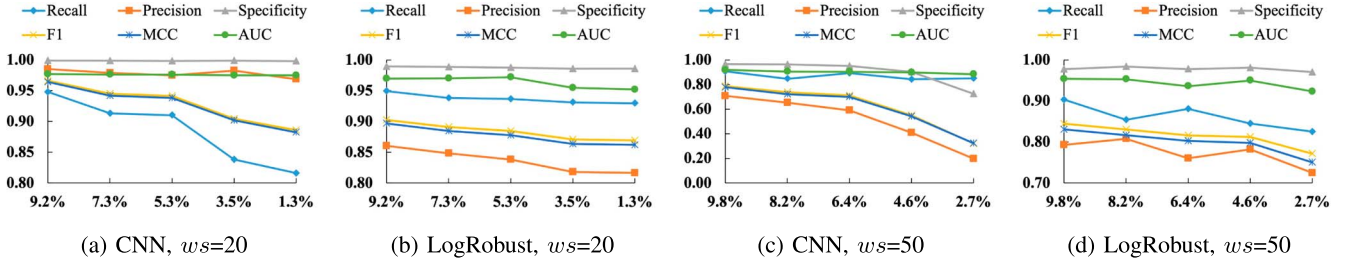


Fig. 2. The performance variations with different anomaly proportions on the BGL dataset, while maintaining data variety.

DLLAD performance, even when the log data variety is kept constant.

Finding 2: DLLAD models perform worse on datasets with more severe class imbalance. Even in datasets like BGL, which have greater log data variety, reducing the anomaly proportion (and thus increasing the severity of class imbalance) leads to a noticeable performance decline in DLLAD models.

B. RQ2. How Does the Resampling Ratio of Normal to Abnormal Data Affect the Performance of DLLAD Approaches?

We comprehensively evaluate the ten data resampling methods for each dataset using quarter-based resampling ratios obtained by multiplying the original ratio of normal data to abnormal data by $1/4$, $1/2$, and $3/4$, as described in Section III-A. The employed data resampling methods are categorized into three groups: OverSampling (comprising ROS_R , $SMOTE$, $ADASYN$, and ROS_F), UnderSampling (encompassing $NearMiss$, IHT , RUS_F , and RUS_R), and HybridSampling, represented by $SMOTEENN$ and $SMOTETomek$. Fig. 3(a) and 3(b) present heatmaps that show the performance of three DLLAD approaches using these resampling methods across datasets with varying window sizes, in terms of F1-score and MCC values. We do not emphasize AUC, as the differences in AUC values across these resampling methods are minimal. Instead, we focus on two comprehensive metrics: F1-score, which balances recall and precision for detected anomalies, and MCC, a fully symmetric metric that considers all four values (TP, TN, FP, and FN) in the confusion matrix. To identify which resampling ratio optimizes performance for DLLAD approaches, we enumerate “hits”, indicating instances where a specific resampling ratio yields the highest performance. For the OverSampling and UnderSampling groups, each cell in the heatmap represents 36 instances (4 oversampling/undersampling methods \times 3 DLLAD approaches \times 3 window sizes). For HybridSampling, each cell represents 18 instances (2 hybrid sampling methods \times 3 DLLAD approaches \times 3 window sizes). If the DLLAD approach achieves the same performance at two different resampling ratios, we do not count it as a “hit”.

In Fig. 3(a) and 3(b), we observe that oversampling methods with a resampling ratio of one-quarter of the original normal-to-abnormal ratio consistently yield superior performance for

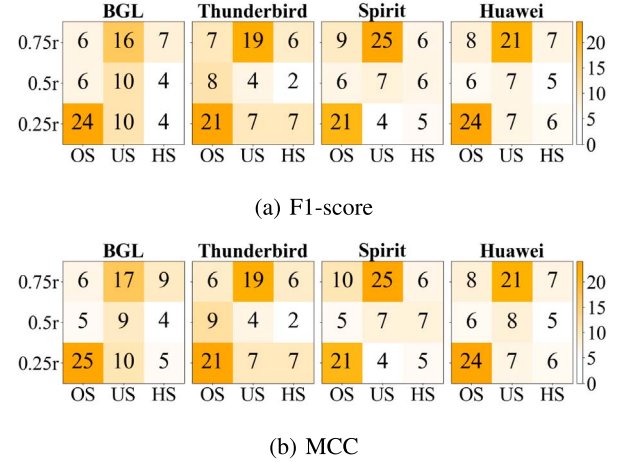


Fig. 3. The heatmaps illustrate the model performance of DLLAD approaches using different data resampling methods (OverSampling (OS), UnderSampling (US), and Hybrid Sampling (HS)) with varying resampling ratios across four datasets and three window sizes. The cells in heatmaps indicate the resampling ratio that achieves the best model performance. r represents the original ratio of normal to abnormal data.

DLLAD approaches. This means creating more abnormal sequences to achieve a target of $1/4$ of the original normal-to-abnormal ratio. This trend is especially pronounced in the BGL and Huawei datasets, with 24 or 25 hits out of a total of 36 for both F1-score and MCC. For the Thunderbird and Spirit datasets, one-quarter of the original ratio results in 21 hits out of 36. In contrast, for undersampling methods, a resampling ratio of three-quarters of the original normal log sequences achieves the best performance for DLLAD approaches. This involves reducing fewer normal sequences to reach a target of $3/4$ of the original normal-to-abnormal ratio. The effect is most evident in the Spirit dataset, with 25 out of 36 hits, followed by Huawei (21 hits), Thunderbird (19 hits), and BGL (16 hits). For hybrid sampling methods, identifying the optimal resampling ratio is challenging. In the Thunderbird, Spirit, and Huawei datasets, at least two resampling ratios show similar effectiveness in achieving the best DLLAD performance. Specifically, the most effective resampling ratios are three-quarters of the original ratio for BGL and Huawei datasets, one-quarter for Thunderbird, and one-half for Spirit. Thus, the optimal resampling ratio varies inconsistently across the four datasets.

Finding 3: To optimize DLLAD approaches with *oversampling*, it is recommended to increase the degree of minority class amplification (i.e., generating more abnormal log sequences). Conversely, for DLLAD approaches employing *undersampling*, it is advisable to reduce the degree of majority class reduction (i.e., removing fewer normal log sequences).

Finding 4: There is no particular preferred resampling ratio for applying *hybrid sampling* methods to DLLAD approaches.

(*SMOTETomek*) methods in 26/9/1 and 30/4/2 out of 36 cases, respectively.

Finding 5: Employing ROS_R , $SMOTE$, $ADASYN$, and ROS_F to alleviate the class imbalance can yield better results compared to *NoSampling* in DLLAD.

Finding 6: Data resampling directly on the raw data generally outperforms resampling in the feature space.

Finding 7: *Oversampling* exhibits better performance compared to *undersampling* and *hybrid sampling*.

C. RQ3. Does Data Resampling Improve the Effectiveness of Existing DLLAD Approaches?

Tables V–VII present the results of the three DLLAD approaches with ten data resampling methods and *NoSampling* on the four datasets. In line with the RQ2 recommendation, the oversampling, undersampling, and hybrid sampling methods apply resampling ratios of one-quarter, three-quarters, and three-quarters of the original normal-to-abnormal ratio, respectively. We utilize three distinct colors to underscore the statistical differences among data resampling methods and *NoSampling* in terms of all evaluation metrics: Recall, Precision, Specificity, F1, MCC, and AUC. These methods are categorized into multiple groups using the Scott-Knott ESD test, often resulting in more than six groups. Darker colors indicate higher-ranked groups, with each color representing two groups (i.e., the darkest purple denotes the first and second groups, moderate purple indicates the third and fourth groups, and the lightest purple represents the fifth and sixth groups). If some data resampling methods fall outside the top six groups but significantly outperform *NoSampling*, the results of these data resampling methods will be bolded.

(1) General Insights on data resampling. We observe that employing ROS_R on all three DLLAD approaches consistently yields superior performance. Specifically, ROS_R ranks within the top two groups (represented by the darkest purple color) in all 36 cases (3 DLLAD approaches \times 4 datasets \times 3 window sizes) in terms of at least two comprehensive metrics. $SMOTE$ is the second recommended data resampling method, demonstrating significantly better performance than *NoSampling* in 34 out of 36 cases across all datasets in terms of at least two comprehensive metrics. Similarly, $ADASYN$ and ROS_F outperforms *NoSampling* in 33 out of 36 cases, followed by RUS_R in 30 out of 36 cases, $SMOTETomek$ in 28 out of 36 cases, and RUS_F in 24 out of 36 cases. In contrast, $SMOTEENN$, *NearMiss*, and *IHT* exhibit poor performance, succeeding in only 14 out of 36, 11 out of 36, and 7 out of 36 cases, respectively. Furthermore, we find that data resampling directly on the raw data generally statistically significantly performs better than data resampling in the feature space. For example, ROS_R and RUS_R perform better, equally, or worse than ROS_F and RUS_F in 17/18/1 and 15/16/5 out of 36 cases, respectively. Additionally, *oversampling* exhibits statistically significantly superior performance compared to *undersampling* and *hybrid sampling*. For instance, the best oversampling method ROS_R performs better, equally, or worse than the best undersampling (RUS_R) and hybrid sampling

(2) The impact of data resampling on the DLLAD approaches across datasets with varying imbalance levels. In datasets characterized by moderate imbalance, such as BGL and Spirit, certain data resampling methods contribute to improved performance in all three DLLAD approaches. Given the consistent results between F1 and MCC, and the minimal changes in AUC in certain cases, our analysis primarily focuses on the outcomes of MCC due to the space limitation. For the BGL dataset with $ws = 20$, CNN (Table V), LogRobust (Table VI), and NeuralLog (Table VII) achieve high performance with *NoResampling*, with MCC values of 0.964, 0.897, and 0.864, respectively. Data resampling methods, excluding *IHT*, slightly improve MCC for CNN by 0.4%–1.2% and LogRobust by 1.3%–3.2%. NeuralLog shows a more significant improvement of 1.8%–8.4% with all resampling methods except *NearMiss* and *IHT*. With $ws = 50$, CNN improves by 5.5%–14.6% with resampling methods, except *IHT*, ROS_F , and RUS_F . LogRobust and NeuralLog see improvements of 3.2%–11.4% and 3.1%–22.4%, respectively, with most resampling methods except *NearMiss* and *IHT*. At a window size of 100, the impact of data resampling becomes more pronounced, especially in enhancing recall by reducing false positives. CNN shows a 41.8%–78.8% improvement with most resampling methods except *IHT*. LogRobust improves by 3.2%–11.4% with all resampling methods, and NeuralLog by 19.0%–69.6%, except for *NearMiss* and *IHT*. For the Spirit dataset across three window sizes, certain data resampling methods improve the performance of CNN and LogRobust, with enhancements in MCC ranging from 4.0% to 26.0% for CNN and 2.0% to 8.9% for LogRobust. These improvements occur with resampling methods, except for *NearMiss*, *IHT*, $SMOTEENN$, $SMOTETomek$, and for LogRobust, also RUS_F . NeuralLog also benefits from data resampling methods, except *NearMiss*, *IHT*, and RUS_F , with improvements of 0.9%–11.7%. **Therefore, all oversampling methods (ROS_R , $SMOTE$, $ADASYN$, and ROS_F) and the undersampling method RUS_R demonstrate statistically significant improvements on three DLLAD approaches in datasets with higher anomaly proportions ($>4.4\%$).**

In the highly imbalanced Thunderbird dataset (anomaly proportion $<0.4\%$), ROS_R and $SMOTE$ stand out, particularly ROS_R , which shows statistically significant improvements across all cases. For CNN (Table V), undersampling methods like RUS_F , RUS_R , *NearMiss*, as well as the hybrid sampling method $SMOTEENN$, uniformly cause the CNN to classify all log sequences as normal sequences, resulting in performance metrics (i.e., Recall, Precision, F1, and MCC) being

reduced to 0. However, ROS_R significantly improves *NoSampling* by 48.2%, 103.2%, and 79.8% for ws set to 20, 50, and 100, respectively. A statistically significant difference in CNN performance is only observed with *SMOTETomek* and ROS_R . LogRobust benefits from some resampling methods, except *NearMiss*, *IHT*, *SMOTEENN*, and *SMOTETomek*, with improvements ranging from 5.0% to 121.0% across three window sizes (shown in Table VI). In contrast to the effects observed on CNN, certain undersampling, oversampling, and hybrid sampling methods except for *ADASYN*, *IHT*, and *SMOTEENN*, can positively impact NeuralLog (Table VII). Particularly, data resampling applied to raw data (i.e., ROS_R and RUS_R) substantially enhances the MCC of *NoSampling* by 6.0% to 59.0%. **Hence, data resampling methods have varied effects on three DLLAD approaches, with only ROS_F being consistently beneficial on severely imbalanced datasets.**

In the highly imbalanced dataset Huawei (anomaly proportion < 0.6%), which has the least variety and volume compared to the public datasets (detailed in Section III-A), CNN and LogRobust perform relatively better than NeuralLog (detailed in Section IV-A). For CNN (Table V), *ADASYN*, ROS_R , and RUS_R improve performance by 4.5%–25.2% across various window sizes. For LogRobust (Table VI), ROS_R , *SMOTE*, *ADASYN*, ROS_F , and *SMOTETomek* boost performance by 3.5%–49.4%. In contrast, NeuralLog (Table VII) shows the most significant improvements, with ROS_R , *SMOTE*, *ADASYN*, and ROS_F enhancing performance by 27.0%–145.1%. The substantial performance gains observed in NeuralLog with oversampling methods suggest that generating synthetic anomalies or duplicating anomalies increases data variety and volume. As mentioned earlier, the limited size and variety of the Huawei dataset may hinder NeuralLog’s complex architecture from achieving effective training. By generating anomalies, oversampling creates a richer and broader dataset, enabling the Transformer-based NeuralLog to better capture complex patterns and relationships in the log data. In addition, we find that copying the embeddings of anomalies (ROS_F) does not always enhance DLLAD model performance (e.g., CNN with $ws = 100$) as effectively as copying the raw text of anomalies (ROS_R). This is likely because oversampling on raw data alters the data distribution, thereby creating a wider dataset that is more effective for generating embeddings and training models. **In short, for severely imbalanced datasets with limited data, oversampling methods, particularly those that create synthetic anomalies (e.g., *ADASYN*) or duplicate raw anomalies (e.g., ROS_R), can significantly enhance the performance of three DLLAD approaches.**

Finding 8: In cases of severe class imbalance (e.g., the abnormal proportion is less than 1%), ROS_R consistently enhances the effectiveness of DLLAD approaches. In cases of moderate class imbalance, ROS_R , *SMOTE*, *ADASYN*, and ROS_F exhibit effectiveness improvements across DLLAD approaches.

Furthermore, there are other noteworthy observations from the analysis. Undersampling methods like *NearMiss* and *IHT*

often perform poorly likely because reducing the majority class can lead to the loss of important information in the log data. Similarly, the hybrid method *SMOTEENN* struggles to enhance DLLAD performance. This method generates synthetic abnormal sequences and then removes log sequences considered noisy or overlapping with the majority class. However, this removal step can also discard valuable sequences that share features with both classes. In contrast, oversampling methods, which focus on generating additional minority class sequences, effectively address class imbalance. Although these methods might introduce some noise, they generally better preserve essential information in the log data. These observations highlight the trade-off between balancing class distributions and retaining critical data when applying data resampling methods in DLLAD.

Finding 9: Undersampling methods *NearMiss* and *IHT* are often ineffective in addressing the data imbalance problem. In addition, the hybrid method *SMOTEENN* generally does not improve the performance of DLLAD approaches.

V. DISCUSSION

A. Why Do(Not) Data Resampling Methods Work?

In RQ3, we investigate the impact of various data resampling methods on the performance of DLLAD approaches. To enhance the interpretability of their performance, we utilize Local Interpretable Model-agnostic Explanations (LIME) [47], which is a widely-used model-agnostic explainable algorithm for explaining deep learning models in software engineering [48], [49], [50], [51], [52], [53], [54]. LIME can identify the tokens associated with the highest attention weights in the neural network model. These tokens are considered highly important for the DLLAD model’s predictions.

Accordingly, we analyze the tokens with the highest attention weights for each abnormal log sequence in the test set and manually filter out less meaningful tokens, such as “from”, “for”, “by”, and “via” to focus on the highly important tokens. For a DLLAD model utilizing a data resampling method, the top ten most important tokens identified in each abnormal log sequence form the important token set [55]. In Table VIII, we use the Spirit dataset for demonstration, which contains the greatest variety of abnormal log sequences. We calculate the frequency of these important tokens, indicating how often each one appears in the important token set. A higher frequency implies more importance for the DLLAD model’s predictions. The table presents the highly important tokens for three DLLAD models with the best four and worst two data resampling methods identified in RQ3. We observe that DLLAD models frequently highlight behavior-related tokens (e.g., *repeated*, *authenticated*, and *request*) and negative tokens (e.g., *no*, *error*, and *cannot*).

In Table VIII, certain tokens, such as *cannot*, *tmreply*, and *request*, exhibit relatively low frequencies when DLLAD models are applied without data resampling. However, models utilizing the data resampling methods tend to focus more on these tokens, resulting in higher frequencies compared

TABLE VIII
THE FREQUENTLY IDENTIFIED IMPORTANT TOKENS IN ABNORMAL LOG SEQUENCES

Model	Resampling	network	no	authenticated	request	error	cannot	repeated	tmreply
CNN	<i>NoSampling</i>	.046	.040	.006	.006	.008	.001	.007	.002
	SMOTE	.026	.047	.007	.010	.011	.005	.010	.006
	ADASYN	.032	.040	.006	.009	.013	.003	.013	.007
	NearMiss	.009	.040	.008	.007	.009	.005	.011	.006
	IHT	.059	.040	.009	.006	.009	.004	.008	.005
	ROS_F	.054	.049	.008	.010	.012	.004	.010	.006
	ROS_R	.028	.044	.009	.008	.011	.004	.012	.007
LogRobust	<i>NoSampling</i>	.039	.044	.007	.004	.007	.002	.010	.004
	SMOTE	.029	.043	.008	.010	.009	.006	.014	.008
	ADASYN	.034	.047	.009	.009	.010	.013	.013	.009
	NearMiss	.017	.042	.009	.007	.007	.004	.011	.005
	IHT	.045	.051	.010	.007	.006	.004	.012	.007
	ROS_F	.040	.050	.007	.009	.008	.015	.011	.010
	ROS_R	.031	.054	.008	.009	.009	.006	.014	.009
NeuralLog	<i>NoSampling</i>	.021	.024	.003	.003	.003	.000	.003	.001
	SMOTE	.018	.033	.004	.005	.004	.008	.007	.002
	ADASYN	.019	.027	.003	.004	.006	.003	.004	.002
	NearMiss	.025	.029	.005	.004	.003	.002	.003	.002
	IHT	.030	.021	.006	.003	.003	.003	.003	.002
	ROS_F	.025	.031	.005	.006	.005	.003	.005	.002
	ROS_R	.022	.025	.002	.004	.004	.003	.004	.003

to those without data resampling. For instance, the token `cannot` shows a notable increase in frequency, with CNN and LogRobust models using *SMOTE* having token frequencies up to 5 and 3 times higher, respectively, than their *NoSampling* counterparts. Similarly, for `tmreply`, the token frequencies with CNN, LogRobust, and NeuralLog using ROS_R are up to 3.5, 2.25, and 3 times higher, respectively, compared to these models without data resampling. Conversely, tokens like `network`, which have higher frequencies in models without data resampling, show reduced frequencies when data resampling methods such as *SMOTE* and *ADASYN* are applied. This shift indicates that these oversampling methods enable models to prioritize more important tokens for the DLLAD model's predictions.

To intuitively explain the observation, Table IX presents an example of LIME results for an actual abnormal log sequence from the Spirit dataset, with the abnormal log event highlighted in bold. To conserve space, continuous repeated log events are denoted by $\times N$. The abnormal log event indicates that the system is unable to complete the expected task response. The key token `cannot` indicates an operation failure, typically signifying an abnormal event. Meanwhile, the subsequent token `tmreply` implies that the reply is not completed, suggesting a potential communication or handling issue between the system and `sadmin2`. This could lead to the task not being executed or properly concluded. The table lists the top five tokens for LogRobust with and without data resampling. We observe that with *NoSampling*, the top five tokens identified by the model do not appear in the abnormal log event, indicating the model's failure to capture key abnormal information within the log sequence. In contrast, when using data resampling methods like ROS_R , *SMOTE*, *ADASYN*, and ROS_F , the model effectively identifies key tokens within the abnormal log event. Notably, `cannot` and `tmreply` rank among the top five tokens when employing the ROS_R and *ADASYN*. These data resampling

methods enable DLLAD models to focus more on the important features of abnormal log events, assigning greater weights to these important tokens and allowing the models to more accurately detect abnormal log sequences.

B. Efficiency of Data Resampling and Model Training

In Section IV, we examine the effectiveness of data resampling methods applied to DLLAD approaches. Additionally, this section evaluates the efficiency of these approaches and their impact on model training time. For this analysis, we use NeuralLog, a model that typically requires more training time, applied across the four datasets with a window size of 20. The average data resampling time ($T_{resample}$) and model training time (T_{train}) are presented in Table X.

Data resampling on raw data methods, such as ROS_R and RUS_R , is extremely efficient, taking less than 1 millisecond, making it negligible. In contrast, data resampling methods applied in the feature space, including ROS_F , RUS_F , *SMOTE*, *ADASYN*, and *NearMiss*, take up to 7 minutes. More complex methods like *IHT*, *SMOTEENN*, and *SMOTETomek* exhibit much longer processing times, ranging from 27 to 89 minutes. Regarding model training time, undersampling methods generally lead to faster training for NeuralLog. For instance, RUS_R and RUS_F reduce training time by approximately 4 minutes compared to *NoSampling*. In contrast, oversampling and hybrid sampling methods increase training time, but the increase remains under 6 minutes compared to *NoSampling*.

C. Implications of Our Findings

Sections IV and V-B offer a comprehensive analysis of the impact of various data resampling methods on the effectiveness and efficiency of DLLAD approaches. Based on the insights, we outline several practical implications for practitioners.

TABLE IX

THE ABNORMAL LOG EVENT WITHIN THE ABNORMAL LOG SEQUENCE IS HIGHLIGHTED IN BOLD. THE TOP FIVE TOKENS IDENTIFIED BY LIT FOR LOGROBUST ARE LISTED BELOW, WITH TOKENS FROM THE ABNORMAL LOG EVENT SHOWN IN BOLD

Log Sequence
from=<#<*>#@#<*>#>, size=<*>, nrpt=<*>(queue active) ×3 to=<#<*>#@#<*>#>, relay=none, delay=<*>, status=deferred (Name service error for name=<#<*># type=MX: Host not found, try again) ×3 mount request from <*>for <*><*>×3 from <*>via <*>network <*>.<*>.<*>/<*>: no free leases ×4 mount request from <*>for <*><*> from <*> from <*>via <*>network <*>.<*>.<*>/<*>: no free leases running on <*>privileges. password <*>for user <*>(<#<*>#@#<*>#). authentication for user <*>accepted. #<*>#, coming from <*>authenticated. from <*>via <*>network <*>.<*>.<*>/<*>: no free leases mount request from <*>for <*><*> closed for user <*> opened for user <*>by <*> closed for user <*> opened for user <*>by <*> publickey for <*>from <*>port <*>ssh<*>×2 repeated <*>times cannot tm_reply to <*>.sadmin2 task <*> closed for user <*> opened for user <*>by <*> closed for user <*> opened for user <*>by <*> publickey for <*>from <*>port <*>ssh<*>×2 from <*>via <*>network <*>.<*>.<*>/<*>: no free leases ×8 CMD (test -x /etc/pbs_stat.py && /etc/pbs_stat.py cron)
Top Five Tokens
NoSampling: name, no, leases, network, closed ROS _R : leases, no, tm_reply , repeated, cannot SMOTE: MX, accepted, free, repeated, tm_reply ADASYN: error, found, root, tm_reply , cannot ROS _F : service, free, network, no, cannot NearMiss: /etc/pbs_stat.py, root, authenticated, leases, free IHT: authenticated, no, network, free, accepted

TABLE X

THE AVERAGE DATA RESAMPLING TIME ($T_{resample}$) AND MODEL TRAINING TIME (T_{train}) FOR NEURALLOG ON FOUR DATASETS WITH $ws = 20$

Resampling	$T_{resample}$	T_{train}	$T_{resample+train}$
NoSampling	-	18m21s	18m21s
ROS _R	< 1ms	19m12s	19m12s
RUS _R	< 1ms	14m10s	14m10s
ROS _F	4m12s	23m48s	28m
RUS _F	2m31s	14m16s	16m47s
SMOTE	4m24s	23m33s	27m57s
ADASYN	6m32s	23m38s	30m10s
NearMiss	1m59s	16m4s	18m3s
IHT	89m16s	19m11s	108m27s
SMOTEENN	30m46s	18m55s	49m41s
SMOTETomek	27m8s	21m10s	48m18s

(1) Effectiveness prioritization: For non-urgently training DLLAD models to detect log anomalies, practitioners prioritize high accuracy in predicted results. Utilizing ROS_R generally enhances DLLAD model performance. Notably, in highly imbalanced datasets like Thunderbird, the performance of three DLLAD models improves consistently when employing ROS_R.

In the Huawei dataset—also highly imbalanced but characterized by lower data volume and less variety—certain oversampling methods (e.g., ROS_R and ADASYN) that duplicate or create abnormal log sequences are beneficial for improving DLLAD performance. Therefore, we recommend using ROS_R with the DLLAD model for effective predictions.

(2) Efficiency prioritization: For scenarios requiring efficient training of DLLAD models to quickly identify log anomalies, CNN and LogRobust are preferable due to their shorter training times. Furthermore, using ROS and RUS methods on raw data for duplication and removal before embedding takes considerably less time than data resampling methods in feature spaces, as indicated in Table X. These methods also perform well, achieving top-1 and top-5 ranks among the ten data resampling methods evaluated (detailed in Section IV-C). Based on our experimental findings, LogRobust outperforms CNN overall. Therefore, employing ROS_R and RUS_R with LogRobust is recommended for efficient prediction.

(3) Further enhancement: As discussed in Section V-A, our analysis identifies certain important tokens that contribute to DLLAD, such as cannot and tmreply in the Spirit dataset. When employing data resampling methods like SMOTE, ADASYN, and ROS_R, these tokens receive increased attention weights, thereby aiding DLLAD models in better predicting log anomalies. Building on these insights, future research may be able to develop enhanced data oversampling methods that focus on generating or duplicating abnormal log sequences containing these important tokens. By prioritizing log sequences with highly important tokens, these methods aim to improve the training process, enabling the model to learn more effectively from relevant abnormal features. This targeted data resampling strategy balances the dataset, thereby amplifying the presence of key abnormal log features and ultimately leading to improved anomaly detection performance.

D. Threats to Validity

(1) Limited models and datasets. One potential concern pertains to our selection of DLLAD approaches, we adapted existing supervised approaches for our empirical investigation. This choice is motivated by several factors: Semi-supervised approaches leverage only a fraction of normal logs, whereas unsupervised methods assume datasets lack labels, diverging from our fully supervised data scenario. Furthermore, some unsupervised and semi-supervised approaches share a model structure similar to NeuralLog, as observed in LAnoBERT [37] and Hades [35]. Previous empirical studies [1], [2] have consistently shown inferior performance of unsupervised and semi-supervised approaches compared to supervised ones. Hence, we deliberately include the latest supervised approaches as our evaluated DLLAD approaches. Another concern arises from the limited availability of datasets. Currently, only a few public datasets are available (e.g., HDFS, BGL, Thunderbird, and Spirit). To enhance our study, we include the recently released Huawei dataset. These datasets exhibit different levels of class imbalance and data variety, allowing our empirical study to generalize its findings more effectively.

(2) **Hyperparameter settings.** The selection of an appropriate resampling ratio of normal to abnormal data constitutes a threat when assessing the effectiveness of data resampling methods. Considering numerous data resampling methods and the impracticality of evaluating all of them across diverse datasets with exhaustive resampling ratio variations, it is crucial to systematically choose the resampling ratio. To address this challenge, we introduce a standardized quarter-based unit for a consistent benchmark across various data resampling methods. On this basis, the general conclusions obtained aid researchers in narrowing their focus for subsequent study phases.

(3) **Generalizability.** We intentionally select ten data resampling methods from three distinct categories and systematically apply them to three representative DLLAD approaches across four publicly available datasets. Additionally, Lyu et al. [54] highlight that the randomness introduced by data resampling (i.e., undersampling the majority class) can result in internal inconsistencies in model interpretations. To alleviate this issue, we average the results over five runs to reduce the impact of randomness and internal inconsistencies. Our objective does not compare the effectiveness of those DLLAD approaches but rather focuses on evaluating the capabilities of different data resampling methods applied to those approaches. As such, our findings aim to highlight general trends in how data resampling affects different DLLAD approaches, with the ultimate goal of offering valuable insights to inform future research endeavors.

VI. RELATED WORK

Deep Learning-Based Log Anomaly Detection. Since the work by He et al. [38] (which evaluated six machine learning-based anomaly detection approaches and compared their accuracy and efficiency on two representative production log datasets), numerous DLLAD models have emerged. These DLLAD models typically fall into several categories, including CNN-based, LSTM-based, and Transformer-based models. Lu et al. [9] employed CNN with three filters to extract local semantic information from log data. Zhang et al. [56], Du et al. [8], and Meng et al. [11] adopted LSTM [57] to capture long-term dependencies in log sequences and learn log patterns for predicting the next log. Vinayakumar et al. [58] employed a stacked-LSTM model to learn temporal patterns using sparse representations. Zhang et al. [10] proposed the LogRobust method that integrated the attention mechanism with a Bi-LSTM model, enabling comprehensive sequence information capture in both directions. Li et al. [59] employed a unified attention-based Bi-LSTM model to learn the patterns for sequential anomaly detection. Le et al. [2] introduced NeuralLog, utilizing BERT for embedding representation and a Transformer encoder for log anomaly detection classification. To the best of our knowledge, only the work by Le et al. [1] has highlighted that DLLAD models trained on highly imbalanced datasets exhibit low precision or recall values. Yet, there is currently no research exploring whether data resampling methods can mitigate class imbalance issues and improve DLLAD model performance. Subsequent work, such as unsupervised

approach LAnoBERT [37], and semi-supervised approaches like PLELog [12], AdaLog [13], and Hades [35], have emerged to alleviate the potential difficulty of acquiring large labeled log datasets for training supervised learning models in log anomaly detection. However, we chose to focus on CNN [9], LogRobust [10], and NeuralLog [2] as the DLLAD models in our study, instead of these recently proposed semi-supervised or unsupervised methods. Our rationale behind this decision is based on the following reasons. (1) Different data scenarios: semi-supervised approaches use only a portion of the normal logs, and unsupervised approaches assume datasets have no labels, which differs from our fully supervised data hypothesis scenario. (2) Similar model structure: some unsupervised and semi-supervised approaches share a model structure similar to NeuralLog [2]. For example, LAnoBERT [37] employed pre-trained BERT for unsupervised learning with a masked language modeling loss function, and Hades utilized FastText for semantic vector representation and Transformer for classification. (3) Performance disparities: previous empirical studies [1], [2] have indicated notably inferior performance of unsupervised and semi-supervised approaches compared to supervised approaches.

Data Resampling for Software Engineering. Data resampling has been widely applied to address the class imbalance issue in the field of software engineering, such as quality prediction [60], bug classification [61], [62], defect prediction [28], [42], [63], [64], [65], [66], [67], [68], and code smell detection [69], [70]. For example, Zheng et al. [61] analyzed the impact of six data resampling methods (e.g., SMOTE, Mahakil [71], and Rose [72]) on multiple classifiers for bug report classification. They found that the combination of Rose with random forest yielded the best performance. Bennin et al. [65] observed that while their investigated data resampling methods have no statistically significant effect on defect prioritization, these methods improve the defect classification performance with regard to Recall and G-mean. Subsequently, they [42] demonstrated that random undersampling and borderline-SMOTE are the more stable data resampling methods in software defect prediction. Li et al. [70] investigated the effects of 31 imbalanced learning methods on machine learning classifiers for code smell detection. Their study revealed varied impacts of these methods across different code smells, with deep forest consistently enhancing performance. Additionally, certain data resampling methods such as CNN, ENN, BSMOTE, and ROS showed superior performance compared to SMOTE. Differently from prior work, in the context of log anomaly detection, which poses unique challenges such as the need for careful selection of window sizes, our research conducts an extensive analysis to explore the influence of data resampling methods on model performance across various window sizes. Furthermore, our study delves into the effects of some data resampling methods on both raw data and the feature space, providing valuable insights for practitioners in this field.

Deep Learning-Based Anomaly Detection in Other Fields. Deep learning-based anomaly detection is applicable not only in software log anomaly detection but also in diverse fields like fraud detection [73], [74], [75], [76], medical diagnosis

[14], [77], [78], [79], manufacturing defect detection [80], and network intrusion detection [81]. While the workflow for anomaly detection in these domains shares similarities with the depicted Fig. 1, specific data preprocessing steps, such as log parsing and grouping, may not be applicable. In these fields, class imbalance also presents a significant challenge that can affect the performance of anomaly detection models. Various data resampling methods have been employed to address this issue. For example, Roy et al. [74] implemented RUS with a recommended resampling ratio [82] of 10:1 for non-fraudulent to fraudulent credit card transaction data, followed by an LSTM model for detecting fraud in credit card transactions. Li et al. [78] employed a CNN-based model with adjusted class weights to analyze phonocardiograms for abnormal heart sound detection. Abdelkhalek et al. [81] proposed a data resampling approach combining ADASYN and Tomek Links in conjunction with diverse deep learning models (such as LSTM and CNN) for improved detection of malicious attacks, aiming to address the class imbalance issue between normal traffic and attack samples. In summary, the aforementioned studies suggest that certain data resampling methods can enhance model performance, aligning with our findings. However, empirical research regarding the most effective data resampling methods, optimal resampling ratio, and their impacts on both raw data and the feature space is currently lacking in these fields.

VII. CONCLUSION

Our study represents a pioneering effort in comprehensively assessing the impact of ten data resampling methods on alleviating class imbalance in DLLAD. Through empirical analysis, we have derived several critical insights. Firstly, severe data imbalances, like in the Thunderbird dataset, often lead to poor performance in DLLAD approaches. Secondly, oversampling methods generally outperform both undersampling and hybrid sampling methods. Moreover, data resampling on raw data yields superior results compared to data resampling in the feature space. Notably, ROS_R exhibits outstanding performance, particularly in scenarios characterized by severe class imbalances. Additionally, certain undersampling and hybrid sampling methods, such as $SMOTEENN$, $NearMiss$, and IHT , show limited effectiveness in most cases. Our exploration of different resampling ratios of normal to abnormal data provides actionable recommendations for optimizing the impact of data resampling on DLLAD approaches. By adopting the recommended methods with specific resampling ratios, the performance of DLLAD approaches can be significantly enhanced. Furthermore, we offer implications for improved data resampling strategies through the oversampling of abnormal log sequences that contain important tokens contributing to DLLAD. Our source code and data are publicly available².

REFERENCES

- [1] V.-H. Le and H. Zhang, "Log-based anomaly detection with deep learning: How far are we?" in *Proc. 44th Int. Conf. Softw. Eng.*, 2022, pp. 1356–1367.
- [2] V.-H. Le and H. Zhang, "Log-based anomaly detection without log parsing," in *Proc. 36th IEEE/ACM Int. Conf. Autom. Softw. Eng. (ASE)*, Piscataway, NJ, USA: IEEE Press, 2021, pp. 492–504.
- [3] Q. Fu, J.-G. Lou, Q. Lin, R. Ding, D. Zhang, and T. Xie, "Contextual analysis of program logs for understanding system behaviors," in *Proc. 10th Work. Conf. Min. Softw. Repositories (MSR)*, Piscataway, NJ, USA: IEEE Press, 2013, pp. 397–400.
- [4] Q. Fu et al., "Where do developers log? An empirical study on logging practices in industry," in *Proc. 36th Int. Conf. Softw. Eng.*, 2014, pp. 24–33.
- [5] H. Jiang, X. Li, Z. Yang, and J. Xuan, "What causes my test alarm? Automatic cause analysis for test alarms in system and integration testing," in *Proc. IEEE/ACM 39th Int. Conf. Softw. Eng. (ICSE)*, Piscataway, NJ, USA: IEEE Press, 2017, pp. 712–723.
- [6] C. Zhi, J. Yin, S. Deng, M. Ye, M. Fu, and T. Xie, "An exploratory study of logging configuration practice in Java," in *Proc. IEEE Int. Conf. Softw. Maintenance Evol. (ICSME)*, Piscataway, NJ, USA: IEEE Press, 2019, pp. 459–469.
- [7] S. He, P. He, Z. Chen, T. Yang, Y. Su, and M. R. Lyu, "A survey on automated log analysis for reliability engineering," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–37, 2021.
- [8] M. Du, F. Li, G. Zheng, and V. Srikumar, "DeepLog: Anomaly detection and diagnosis from system logs through deep learning," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, 2017, pp. 1285–1298.
- [9] S. Lu, X. Wei, Y. Li, and L. Wang, "Detecting anomaly in big data system logs using convolutional neural network," in *Proc. IEEE 16th Int. Conf. Dependable Autonom. Secur. Comput., 16th Int. Conf. Pervasive Intell. Comput., 4th Int. Conf. Big Data Intell. Comput. Cyber Sci. Technol. Congr. (DASC/PiCom/DataCom/CyberSciTech)*, Piscataway, NJ, USA: IEEE Press, 2018, pp. 151–158.
- [10] X. Zhang et al., "Robust log-based anomaly detection on unstable log data," in *Proc. 27th ACM Joint Meet. Eur. Softw. Eng. Conf. Symp. Found. Softw. Eng. (ESEC/SIGSOFT FSE)*, 2019, pp. 807–817.
- [11] W. Meng et al., "LogAnomaly: Unsupervised detection of sequential and quantitative anomalies in unstructured logs," in *Proc. 28th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2019, pp. 4739–4745.
- [12] L. Yang et al., "Semi-supervised log-based anomaly detection via probabilistic label estimation," in *Proc. IEEE/ACM 43rd Int. Conf. Softw. Eng. (ICSE)*, Piscataway, NJ, USA: IEEE Press, 2021, pp. 1448–1460.
- [13] X. Ma, J. Keung, P. He, Y. Xiao, X. Yu, and Y. Li, "A semi-supervised approach for industrial anomaly detection via self-adaptive clustering," *IEEE Trans. Ind. Inform.*, vol. 20, no. 2, pp. 1687–1697, Feb. 2024.
- [14] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, pp. 1–54, 2019.
- [15] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano, "Experimental perspectives on learning from imbalanced data," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 935–942.
- [16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
- [17] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intell.)*, Piscataway, NJ, USA: IEEE Press, 2008, pp. 1322–1328.
- [18] I. Mani and I. Zhang, "KNN approach to unbalanced data distributions: A case study involving information extraction," in *Proc. Workshop Learn. Imbalanced Datasets*, vol. 126, ICML, 2003, pp. 1–7.
- [19] M. R. Smith, T. Martinez, and C. Giraud-Carrier, "An instance level analysis of data complexity," *Mach. Learn.*, vol. 95, pp. 225–256, Nov. 2014.
- [20] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newslett.*, vol. 6, no. 1, pp. 20–29, 2004.
- [21] Q. Lin, H. Zhang, J.-G. Lou, Y. Zhang, and X. Chen, "Log clustering based problem identification for online service systems," in *Proc. 38th Int. Conf. Softw. Eng. Companion*, 2016, pp. 102–111.
- [22] P. He, J. Zhu, Z. Zheng, and M. R. Lyu, "Drain: An online log parsing approach with fixed depth tree," in *Proc. IEEE Int. Conf. Web Serv. (ICWS)*, Piscataway, NJ, USA: IEEE Press, 2017, pp. 33–40.
- [23] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "FastText.zip: Compressing text classification models," 2016, *arXiv:1612.03651*.
- [24] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, 1988.

²<https://github.com/ResamplingDLLAD/ResamplingEmpirical>

- [25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, [arXiv:1810.04805](https://arxiv.org/abs/1810.04805).
- [26] S. Chakraborty, R. Krishna, Y. Ding, and B. Ray, "Deep learning based vulnerability detection: Are we there yet?," *IEEE Trans. Softw. Eng.*, vol. 48, pp. 3280–3296, 2022.
- [27] X. Yang, S. Wang, Y. Li, and S. Wang, "Does data sampling improve deep learning-based vulnerability detection? Yeas! and Nays!" in *Proc. IEEE/ACM 45th Int. Conf. Softw. Eng. (ICSE)*, Piscataway, NJ, USA: IEEE Press, 2023, pp. 2287–2298.
- [28] L. Pelayo and S. Dick, "Evaluating stratification alternatives to improve software defect prediction," *IEEE Trans. Rel.*, vol. 61, no. 2, pp. 516–525, Jun. 2012.
- [29] R. Malhotra and S. Kamal, "An empirical study to investigate oversampling methods for improving software defect prediction using imbalanced data," *Neurocomputing*, vol. 343, pp. 120–140, May 2019.
- [30] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.
- [31] D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-2, no. 3, pp. 408–421, Jul. 1972.
- [32] I. Tomek, "Two modifications of CNN," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-6, no. 11, pp. 769–772, 1976.
- [33] S. He, J. Zhu, P. He, and M. R. Lyu, "Loghub: A large collection of system log datasets towards automated log analytics," 2020, [arXiv:2008.06448](https://arxiv.org/abs/2008.06448).
- [34] A. Oliner and J. Stearley, "What supercomputers say: A study of five system logs," in *Proc. 37th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw. (DSN'07)*, Piscataway, NJ, USA: IEEE Press, 2007, pp. 575–584.
- [35] C. Lee, T. Yang, Z. Chen, Y. Su, Y. Yang, and M. R. Lyu, "Heterogeneous anomaly detection for software systems via semi-supervised cross-modal attention," in *Proc. IEEE/ACM 45th Int. Conf. Softw. Eng. (ICSE)*, 2023, pp. 1724–1736.
- [36] X. Liu et al., "LogNADS: Network anomaly detection scheme based on log semantics representation," *Future Gener. Comput. Syst.*, vol. 124, pp. 390–405, Nov. 2021.
- [37] Y. Lee, J. Kim, and P. Kang, "LAnoBERT: System log anomaly detection based on bert masked language model," *Appl. Soft Comput.*, vol. 146, Oct. 2023, Art. no. 110689.
- [38] S. He, J. Zhu, P. He, and M. R. Lyu, "Experience report: System log analysis for anomaly detection," in *Proc. IEEE 27th Int. Symp. Softw. Rel. Eng. (ISSRE)*, Piscataway, NJ, USA: IEEE Press, 2016, pp. 207–218.
- [39] Q. Song, Y. Guo, and M. Shepperd, "A comprehensive investigation of the role of imbalanced learning for software defect prediction," *IEEE Trans. Softw. Eng.*, vol. 45, no. 12, pp. 1253–1269, Dec. 2019.
- [40] J. Yao and M. Shepperd, "Assessing software defection prediction performance: Why using the Matthews correlation coefficient matters," in *Proc. 24th Int. Conf. Eval. Assess. Softw. Eng.*, 2020, pp. 120–129.
- [41] R. Moussa and F. Sarro, "On the use of evaluation measures for defect prediction studies," in *Proc. 31st ACM SIGSOFT Int. Symp. Softw. Testing Anal.*, 2022, pp. 101–113.
- [42] K. E. Bennin, J. W. Keung, and A. Monden, "On the relative value of data resampling approaches for software defect prediction," *Empirical Softw. Eng.*, vol. 24, pp. 602–636, Apr. 2019.
- [43] K. E. Bennin, A. Tahir, S. G. MacDonell, and J. Börstler, "An empirical study on the effectiveness of data resampling approaches for cross-project software defect prediction," *IET Softw.*, vol. 16, no. 2, pp. 185–199, 2022.
- [44] C. Tantithamthavorn, S. McIntosh, A. E. Hassan, and K. Matsumoto, "An empirical comparison of model validation techniques for defect prediction models," *IEEE Trans. Softw. Eng.*, vol. 43, no. 1, pp. 1–18, Jan. 2017.
- [45] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 559–563, 2017.
- [46] B. Yu et al., "Deep learning or classical machine learning? An empirical study on log-based anomaly detection," in *Proc. 46th IEEE/ACM Int. Conf. Softw. Eng.*, 2024, pp. 1–13.
- [47] M. T. Ribeiro, S. Singh, and C. Guestrin, "why should i trust you? Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 1135–1144.
- [48] C. Pornprasit, C. Tantithamthavorn, J. Jiarapakdee, M. Fu, and P. Thongtanunam, "PyExplainer: Explaining the predictions of just-in-time defect models," in *Proc. 36th IEEE/ACM Int. Conf. Autom. Softw. Eng. (ASE)*, Piscataway, NJ, USA: IEEE Press, 2021, pp. 407–418.
- [49] C. K. Tantithamthavorn and J. Jiarapakdee, "Explainable AI for software engineering," in *Proc. 36th IEEE/ACM Int. Conf. Autom. Softw. Eng. (ASE)*, Piscataway, NJ, USA: IEEE Press, 2021, pp. 1–2.
- [50] B. Ledel and S. Herbold, "Studying the explanations for the automated prediction of bug and non-bug issues using lime and shap," 2022, [arXiv:2209.07623](https://arxiv.org/abs/2209.07623).
- [51] M. Fan, W. Wei, X. Xie, Y. Liu, X. Guan, and T. Liu, "Can we trust your explanations? Sanity checks for interpreters in android malware analysis," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 838–853, 2021.
- [52] J. Feichtner and S. Gruber, "Understanding privacy awareness in android app descriptions using deep learning," in *Proc. 10th ACM Conf. Data Appl. Secur. Privacy*, 2020, pp. 203–214.
- [53] A. Warnecke, D. Arp, C. Wressnegger, and K. Rieck, "Evaluating explanation methods for deep learning in security," in *Proc. IEEE Eur. Symp. Secur. Privacy (EuroS&P)*, Piscataway, NJ, USA: IEEE Press, 2020, pp. 158–174.
- [54] Y. Lyu, G. K. Rajbahadur, D. Lin, B. Chen, and Z. M. Jiang, "Towards a consistent interpretation of AIOps models," *ACM Trans. Softw. Eng. Methodol. (TOSEM)*, vol. 31, no. 1, pp. 1–38, 2021.
- [55] B. Steenhoek, M. M. Rahman, R. Jiles, and W. Le, "An empirical study of deep learning models for vulnerability detection," in *Proc. IEEE/ACM 45th Int. Conf. Softw. Eng. (ICSE)*, Piscataway, NJ, USA: IEEE Press, 2023, pp. 2237–2248.
- [56] K. Zhang, J. Xu, M. R. Min, G. Jiang, K. Pelechris, and H. Zhang, "Automated it system failure prediction: A deep learning approach," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Piscataway, NJ, USA: IEEE Press, 2016, pp. 1291–1300.
- [57] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [58] R. Vinayakumar, K. Soman, and P. Poornachandran, "Long short-term memory based operation log anomaly detection," in *Proc. Int. Conf. Adv. Comput. Commun. Inform. (ICACCI)*, Piscataway, NJ, USA: IEEE Press, 2017, pp. 236–242.
- [59] X. Li, P. Chen, L. Jing, Z. He, and G. Yu, "SwissLog: Robust and unified deep learning based log anomaly detection for diverse faults," in *Proc. IEEE 31st Int. Symp. Softw. Rel. Eng. (ISSRE)*, Piscataway, NJ, USA: IEEE Press, 2020, pp. 92–103.
- [60] C. Seiffert, T. M. Khoshgoftaar, and J. Van Hulse, "Improving software-quality predictions with data sampling and boosting," *IEEE Trans. Syst., Man, Cybern.-Part A: Syst. Hum.*, vol. 39, no. 6, pp. 1283–1294, Nov. 2009.
- [61] W. Zheng, Y. Xun, X. Wu, Z. Deng, X. Chen, and Y. Sui, "A comparative study of class rebalancing methods for security bug report classification," *IEEE Trans. Rel.*, vol. 70, no. 4, pp. 1658–1670, Dec. 2021.
- [62] R. Shu, T. Xia, J. Chen, L. Williams, and T. Menzies, "How to better distinguish security bug reports (using dual hyperparameter optimization)," *Empirical Softw. Eng.*, vol. 26, pp. 1–37, Apr. 2021.
- [63] Y. Kamei, A. Monden, S. Matsumoto, T. Kakimoto, and K.-i. Matsumoto, "The effects of over and under sampling on fault-prone module detection," in *Proc. 1st Int. Symp. Empirical Softw. Eng. Meas. (ESEM 2007)*, Piscataway, NJ, USA: IEEE Press, 2007, pp. 196–204.
- [64] M. Tan, L. Tan, S. Dara, and C. Mayeux, "Online defect prediction for imbalanced data," in *Proc. IEEE/ACM 37th IEEE Int. Conf. Softw. Eng.*, vol. 2, Piscataway, NJ, USA: IEEE Press, 2015, pp. 99–108.
- [65] K. E. Bennin, J. Keung, A. Monden, P. Phannachitta, and S. Mensah, "The significant effects of data sampling approaches on software defect prioritization and classification," in *Proc. ACM/IEEE Int. Symp. Empirical Softw. Eng. Meas. (ESEM)*, Piscataway, NJ, USA: IEEE Press, 2017, pp. 364–373.
- [66] S. Huda et al., "An ensemble oversampling model for class imbalance problem in software defect prediction," *IEEE Access*, vol. 6, pp. 24184–24195, 2018.
- [67] H. Xu, R. Duan, S. Yang, and L. Guo, "An empirical study on data sampling for just-in-time defect prediction," in *Proc. Artif. Intell. Secur.: 7th Int. Conf. (ICAIS) Proc., Part II 7*, Dublin, Ireland: Springer, July 19–23, 2021, pp. 54–69.
- [68] R. Yedida and T. Menzies, "On the value of oversampling for deep learning in software defect prediction," *IEEE Trans. Softw. Eng.*, vol. 48, no. 8, pp. 3103–3116, Aug. 2022.

- [69] F. Pecorelli, D. Di Nucci, C. De Roover, and A. De Lucia, "A large empirical assessment of the role of data balancing in machine-learning-based code smell detection," *J. Syst. Softw.*, vol. 169, Nov. 2020, Art. no. 110693.
- [70] F. Li, K. Zou, J. W. Keung, X. Yu, S. Feng, and Y. Xiao, "On the relative value of imbalanced learning for code smell detection," *Softw.: Pract. Exp.*, vol. 53, pp. 1902–1927, Oct. 2023.
- [71] K. E. Bennin, J. Keung, P. Phannachitta, A. Monden, and S. Mensah, "MAHAKIL: Diversity based oversampling approach to alleviate the class imbalance issue in software defect prediction," *IEEE Trans. Softw. Eng.*, vol. 44, no. 6, pp. 534–550, Jun. 2018.
- [72] N. Lunardon, G. Menardi, and N. Torelli, "ROSE: A package for binary imbalanced learning," *R J.*, vol. 6, no. 1, pp. 79–89, Jun. 2014.
- [73] Y. Heryadi and H. L. H. S. Warnars, "Learning temporal representation of transaction amount for fraudulent transaction recognition using CNN, stacked LSTM, and CNN-LSTM," in *Proc. IEEE Int. Conf. Cybern. Comput. Intell. (CyberneticsCom)*, Piscataway, NJ, USA: IEEE Press, 2017, pp. 84–89.
- [74] A. Roy, J. Sun, R. Mahoney, L. Alonzi, S. Adams, and P. Beling, "Deep learning detecting fraud in credit card transactions," in *Proc. Syst. Inf. Eng. Des. Symp. (SIEDS)*, Piscataway, NJ, USA: IEEE Press, 2018, pp. 129–134.
- [75] P. Raghavan and N. El Gayar, "Fraud detection using machine learning and deep learning," in *Proc. Int. Conf. Comput. Intell. Knowl. Economy (ICCIKE)*, Piscataway, NJ, USA: IEEE Press, 2019, pp. 334–339.
- [76] W. Hilal, S. A. Gadsden, and J. Yawney, "Financial fraud: a review of anomaly detection techniques and recent advances," *Expert Syst. Appl.*, vol. 193, May 2022, Art. no. 116429.
- [77] K. M. Tsiouris, V. C. Pezoulas, M. Zervakis, S. Konitsiotis, D. D. Koutsouris, and D. I. Fotiadis, "A long short-term memory deep learning network for the prediction of epileptic seizures using EEG signals," *Comput. Biol. Med.*, vol. 99, pp. 24–37, Aug. 2018.
- [78] F. Li, H. Tang, S. Shang, K. Mathiak, and F. Cong, "Classification of heart sounds using convolutional neural network," *Appl. Sci.*, vol. 10, no. 11, 2020, Art. no. 3956.
- [79] T. Fernando, H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Deep learning for medical anomaly detection—A survey," *ACM Comput. Surv. (CSUR)*, vol. 54, no. 7, pp. 1–37, 2021.
- [80] J. Yang, S. Li, Z. Wang, H. Dong, J. Wang, and S. Tang, "Using deep learning to detect defects in manufacturing: A comprehensive survey and current challenges," *Materials*, vol. 13, no. 24, 2020, Art. no. 5755.
- [81] A. Abdelkhalek and M. Mashaly, "Addressing the class imbalance problem in network intrusion detection systems using data resampling and deep learning," *J. Supercomput.*, vol. 79, pp. 1–34, Feb. 2023.
- [82] X.-w. Chen and M. Wasikowski, "FAST: A roc-based feature selection metric for small samples and imbalanced data classification problems," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2008, pp. 124–132.