IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 27 March 2024, accepted 27 April 2024, date of publication 6 May 2024, date of current version 13 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3396926

RESEARCH ARTICLE

Severity-Oriented Multiclass Drone Flight Logs Anomaly Detection

SWARDIANTARA SILALAHI[1], (Graduate Student Member, IEEE), TOHARI AHMAD[1], (Member, IEEE), HUDAN STUDIAWAN[1], (Member, IEEE), EIRINI ANTHI[2], AND LOWRI WILLIAMS[2]

[1]Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia
[2]School of Computer Science and Informatics, Cardiff University, CF24 4AG Cardiff, U.K.

Corresponding author: Tohari Ahmad (tohari@if.its.ac.id)

This work was supported by Institut Teknologi Sepuluh Nopember (ITS), Pendidikan Magister Menuju Doktor untuk Sarjana Unggul (PMDSU), Peningkatan Kualitas Publikasi Internasional (PKPI) Scholarship from the Ministry of Education, Culture, Research and Technology, Indonesia, and the School of Computer Science and Informatics, Cardiff University, U.K.

ABSTRACT The availability of log data recorded by computer-based systems such as operating system and network logs, makes it possible for the stakeholder to look after the system for monitoring, evaluation, and improvement purposes. If an incident happens to the system, the log is the first and most important artefact to recover so that investigations may be performed to gather an understanding of why such incidents may have occurred. Log-based anomaly detection is one of the common approaches to uncovering incident scenarios and finding the root cause of such incidents. In the context of drone flight, incidents reported in logs include errors during take-off, flight range issues, and cancellations of actions. Existing studies employ sequence anomaly detection to check whether an event during a drone flight is anomalous. It needs several preceding events and includes deciding if the following event is legitimate or malicious. However, one single log record can have no relationship to other log events and be malicious at the same time. Thus, several studies explored point anomaly detection, where one log record is the only feature needed. Dividing the anomalies into two categories can be overwhelming as the number of logs generated by a system is large. At the same time, it can be helpful to separate critical anomalies from the less severe ones. Therefore, this study proposes **DroLoVe**, a severity-oriented multiclass anomaly detection approach for drone flight log data. In accordance with the dataset characteristics, where the samples from different severity levels share common features, this paper employs a multitask-based label vector representation to train deep neural network models. After an extensive experiment on several baselines, the proposed scenario outperforms other models from existing studies with promising results. The proposed representation of the label improves the prediction confidence score on various encoder types with 8.6% and 1.8% from focal and cross-entropy scenarios on average, respectively.

INDEX TERMS Anomaly detection, digital forensics, drone forensics, multitask learning, transformer encoder, information security.

I. INTRODUCTION

The availability of digital data produced by computer systems continues to increase exponentially. It is followed by the advancement in many research areas to make use of the data, such as natural language processing, which is used to analyse textual data. This type of data can be found in several contexts stored in computer storage devices, such as runtime logs that are constantly generated during the operational period of the device. The information recorded in log data is highly valuable for many purposes, including running system monitoring, running process conformance checks, and overall system evaluation [1].

The use of log data is critical when incident cases occur, as the empirical events and incident scenarios can be discovered by analysing the log artefacts. Assuming that

The associate editor coordinating the review of this manuscript and approving it for publication was Pedro R. M. Inácio.

the integrity of the log is guaranteed, log data is one of the artefacts with the highest priority for investigating various types of incidents, such as collision, crash landing, cyber-attack, payload delivery issue, and weather-related incidents [2]. It is exemplified by the research effort that has been made by the community on log-based anomaly detection using various approaches, including machine learning- [3], [4], deep learning- [4], [5], [6], and graph-based methods [7], [8].

Different types of systems generate log records in different formats and structures, which are strongly affected by the log statements or templates normally used in each of the systems. Certain systems use highly technical log statements, which consist of domain-specific log elements. For instance, operating system logs have common elements such as host names, Internet Protocol (IP) addresses, protocols, ports, and messages [9]. Thus, analysing a particular system log may need different techniques from one type to another.

Generally, log-based anomaly detection can be performed using either point, collective, or contextual approaches. In the log-based anomaly detection literature, most of the existing studies propose either a contextual or collective approach, where the decision is taken by examining a sequence of log events. However, detecting abnormalities in log data can be performed using a point-based approach, where one single log record is the only feature needed. Therefore, several studies explored utilising a point-based approach to detect anomalous events on logs data, such as in operating systems [9], [10], drone devices [11], and distribution systems [9], [10]. In the drone context, a log message contains a description of occurring events which are triggered by various components such as sensors, peripherals, and firmware. Analysing the log message means analysing events from all aspects of the drone, including sensors, components and features [2].

Typically, existing log-based anomaly detection studies classify log events into two categories: normal and anomaly [3]. This is helpful in an online setting, where the number of detected abnormalities in a certain period of time is considerably small. However, in a digital forensic setting, where the detection is performed on all collected artefacts which might be large in size, the number of the abnormalities is likely to be large [12]. Thus, analysing the detection result in a binary setting is impractical for the investigator. Moreover, out of all detected peculiarities, there might be several negligible ones with less severe impacts on the system or less likely to be related to an incident. Extending binary to multiclass anomaly detection can provide more detailed and contextual detection results to assist in an investigation and analysis [13], [14], [15]. Considering the severity levels of an anomaly, an investigator can adjust the analysis objectives targeting a certain level only, depending on the needs and cases.

As shown in Fig. 1, the drone experiences several events during a flight. When analysing the flight logs from a forensic perspective, distinguishing the severity levels can help the investigator pinpoint critical anomalies to less severe

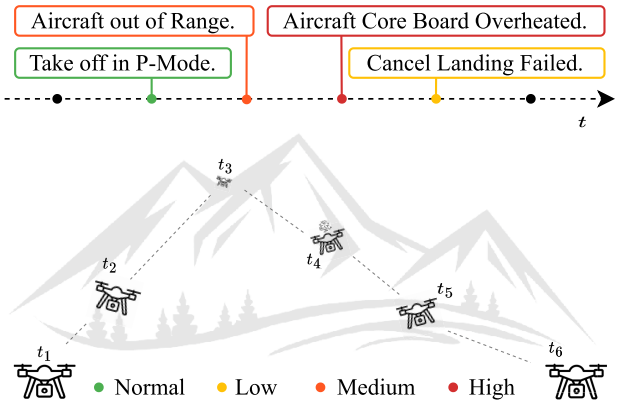


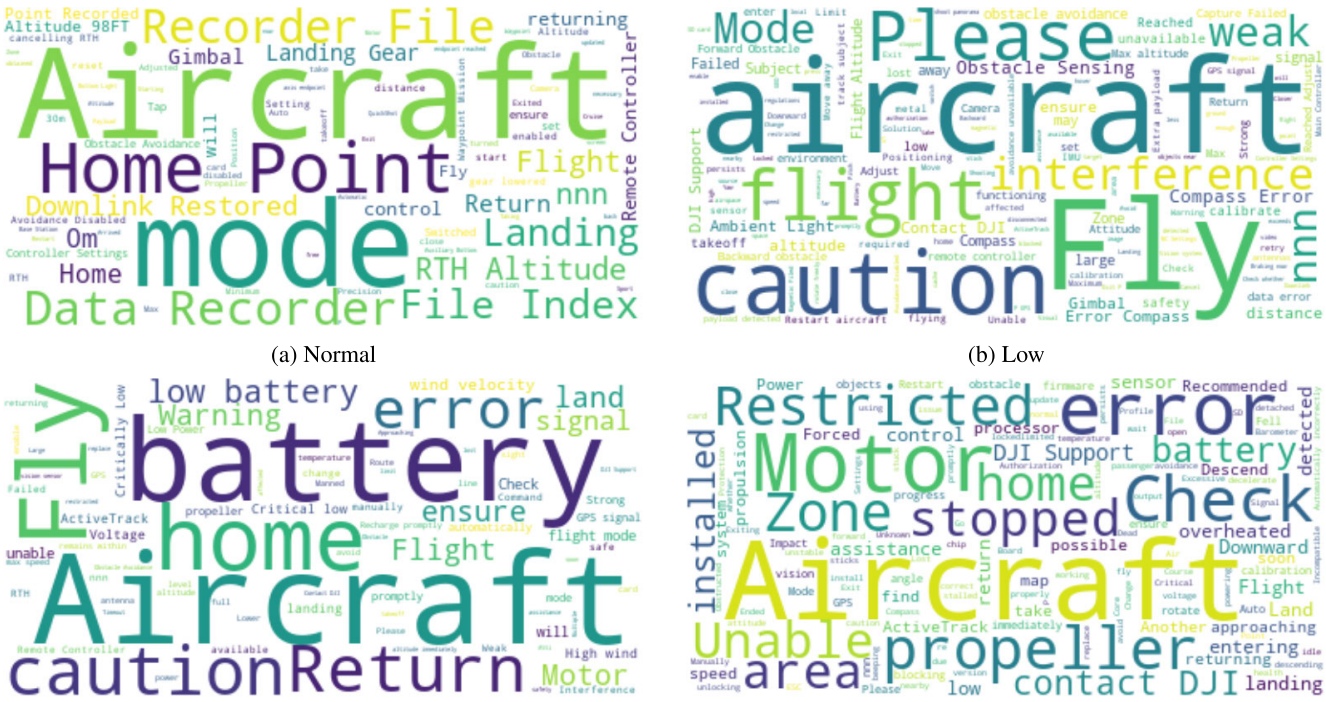
FIGURE 1. An illustration of a flight scenario where several anomalous events with different severity levels happen.

ones. The challenge is that even though the severity level is different, the samples share a common word or even phrase, as depicted in Fig. 2. It makes it challenging to build a point anomaly detection model that depends only on the semantics of the log message. Fig. 3 shows the visualisation of the semantic feature vectors of the samples in the drone flight log datasets obtained from the pre-trained Bidirectional Encoder Representations from Transformers (BERT) [16] after undergoing a dimensional reduction using the t-SNE [17].

This paper is a further experiment of the previous initial study in multiclass anomaly severity detection on drone flight logs data [18]. It is strongly inspired by the work in [9] and [10] where detecting anomalies on logs data is seen as a sentiment analysis task. In this work, the distinctive feature lies within the employment of a sequence classification model with a multitask label to train a detection model for better performance. An extensive experiment is conducted on several different encoders commonly used in log-based anomaly studies, including long short-term memory (LSTM), gated recurrent unit (GRU), transformer, and fine-tuned large language model (LLM). To provide an objective performance comparison of the proposed framework, several baselines are constructed, including those that are proposed in published works.

This paper proposes a transformer encoder-based log-based anomaly detection optimised using multitask label representation to help the model learn from the overlapped features shared by the samples from different severity levels. A domain-specific decoding procedure is also proposed to infer the prediction result. In-depth analyses are performed to evaluate the proposed framework thoroughly. The main **contributions** of this paper are summarised as follows:

1) Propose **Dro**ne **Lo**g **Sev**erity (DroLoVe), a severity-oriented multiclass anomaly detection approach on drone flight log data.
2) Propose a data-driven vector representation of the label and a severity-oriented decoding procedure of the



(a) Normal (b) Low

(c) Medium (d) High

FIGURE 2. A word cloud visualisation to see that the samples from different severity levels share common words and phrases.

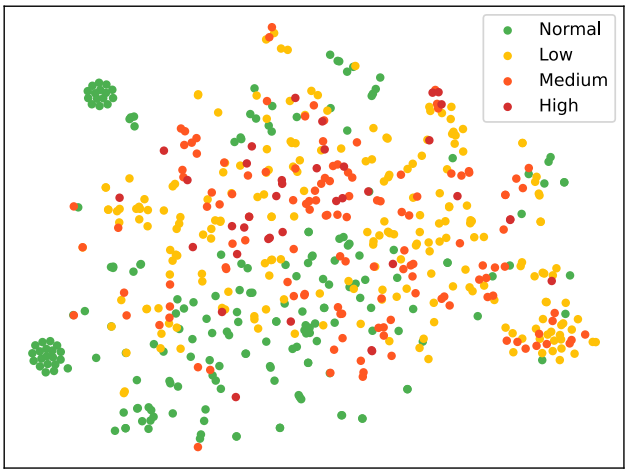


FIGURE 3. A 2D visualisation of the dataset using BERT [16] and t-SNE [17]. It shows that the samples are not linearly separated; samples belong to different classes are close to one another.

label to perform multiclass anomaly severity detection within drone flight log messages.

3) Provide in-depth discussions and analyses of the performance evaluation supported by an extensive experiment on a large number of hyperparameter search spaces.
4) Open the experimental code and results publicly available on a GitHub repository[1] to promote transparent, verifiable, and reproducible research.

[1]https://github.com/swardiantara/DroLoVe

The remainder of the paper is structured as follows: Section II discusses relevant previous studies related to log-based anomaly detection and how to deal with class imbalance problems in log-based anomaly detection task, the proposed framework is explained in detail in Section III, Section IV provides the experimental setup, followed by Section V where the experimental results along with comprehensive analyses and discussions are presented. Section VI discloses the challenges, limitations, and threats to the validity of the study. The conclusion and future works are elaborated in Section VII.

II. RELATED WORK

This section discusses closely relevant prior studies on log-based anomaly detection and different approaches to handle class imbalance problems.

A. ANOMALY DETECTION ON LOGS DATA

Log data recorded by a system store valuable information that can be utilised to perform analysis for investigative purposes, such as anomaly detection, which is commonly referred to as log-based anomaly detection. Recently, with the advancements in Machine Learning (ML) and Deep Learning (DL) models, many studies proposed ML-based [3], [4], DL-based [4], [5], [6], and graph neural network (GNN)-based [7] solutions for detecting abnormalities in log data. Among the published studies, various modelling techniques were used, including point, contextual, and collective. These modelling methods refer to the input needed by the model

TABLE 1. Summary of related works in log-based anomaly detection.

Ref.	Parsing	Pre-process	Input	Model	Handle CIP	Multi Class
DeepSyslog [19]	✓	Cleansing	Sequence	LSTM	✗	✗
DronLomaly [20]	✗	Normalisation	Sequence	LSTM	✗	✗
LayerLog [21]	✗	Cleansing	Sequence	LSTM	✗	✗
LogEncoder [22]	✓	✗	Sequence	BiLSTM	✗	✗
LogGraph [8]	✓	✗	Sequence	GNN	✗	✗
SwissLog [23]	✓	✗	Sequence	BiLSTM	✗	✗
Loader [24]	✗	✗	Sequence	Transformer	✗	✗
NeuralLog [25]	✗	✗	Sequence	Transformer	✗	✗
Pylogsentiment [9]	✓	✗	Point	GRU	Tomek Link	✗
SentiLog [26]	✗	Cleansing	Point	BiLSTM	✗	✗
TransSentLog [10]	✓	✗	Point	Transformer	Tomek link	✗
DroLoVe (Ours)	✗	✗	Point	Transformer	Loss Weight	✓

to perform anomaly detection [9]. In a point setting, the detection is performed against individual log records. The model depends solely on the features from a single log entry. On the contrary, contextual and collective settings utilise a group of log events to detect the presence of abnormalities.

Common problems encountered in log-based anomaly detection include the unstructured nature of log messages, each system having its own log characteristics, being less human-readable, and containing many special or technical terms [22], [23]. To deal with these problems, a typical ML/DL-based log-based anomaly detection comprises several stages, i.e., log cleansing, log parsing, feature extraction, model training, model testing, and model evaluation. The role of log parsing in log analysis and anomaly detection has been critical to perform both online and offline detection [27]. Performing log parsing aims to extract the core features of logs and reduce the noise. However, employing parsing can remove valuable information within the log messages [25]. To prevent parsing errors from being propagated to the next detection phase, utilising a contextual embedding, such as BERT, to extract the semantic features of logs data without performing parsing can be a solution [21]. Therefore, all the features extracted from each log record are preserved as they are. Nevertheless, similar but contradicting log events ended up having features that are close to one another in the latent space. Dealing with such an issue, Qi et al. [22] propose a contrastive-based approach which consists of a representation learning model to provide a decent input to a one-class classifier to distinguish the normal from the abnormal log samples. Instead of removing the parameter values in a log message when performing log parsing, the information can be used as an additional feature along with the metadata of the logs to improve the performance of the models [19].

Log-based anomaly detection has been applied widely to various systems, such as operating systems [9], [10], parallel file systems [26], drones [11], [18], [20], internet of things [28], and industrial control systems [29]. Among these studies, a sequence-based approach is the common modelling technique used, where the detection is observed on a collection of log events. For that reason, a recurrent-based network is employed to capture the sequential dependency and relationship between the occurrences of the log events. For instance, an LSTM [22] and GRU [9] model is used to capture the contextual features of the log sequence. The same approach can also be used in a point-based setting, where the sequence of words or parameters in a log message is the features [9]. Since part of the log record contains a human-readable message, the transformer model can be utilised to extract the semantic information between the words and parameters within a log record to perform point anomaly detection [10], [24]. In this case, given that drone log data also includes natural language, a point-based anomaly detection approach is adopted in this paper. Table 1 presents the summary of the previous related works.

B. CLASS IMBALANCE PROBLEM (CIP) IN LOG-BASED ANOMALY DETECTION

It is assured that in log-based anomaly detection, the number of anomalous samples is significantly less than the normal ones. Employing a supervised-based technique is prone to bias, as the model tends to learn from the majority samples. To overcome CIP, several studies proposed data-level and algorithm-level solutions [30]. Moreover, in a certain case, there are no anomalous samples available. In this particular situation, a one-class approach can be used to construct a normal baseline model. During the detection, an anomaly score is used to decide if an input event is anomalous based on a threshold value [20], [22].

A practical solution to overcome CIP is by controlling the class distribution in the dataset, either by oversampling the minority class, undersampling the majority class, or adding more data to the existing dataset [31]. Overall, either approach can improve the performance of the model, depending on the dataset characteristics [28], [32]. For instance, generating more samples of minority classes using the Synthetic Minority Oversampling Technique (SMOTE) can improve the detection performance of a deep Q network (DQN)-based [33] and deep neural network (DNN)-based [34] models. Instead of duplicating the minority class to add more samples, as in random oversampling, SMOTE uses the k-nearest neighbour as an anchor and creates a new sample that is close to those k samples. It helps the

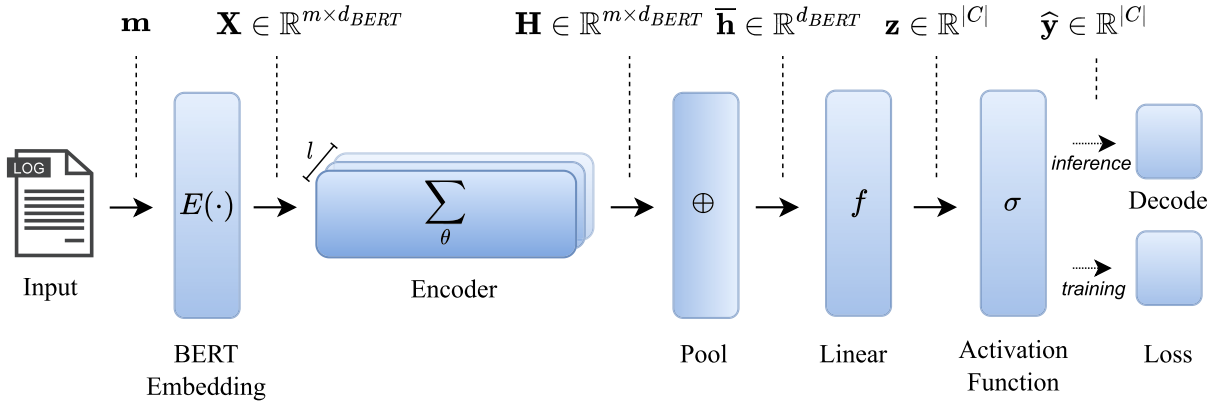


FIGURE 4. The proposed approach overall architecture.

model to learn from minority classes better, instead of from redundant samples produced by random oversampling [34]. However, in other cases, normal samples can be unnecessarily abundant, where removing a certain amount of samples does not reduce the informational value from the dataset. Random undersampling (RUS) is a simple way to eliminate several majority samples to balance the class distribution [35]. Instead of picking random samples to remove, Tomek Link can be used to choose majority samples that are close to minority samples to remove. It can increase the class separability of the dataset, which then helps the model to distinguish between the normal and anomalous events and increases the performance of the model [9]. Performing minority oversampling and majority undersampling simultaneously also makes it possible to balance the class proportion, which then helps improve the performance of the model [36].

Overcoming CIP can also be achieved by using a method-level solution, which revolves around selecting an appropriate model or designing a training procedure. For instance, Qi et al. [37] utilised a bidirectional generative adversarial network to obtain the reconstruction loss and discrimination loss as the features for an n-stacked ensemble classifier to perform anomaly detection., which resulted in an increase in recall. Ensemble models have been shown to perform better than a single classifier, exemplified by [38], who proposed an isolation forest (IF)-based method to perform anomaly detection with various contamination ratios. Similar to the ensemble, training two detection models for detecting the normal and anomalous events separately can reduce false positive cases [39]. A common challenge in using ensemble-based models is how to perform final predictions, considering each sub-model has its own predictions. Therefore, employing self-supervised contrastive learning can be a solution to pre-train a model that can produce distinct features of normal to abnormal samples. Using the decent features from the pre-trained model, a standard clustering is performed to decide if a test event is anomalous based on the Mahalanobis distance score with respect to a threshold value [40]. Clustering can also be used to estimate the unlabeled samples' label probability, which is then used to train a discriminative model. A semi-supervised approach also shows a positive impact on the performance evaluation [41].

Unlike the previously discussed studies, inspired by an existing study from another domain [42], this paper employs loss weighting to train a neural model using an imbalanced dataset. Performing data augmentation at the message level to overcome CIP is impractical in this study as the augmented messages do not exist in the actual scenario. Therefore, this paper explores the effect of using various class weighting schemas and loss functions.

III. PROPOSED FRAMEWORK: DROLOVE

This section presents the proposed method towards a severity-oriented multiclass anomaly detection approach for drone flight log data. The overall flow of the proposed framework is depicted in Fig 4. The following section describes the approach in more detail.

A. OVERVIEW

Performing point log-based anomaly detection that takes human-readable log messages as input is similar to conducting sequence or text classification. As mentioned in Section I, this task may be interpreted as a sentiment analysis task that aims to detect negative sentiments within log messages. Taking an n-length log message $\mathbf{m} = [w_1, w_2, \ldots, w_n]$ from the dataset $\mathbf{D} = \{(\mathbf{m}_i, c_i)\}_{i=1}^{|\mathbf{D}|}$ as the input, where c is the label class name, BERT is used to tokenise the input into a fixed length sequence with m maximum length and retrieve the contextual embeddings, resulting in an input matrix $\mathbf{X} \in \mathbb{R}^{m \times d_{BERT}}$ which is then paired with the encoded label $\mathbf{y}$. Thus, the numerical input becomes $(\mathbf{X}, \mathbf{y})$. This study aims to train a neural network model $\mathcal{M}_\theta$ to classify the message $\mathbf{m}$ into one of the predefined classes $c \in C = \{\texttt{high}, \texttt{medium}, \texttt{low}, \texttt{normal}\}$, or can be written as $\mathcal{M}_\theta : \mathbf{m} \rightarrow c$.

Based on past literature discussed in Section II, processing the sequence of words and performing classification based on the sequence-level features is best performed using modern deep learning models such as LSTM, GRU, and Transformer.

In this study, these models are used in the experiment to identify the best-performing one. The encoder takes the input matrix $\mathbf{X}$ to learn the contextual dependencies among the tokens in the input sequence, yielding the encoder hidden states $\mathbf{H} \in \mathbb{R}^{m \times 768}$, as the BERT-base model produced a 768-dimensional vector. Note that the encoder can be stacked in layers depending on the needs. In this study, the number of layers varies between one and three to prevent the model from being too complex.

Before passing the hidden states from the encoder to the final linear classifier, a pooling is performed to aggregate the features from each token within the sequence to get the final sequence-level representation. To this end, various pooling techniques are used: maximum, average, CLS (Classify token), and *last* as depicted in Fig. 5. Max and average pooling are performed element-wise, meaning that each vector element of the token embedding in the corresponding position is aggregated. CLS is a special token from the BERT pre-training task used to represent the sequence features. Pooling CLS means taking the vector representation of the CLS token as the final feature. As for the *last* refers to the representation of the last token in the sequence; this only applies to LSTM and GRU models. This aligns with the nature of the recurrent model, where the hidden state of the last token is considered to be the sequence representation. When the model employs bi-directionality, the hidden state of the last token from the forward direction and the hidden state of the first token from the backward direction are concatenated to form the same size of the final hidden state as the unidirectional one. The pooling layer takes the hidden state $\mathbf{H}$ as the input and resulting the vector $\bar{\mathbf{h}} \in \mathbb{R}^{768}$.

The next step after performing pooling is to feed the vector $\bar{\mathbf{h}}$ to the linear layer, yielding the unnormalised logits $\mathbf{z} \in \mathbb{R}^{|C|}$, where $|C|$ denotes the number of target class in the dataset. During the training phase, these vectors are used to compute the loss and update the parameter of the model after passing through a normalisation function. For the typical one-hot encoding representation of the label, the standard cross-entropy is used to compute the loss, as defined in the following equation:

$$\mathcal{L}^{\text{CE}} = -\sum_{i=1}^{|C|} y_i \cdot \log(\hat{y}_i) \tag{1}$$

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{z}) \tag{2}$$

$$\hat{y}_i = \frac{\exp(z_i)}{\sum_j^{\dim(\mathbf{z})} \exp(z_j)} \tag{3}$$

where $\mathbf{y}$ and $\hat{\mathbf{y}}$ are the true label vector and the prediction probability distribution, respectively. The prediction probability $\hat{\mathbf{y}}$ is obtained from Eq 2, where each of the element in $\hat{\mathbf{y}}$ is computed using Eq 3.

B. HANDLING CLASS IMBALANCED PROBLEM

One of the common problems when training a neural network is when a dataset has an imbalanced proportion between the classes, which happens to be the case in this study. The distribution of the sample in the dataset is shown in Table 2. As discussed in Section II-B, several techniques can be used to deal with this situation, either by balancing the sample distribution or incorporating class weights during the training [30]. In this study, it is impractical to perform oversampling on the dataset, especially in the form of natural language, as is the case in other domains where techniques such as synonym replacement or structure re-arrangement are applied. This is due to the nature of the domain, where oversampling the drone flight log messages does not reflect the real-world case and situation. Therefore, incorporating a weighting schema during the training is a feasible option.

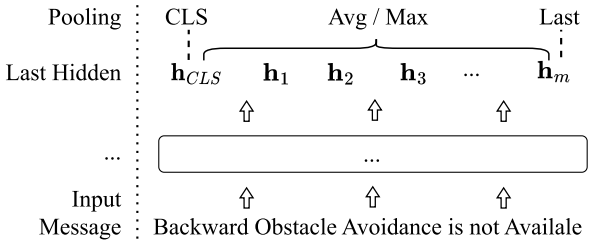


FIGURE 5. An illustration for the pooling mechanism used in the experiment. Note that the *last* pooling is only applicable to the LSTM and GRU encoder.

Preventing a neural model from learning from the majority of samples can be achieved by weighting the loss of each class based on the frequency distribution. Therefore, the importance of the minority class is considered equal [42] to the majority class. Incorporating a class weight into the cross-entropy loss as defined in Eq 1 where weighting the per sample loss is written as the following Eq 4:

$$\mathcal{L}^{\text{CE}^*} = -\sum_{i=1}^{|C|} \alpha_i \cdot y_i \cdot \log(\hat{y}_i) \tag{4}$$

where α_i denotes the weight for the i-th class. Another way of performing weighting to loss function is to penalise the prediction with a low confidence score, called Focal loss [42]. This is achieved by transforming the ratio of the loss value between high-confidence and low-confidence prediction probabilities. Focal loss is computed using the following Eq 5:

$$\mathcal{L}^{\text{Focal}} = -\sum_{i=1}^{|C|} \alpha_i \cdot (1 - \hat{y}_i)^{\gamma} \cdot \log(\hat{y}_i) \tag{5}$$

where γ is a hyperparameter to control the range of the prediction probability value being penalised. Increasing $\gamma > 0$ weakens the relative loss from samples with high prediction confidence scores. Therefore, the model is forced to focus more on hard-to-classify samples [42].

Computing the class weight can be challenging, as the importance of a particular class could be vague relative to the other classes in the label set. A common practice is using frequency-based weighting, where the class weight

	One-Hot Encode	Multitask-110
High	[0 0 0 1]	[0 0 1 1]
Medium	[0 0 1 0]	[0 1 1 0]
Low	[0 1 0 0]	[1 1 0 0]
Normal	[1 0 0 0]	[1 0 0 0]
	(a)	(b)

FIGURE 6. Proposed vector representation of the label based on the dataset characteristics.

is computed based on the frequency distribution of the class in the dataset. In this study, three class weights are explored and used to train all encoder types: uniform, inverse, and balanced. Uniform refers to equal weighting on all classes, which means no weighting is used, while inverse and balanced can be calculated using the Eq 6 and Eq 7, respectively, as in the following:

$$\alpha_c^{\text{inv}} = \left(\frac{|\mathbf{D}_c|}{|\mathbf{D}|}\right)^{-1} \tag{6}$$

$$\alpha_c^{\text{bal}} = \frac{|\mathbf{D}|}{|C| \cdot |\mathbf{D}_c|} \tag{7}$$

where $|\mathbf{D}|$ denotes the total samples in the dataset and $|\mathbf{D}_c|$ represents the number of the samples belong to class c with $\mathbf{D}_c \subset \mathbf{D}$.

C. SEVERITY-ORIENTED LABEL REPRESENTATION

Training a neural network to perform a multiclass classification task typically converts the class names into a one-hot encoding vector, where each vector element represents a particular class which relies on the assumption that the samples from different classes are mutually exclusive [43]. As discussed in Section I, and shown in Fig. 2 and Fig. 3, the samples from distinct classes in the dataset are instead mutually inclusive. Therefore, this study proposes a vector representation of the label inspired by the multitask learning paradigm. Fig. 6 shows the proposed label along with the standard one-hot encoding. The label relies on the assumption that samples belonging to a higher severity level with one class share common low-level features with samples in a lower severity level. We call it a severity-oriented label as the model is trained on two alternative labels instead of one exclusive label only. Therefore, in case of misclassification, the prediction is expected to fall one level under the true label.

Aligning to the nature of the label representation, the loss function used in the training is log loss for multiclass classification, which can be computed using Eq 8, as written in the following:

$$\mathcal{L}^{\text{Log}} = -\sum_{i=1}^{|C|} \alpha_i \cdot [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{8}$$

where α_i is the same term as in Eq 4. With this loss function, the model is forced to learn from the shared features among the two neighbouring classes. This assumption originated from the multi-object detection task, where a single image contains multiple objects to detect, and those different objects share low-level features. Therefore, predicting a certain object benefited from the features of other objects within the same input sample [44]. The same assumption underlies the design for the label representation in this study.

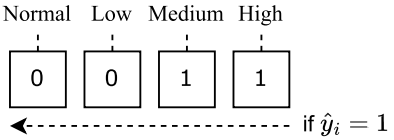


FIGURE 7. Severity-oriented decoding process of the predicted label vector into the class name, which is done in a backward direction.

Inferring the prediction result with the proposed label is different from the typical one-hot encoded label. The index of the element with the highest probability value that is obtained from arg max is used as the final prediction. Then, the integer index is decoded back to the class name using the same function used to convert the class name into the integer index in the label encoding process. Conversely, each vector element in the proposed label represents a particular class, as shown in Fig. 6. Therefore, to normalise the logits from the model into probability, sigmoid is used on each of the prediction vector elements and is written as Eq 9 in the following:

$$\hat{y}_i = \frac{1}{1 + \exp(-z_i)} \tag{9}$$

where $\hat{y}_i$ is the prediction probability for the i-th class and $\hat{y}_i \in [0, 1]$. To get the vector of the prediction label, a prediction threshold $\lambda = 0.5$ is applied, making the element with a value more than or equal to the threshold equal to 1, otherwise it is 0. Finally, severity-oriented decoding is used to get the label class name back by checking the prediction vector from the last element and moving backward to the first element, as illustrated in Fig. 7. If the current element is marked as 1, the class name in that position becomes the predicted class.

IV. EXPERIMENTAL SETUP

This section provides details about the dataset used in the experiment presented herein, as well as describes the baseline models used for performance comparison, the experiment environment, and the hyperparameter settings used.

A. DATASET

In this study, one of the artefacts of drone devices, the flight log message, is used as the source of the evidence to perform multiclass anomaly severity detection. These human-readable messages, such as ''Aircraft Core Board Overheated'' and ''Compass error, calibration required'', can be found in most DJI-make flight logs in three columns: warning, tip, and message. Originally, the flight log was in encrypted.CSV files [45]. Thus, the drone phantom help[2] is needed to decrypt the flight log files and then extract the contents. The

[2]https://www.phantomhelp.com/logviewer/upload

TABLE 2. Summary of the distribution of the class in the dataset.

Dataset	Split	High	Medium	Low	Normal	Total
Filtered	Train	30	104	173	161	468
	Test	8	26	43	40	117
	Total	38	130	216	201	585
Unfiltered	Train	32	632	433	696	1,793
	Test	8	159	108	174	449
	Total	40	791	541	870	2,242

dataset used in this study is also used in [18], where the log messages are collected from two sources: VTO Labs [45] and Drone Wiki [46]. Instead of taking the average over 5 folds, one of the folds (5^{th} fold) is directly used since the standard deviation signified the performance evaluation on 5 different folds is insignificant [18]. The filtered dataset refers to the log messages after performing a unique filtering process, while the unfiltered is the actual collection of drone flight log messages gathered from the two sources mentioned previously. The dataset is split into training and testing with an 80:20 ratio. The summary of the per-class distribution in the dataset is presented in Table 2.

B. BASELINES

In this study, an extensive experiment with various types of encoders is conducted to construct strong baselines for performance comparison. Ensuring the objectivity of the comparison, the hyperparameters of these baselines are fine-tuned using grid search on a finite search space, as shown in Table 3. The details of the baseline construction are described in the following subsections.

1) BASELINE FROM DIVERSE SUITABLE ENCODERS

The distinction between the baseline and the proposed method is based on the pair of the label representation and the loss function used during the training. Two baselines are defined: one-hot encoding with cross-entropy loss and one-hot encoding with focal loss. The proposed scenario employs multitask encoding with log loss. Using these three scenarios, a significant number of models using the following encoders are trained:

1) **None** refers to fine-tuning BERT [16] on the dataset. Different pooling mechanisms and class weighting strategies are explored whilst using this baseline, resulting in 54 scenarios.
2) **Transformer** implies the transformer encoder sub-module in the transformer architecture, proposed in [47]. From the search space in Table 3, bi-directionality is the only irrelevant hyperparameter during the grid search, generating 486 scenarios.
3) **LSTM** [48] and **GRU** [49] allude to recurrent neural network families that are common in sequence classification tasks. In these two models, the number of attention heads is the only irrelevant hyperparameter during the grid search, yielding 324 scenarios each.

TABLE 3. The search space for the hyperparameter tuning.

Variable	Value Space			
	None	Transformer	LSTM	GRU
Loss	Cross-Entropy, Focal, Log			
Num. of Layers	-	1, 2, 3		
Attention Heads	-	4, 6, 8	-	
Pooling	Max, Avg, CLS		Max, Avg, Last	
Class Weight	Uniform, Balanced, Inverse			
Bidirectionality	-		True, False	

Therefore, the overall scenarios are 1,188 in total. BERT is chosen as the embedding model for all scenarios as it is widely used in diverse domains and proven to be better compared to other contextual embedding models [50].

2) BASELINE FROM PREVIOUS WORKS

Among the relevant published studies in the log-based anomaly detection space, sequence-based is one of the most commonly used approaches. Therefore, there are limited relevant references as this study employs a point-based approach. Below are the relevant baselines from previous research:

1) **Pylogsentiment** [9] is the first study which employs sentiment analysis-based anomaly detection on operating system logs using GRU and GloVe embedding.
2) **SentiLog** [26], similar to [9], use a two-layered BiLSTM and GloVe embedding model to perform anomaly detection on parallel file system logs.
3) **TransSentLog** [10] is a further development of [9] and uses a two-layered transformer encoder which used two attention heads and GloVe embedding along with integrated gradients to add explainability to the trained model.
4) **NeuralLog** [25] is a one-layered transformer encoder-based model trained on various log anomaly benchmarks using 12 attention heads and BERT as the embedding. Contrary to the other three baselines, this study performed the detection on the sequence of log records instead of point-based.

BERT has been demonstrated to be a better embedding compared to GloVe [50], as BERT produces a contextual feature vector based on the relationship among the words within a sentence. In this study, BERT is used as the embedding model when reproducing the performance of the baseline. Additionally, hyperparameter tuning is also performed to make the performance comparison as objective and fair as possible.

C. EXPERIMENT SETTINGS

The experiment is conducted on a personal computer equipped with a 16GB VRAM GPU. The method is implemented in Python version 3.10.13 with the help of the Pytorch version 2.0.1 library. The model is trained and tested only once by setting the random seed value to guarantee reproducibility over multiple runs and on different

TABLE 4. The best performing model for each encoder type and loss function based on accuracy and F1 score tested on the unfiltered dataset after performing hyperparameter tuning.

Encoder	Loss	#Layer	#Head	Pooling	Class Weight	Best Epoch	Acc (%)	Pre (%)	Rec (%)	F1 (%)	Conf (μ_σ)
BiGRU*	CE	1	-	Last	Uniform	9	96.429	96.527	96.429	96.470	$0.985_{0.07}$
BiGRU	Focal	2	-	Last	Uniform	12	96.205	96.430	96.205	96.303	$0.968_{0.07}$
BiGRU	Log	3	-	Max	Uniform	4	96.429	96.556	96.429	96.449	$0.984_{0.05}$
LSTM	CE	1	-	Avg	Uniform	12	96.205	96.486	96.205	96.315	$0.995_{0.03}$
LSTM*	**Focal**	**2**	**-**	**Last**	**Uniform**	**10**	**96.429**	**96.604**	**96.429**	**96.498**	$\mathbf{0.965_{0.07}}$
LSTM	Log	1	-	Avg	Balanced	11	96.205	96.535	96.205	96.328	$0.991_{0.05}$
None*	CE	-	-	Avg	Balanced	6	95.759	96.370	95.759	95.985	$0.988_{0.05}$
None	Focal	-	-	Avg	Uniform	15	95.759	95.837	95.759	95.778	$0.974_{0.09}$
None	Log	-	-	Avg	Uniform	7	95.759	95.864	95.759	95.786	$0.993_{0.05}$
Transformer	**CE**	**3**	**6**	**Avg**	**Uniform**	**11**	**96.429**	**96.608**	**96.429**	**96.487**	$\mathbf{0.993_{0.04}}$
Transformer	Focal	3	4	Max	Uniform	10	96.429	96.400	96.429	96.366	$0.967_{0.08}$
Transformer*	**Log**	**1**	**6**	**Avg**	**Balanced**	**7**	**96.875**	**96.856**	**96.875**	**96.851**	$\mathbf{0.995_{0.03}}$

Best-performing model from each loss function
*Best-performing model from each encoder type

devices. During the experiment, as each encoder type has its hyperparameter, the search space is not the same as one another. The details of the search space are presented in Table 3. Other than the ones listed in the table, the hyperparameter values are set as follows: learning rate is $\eta = 2e-5$, 15 epochs, $\gamma = 2$ [42] for the focal loss, and 8 batch size on train and test.

During the training, the best-performing checkpoint is saved and used in the evaluation. The first criterion for choosing the best checkpoint is by comparing both the accuracy and F1 score of the current execution with the previous best run. If the accuracy and F1 score of the current execution is not higher than the previous score, then the second criterion is to check if the F1 score is improved while the accuracy is stagnant. If the second criterion does not hold, it is observed if the accuracy is improved while the F1 score is stagnant. Other than these criteria, the current execution is ignored. The position of the best checkpoint is recorded as the best epoch for performance evaluation.

The accuracy, weighted average precision, recall, and F1 scores are reported as the performance evaluation metrics. Additionally, the mean prediction probability is recorded to measure how confident the model is in predicting the test data on average. To verify the importance of each component in the model, an ablation study is performed on several aspects. To provide an in-depth analysis from a domain-specific perspective, error analyses are also conducted by investigating the the misclassification cases. Finally, the learned representation of the dataset is examined to check if the model successfully learns a decent representation during the training.

V. PERFORMANCE EVALUATION AND DISCUSSION

Following the procedure and details in the previous section, this section reports the experimental results along with in-depth discussions and analyses.

A. BEST PERFORMING SCENARIO

1) EVALUATION METRIC

After experimenting with all the designed scenarios, the models' performance is evaluated on the test dataset. The best-performing model from each pair of encoder type and loss function is reported in Table 4 and Table 5 for unfiltered and filtered datasets. Based on the evaluation metrics shown in both tables, the proposed scenario outperforms all baseline scenarios on all encoder types tested on the unfiltered dataset, with an accuracy of 96.875 and an F1 score of 96.851. As for the filtered dataset, the transformer encoder trained on the baseline scenario outperforms the other scenarios on all encoder types with an accuracy of 83.761 and an F1 score of 82.967. The proposed scenario only performs better on the GRU and LSTM encoders than the baseline scenarios.

2) PREDICTION CONFIDENCE

Evaluating the performance of a neural model strongly depends on the task and domain problem. In a particular task, accuracy can be the only important metric. However, in some other domains, having a high accuracy does not suffice. For example, in this study, the model needs to be sure when predicting if a log is not an anomaly. To check if a model is certain of the prediction, we can investigate the prediction probability. Therefore, we record the prediction probability of each scenario during the testing, taking the mean (μ) and the standard deviation (σ) to analyse the confidence score of the prediction of the model. The distribution of the mean prediction confidence is presented in Fig. 8 based on the loss function. From the figure, it can be observed that the confidence score of log loss is significantly higher and more stable than the other two losses, indicating that the proposed label increases the prediction confidence on all encoder types on average.

3) CONVERGENCE SPEED

The best epoch in the Table 4 and Table 5 indicate the iteration position when the model reaches the best performance from a total of 15 epochs. It can be used to evaluate the training behaviour in different scenarios. From the unfiltered evaluation, the proposed scenario reached the best performance at the 7^{th} iteration, relatively faster than the other two best-performing scenarios, which need 11 and 10 iterations for cross-entropy and focal, respectively. On the contrary, the result of the filtered dataset shows an opposite

TABLE 5. The best performing model for each encoder type and loss function based on accuracy and F1 score tested on the filtered dataset after performing hyperparameter tuning.

Encoder	Loss	#Layer	#Head	Pooling	Class Weight	Best Epoch	Acc (%)	Pre (%)	Rec (%)	F1 (%)	Conf (μ_σ)
GRU	CE	3	-	Avg	Uniform	9	81.197	81.176	81.197	80.863	$0.973_{0.07}$
GRU	Focal	3	-	Avg	Uniform	9	81.197	80.738	81.197	80.723	$0.893_{0.11}$
GRU*	Log	1	-	Avg	Uniform	9	82.051	81.950	82.051	81.429	$0.978_{0.06}$
LSTM	CE	3	-	Avg	Uniform	9	82.051	81.621	82.051	81.325	$0.931_{0.11}$
BiLSTM	Focal	3	-	Last	Inverse	12	82.051	81.816	82.051	81.618	$0.869_{0.11}$
LSTM*	**Log**	**2**	**-**	**Last**	**Inverse**	**13**	**82.051**	**81.699**	**82.051**	**81.859**	$\mathbf{0.976_{0.07}}$
None*	CE	-	-	CLS	Balanced	11	80.342	82.251	80.342	80.873	$0.958_{0.10}$
None	Focal	-	-	CLS	Balanced	7	80.342	80.914	80.342	80.602	$0.833_{0.16}$
None	Log	-	-	Avg	Uniform	14	78.632	79.110	78.632	78.846	$0.975_{0.07}$
Transformer*	**CE**	**3**	**4**	**CLS**	**Uniform**	**13**	**83.761**	**84.699**	**83.761**	**82.963**	$\mathbf{0.970_{0.09}}$
Transformer	**Focal**	**1**	**4**	**CLS**	**Inverse**	**12**	**82.906**	**82.663**	**82.906**	**82.315**	$\mathbf{0.876_{0.15}}$
Transformer	Log	2	8	Max	Balanced	11	81.197	80.756	81.197	80.876	$0.968_{0.09}$

Best-performing model from each loss function
*Best-performing model from each encoder type

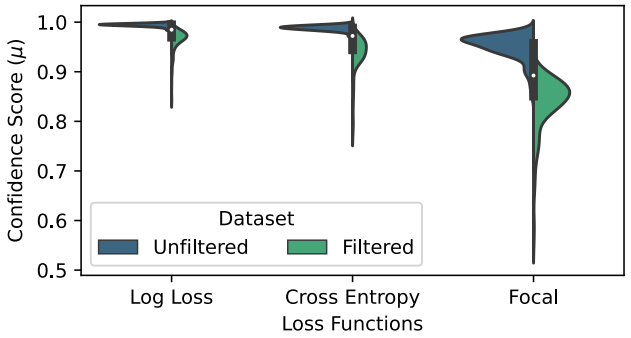


FIGURE 8. The distribution of the mean (μ) prediction confidence score from each loss function. Log loss is significantly higher and more stable than CE and focal on both datasets.

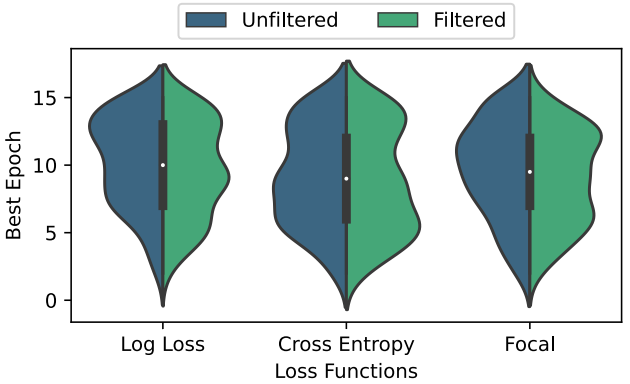


FIGURE 9. The distribution of the best epoch from each loss function. Log loss tends to need more epochs compared to cross-entropy and focal loss.

tendency. The higher-performing scenarios tend to take more iterations than the lower-performing ones. However, the proposed scenario outperforms the other two baseline scenarios on the GRU encoder with the same number of iterations. When using a transformer encoder, the proposed scenario underperforms the other two baselines significantly while taking an almost similar number of iterations. Fig. 9 shows the distribution of the best epoch on each loss function, where log loss tends to take more epochs to reach the best checkpoint.

B. HYPERPARAMETER ANALYSIS ON EACH ENCODER

Considering the total number of experimented scenarios, it is not possible to discuss all the details in this paper. Therefore, we perform a chi-square test to determine which hyperparameter is significant towards the accuracy of the models. Note that this test measures the difference in the mean of accuracy among different groups based on the categorical value in each hyperparameter. Thus, the significant hyperparameter can be different from one encoder to another. Fig. 10 shows the test result, where the heat map colour indicates the statistical test score. The exact test score is shown for each pair where the p-value is less than 0.05, which is the significant value. Chi-square can be computed using $\chi^2 = \sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ where O_{ij} and E_{ij} are the observed and expected frequencies in cell (i, j) in a contingency table constructed from each hyperparameter paired with the accuracy. Since the accuracy is a continuous variable, the value is converted into several groups by binning the value into several ranges. Having the chi-square test result presented in Fig 10, it can be seen that the significant hyperparameters differ from one encoder to another. Note that this test does not reflect the direction of the dependency. Instead, it shows which hyperparameter the accuracy depends on. The dependency can be either increasing or decreasing the accuracy. Nevertheless, it can help choose which hyperparameter to modify in the next experiments.

C. COMPARISON WITH STATE-OF-THE-ART MODELS

In this study, several baselines from previous studies are reproduced on the dataset to perform performance comparisons. The reproduced performance from an experimental procedure explained in subsection IV-B2 is presented in Table 6. On both datasets, the proposed scenario outperforms all the baselines with an improvement of 0.423 and 0.665 in the F1 score on the filtered and unfiltered datasets, respectively. Considering that Pylogsentiment used a GRU-based encoder and SentiLog used an LSTM-based encoder, the proposed scenarios on these two encoders

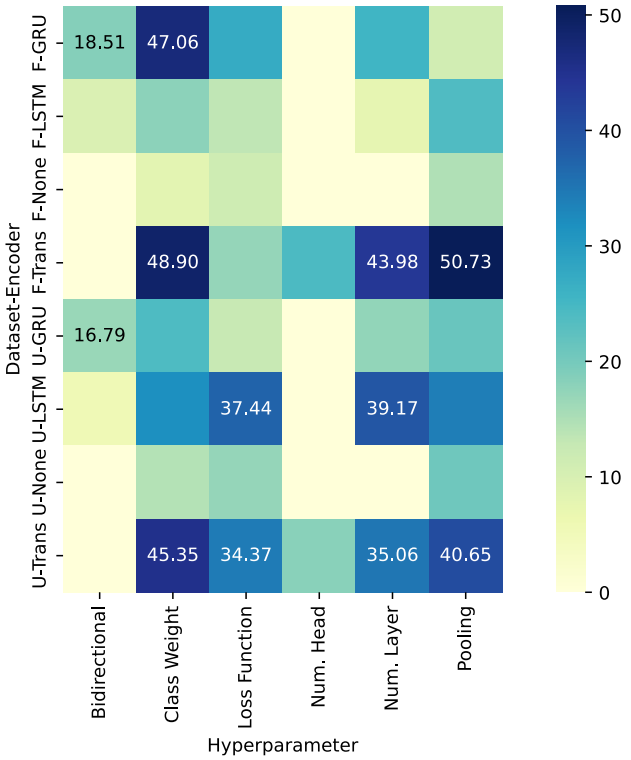


FIGURE 10. A significance test using the chi-square test of independence to check which hyperparameter is significant towards the accuracy. Only those with p-values less than 0.05 are shown.

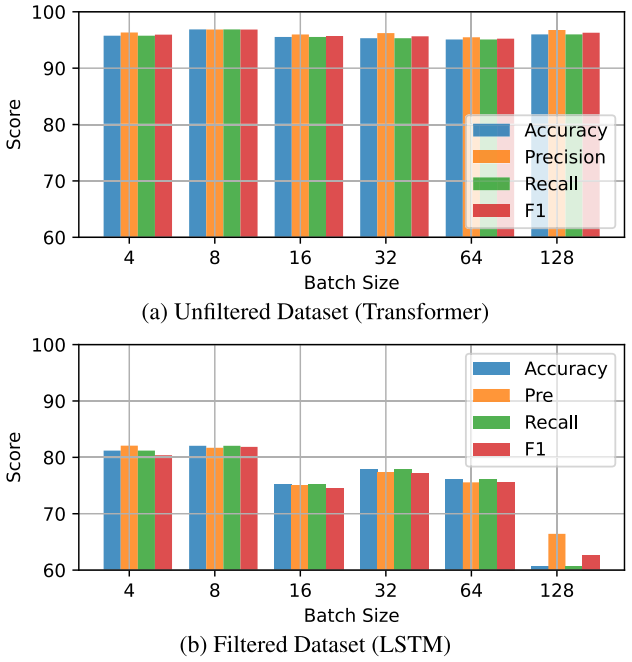


FIGURE 11. Analysis of the effect of different batch sizes on the evaluation scores of the best-performing models.

consistently outperform the performance of the baselines tested on the filtered dataset. Note that we also perform hyperparameter tuning on these baselines to ensure the validity of the performance comparison. As for transformer-based baselines, which are NeuralLog and TransSentLog, the proposed scenario achieves lower performance when tested on the filtered dataset. However, the transformer-based proposed scenario achieves the highest performance on the unfiltered dataset. Based on these findings, it is verified that the proposed scenario can improve the detection performance of the model.

D. ABLATION STUDY

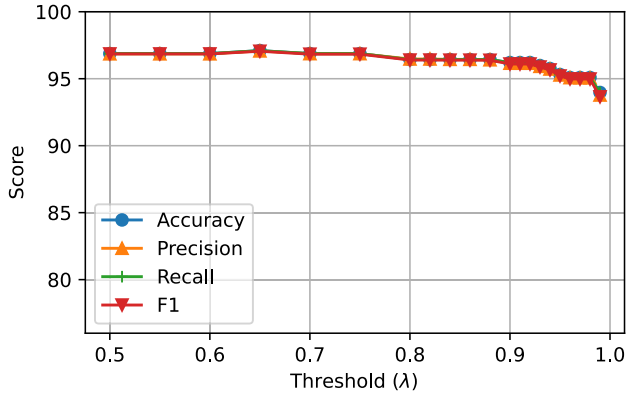
Following the rapid advancement in neural network research, more complex and sophisticated architectures are emerging. A model can consist of layers of components that have a specific role in the learning process. After conducting an experiment on a certain dataset and task, verifying and checking which component in the model contributes positively to the performance is crucial. Therefore, we perform an ablation study to explain which part of the proposed approach has a significant impact on the performance. Note that only the best-performing scenario on each dataset is explored. First, we investigate the significance of the CLS token from the BERT embedding. Secondly, we analyse the impact of freezing the BERT's parameter during training. Thirdly, we vary the batch size during the training. Finally, we examine the effect of increasing the prediction threshold used in our proposed approach on the accuracy and F1 scores.

As shown in Table 6, excluding the CLS token embedding before feeding the input matrix to the encoder significantly decreases performance. Also, freezing the BERT's parameter during the training caused a striking drop in the performance scores.

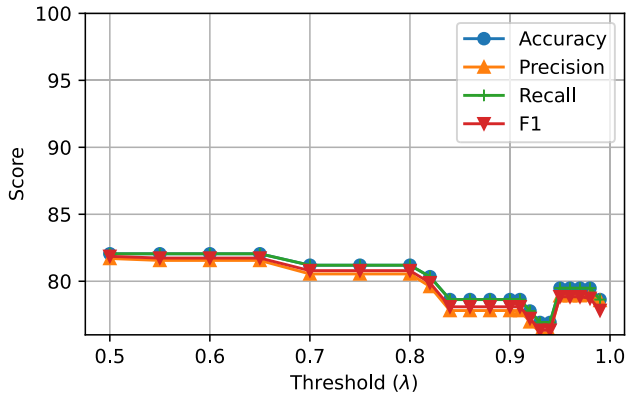
Training a neural model using a batched sample is a common practice to shorten the training time and enforce the model to learn from representative samples. The size of a batch can have an impact on the performance of a model. As shown in Fig 11, varying the batch size has a notable impact on the filtered dataset but insignificant on the unfiltered dataset. From the figure, it can be seen that the 8 is the best option to use. During the inference, when using multitask encoding and severity-oriented decoding, the threshold plays a crucial role in the prediction evaluation. Fig 12 shows the effect of increasing the threshold on the accuracy and F1 scores of the best-performing model tested on both datasets. An interesting case is shown in Fig 12a when $\lambda = 0.65$, the accuracy reaches 97%, exceeding the best model's accuracy when $\lambda = 0.5$. This happens when the prediction confidence score surpasses the threshold while the true label's severity is lower than the prediction label's. Thus, following the decoding procedure in section III-C, the predicted label is incorrect. This phenomenon also happens on the filtered dataset when $\lambda > 0.94$. Overall, the decrement in the accuracy and F1 scores is insignificant until $\lambda = 0.9$ and $\lambda = 0.8$ on the unfiltered and filtered datasets, respectively. Thus, it confirms the prediction confidence distribution that is depicted in Fig 8.

TABLE 6. Performance comparison with several baselines from previous works.

Ref.	Filtered Dataset					Unfiltered Dataset				
	Acc (%)	Pre (%)	Rec (%)	F1 (%)	Conf (μ_σ)	Acc (%)	Pre (%)	Rec (%)	F1 (%)	Conf (μ_σ)
Pylogsentiment [9]	80.342	81.400	80.342	80.604	$0.944_{0.11}$	95.536	95.575	95.536	95.534	$0.988_{0.07}$
SentiLog [26]	79.487	79.621	79.487	79.458	$0.968_{0.08}$	95.759	95.962	95.759	95.842	$0.993_{0.04}$
NeuralLog [25]	82.051	81.506	82.051	81.436	$0.819_{0.17}$	96.205	96.208	96.205	96.186	$0.985_{0.07}$
TransSentLog [10]	81.197	81.120	81.197	81.127	$0.957_{0.11}$	95.982	96.289	95.982	96.104	$0.989_{0.05}$
DroLoVe (GRU)	82.051	81.950	82.051	81.429	$0.978_{0.06}$	96.429	96.556	96.429	96.449	$0.984_{0.05}$
DroLoVe (LSTM)	**82.051**	**81.699**	**82.051**	**81.859**	$\mathbf{0.976_{0.07}}$	96.205	96.535	96.205	96.328	$0.991_{0.05}$
w/o CLS vector	78.632	77.957	78.632	78.206	$0.978_{0.06}$	95.536	96.047	95.536	95.717	$0.996_{0.03}$
freeze BERT's params	64.103	71.658	64.103	66.186	$0.864_{0.15}$	89.063	94.805	89.063	91.414	$0.976_{0.07}$
DroLoVe (Transformer)	81.197	80.756	81.197	80.876	$0.968_{0.09}$	**96.875**	**96.856**	**96.875**	**96.851**	$\mathbf{0.995_{0.03}}$
w/o CLS vector	78.632	78.491	78.632	78.107	$0.952_{0.10}$	95.759	96.096	95.759	95.877	$0.992_{0.04}$
freeze BERT's params	70.940	70.038	70.940	70.391	$0.845_{0.15}$	93.080	94.284	93.080	93.541	$0.972_{0.08}$



(a) Unfiltered Dataset (Transformer)



(b) Filtered Dataset (LSTM)

FIGURE 12. Analysis of the effect of increasing the prediction threshold on the evaluation scores of the best-performing models.

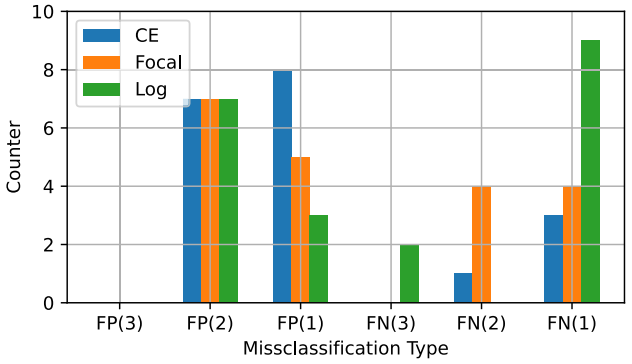


FIGURE 13. The number of misclassified samples from the best-performing model on each loss function tested on the filtered dataset.

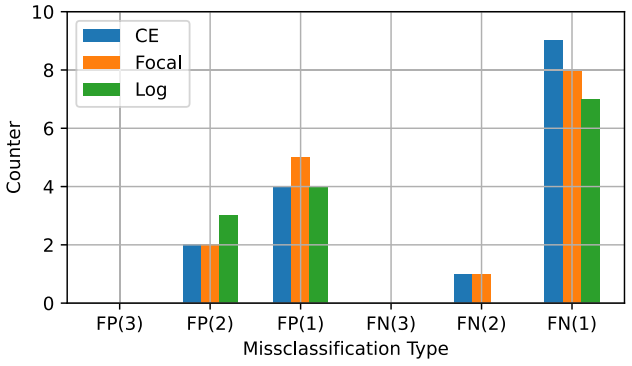


FIGURE 14. The number of misclassified samples from the best-performing model on each loss function tested on the unfiltered dataset.

TABLE 7. Custom confusion matrix for error analysis.

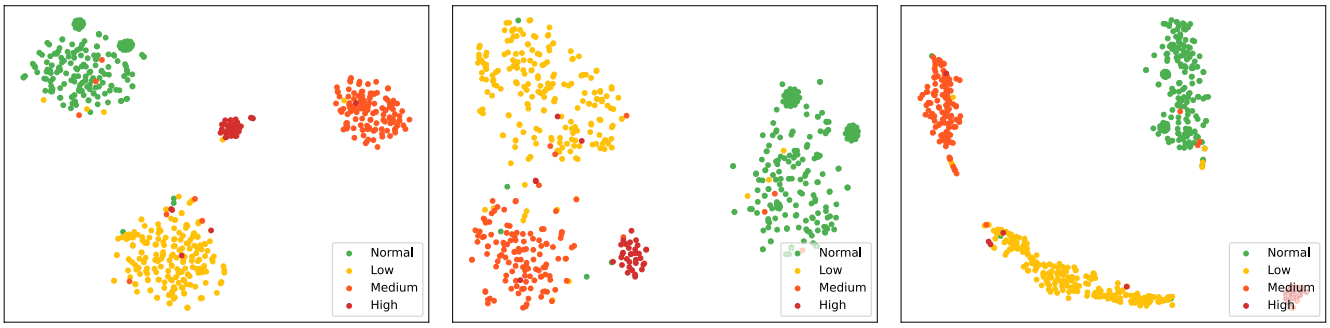
	Normal	Low	Medium	High	
Normal	TP	FN(1)	FN(2)	FN(3)	True Class
Low	FP(1)	TP	FN(1)	FN(2)	
Medium	FP(2)	FP(1)	TP	FN(1)	
High	FP(3)	FP(2)	FP(1)	TP	
	Predicted Class				

E. ERROR ANALYSIS

In the previous section, the performance evaluation has been discussed thoroughly from a quantitative perspective. Here, we investigate the samples from the test set that are predicted incorrectly by the best-performing model from each scenario. In a binary setting, let the Normal class be positive and the Anomaly is negative. Then, False Positive (FP) refers to samples with a true label negative but predicted as positive. At the same time, a False Negative (FN) is a case when a sample belongs to a positive class but is predicted as negative. However, in this study, we define the misclassification cases differently:

- **False Positive** refers to a case when the true class is higher than the predicted class.
- **False Negative** is a misclassification case where the true class is lower than the predicted class.

Table 7 shows the position of FP and FN based on the above definition in a multiclass confusion matrix.



(a) Cross Entropy (b) Focal (c) Log Loss

FIGURE 15. A 2D visualisation of the filtered dataset representation obtained from the best model from each loss function tested on the filtered dataset. t-SNE [17] is used to reduce the dimension of the embeddings.

In an anomaly detection setting, FP is more important than FN since detecting anomalous events as normal is a critical error. Having a high number of FPs can endanger the system with a serious impact, especially for FP(3) cases, which means the true label is High and the predicted label is Normal. At the same time, having a high number of FNs can cause many false alarms, but they are not as critical as FPs. Following the definition in the previous paragraph, we plot the frequency of each misclassification type in Fig 13 and Fig 14. From the figures, it can be seen that the proposed approach performs better on the unfiltered dataset despite having a high number of FN(1) and FN(3). However, on the filtered dataset, the proposed approach has more FP(2) than the other two scenarios, even though it has a smaller number of FP(1), FN(1), and FN(2).

Other than analysing the misclassified samples, we also analyse the learned representation of the dataset by each of the best-performing models. The filtered dataset is used to investigate the learned representation plotted into 2D graphs, as shown in Fig 3. From the visualisation, it can be seen that the samples that belong to High class are still close to the Normal class in Fig 15a and Fig 15b. In Fig 15c, the High class is far from the Normal class but even further from the Medium class. Even though all scenarios result in a well-separated representation, none of the scenarios reflects the ordinal nature of the label. One would expect the High class to be far from the Low class, representing the distance between the severity of the anomalies.

VI. CHALLENGES, LIMITATIONS, AND THREATS TO VALIDITY

In the previous sections, the performance of the proposed framework has been discussed and analysed. In this section, challenges encountered during the study are disclosed, along with limitations and threats to validity. As drone forensics is an emerging topic, there are very limited public datasets available. The log messages used in this study are mainly acquired from DJI-made devices. There is yet to come a publicly available dataset of log messages from other drone manufacturers. Several publicly available datasets are mainly about sensor data and multimedia artefacts [51]. While in this study, we solely depend on the human-readable messages generated by the drone during a flight. Given the condition where a small number of unique messages and most of them are acquired from DJI drones, the generality of the proposed model remains untested. Considering the performance evaluation score, where the highest accuracy is under 85%, the model needs further improvement so that the detection validity can be enhanced and that the detection results of the model can be convincing and accountable enough to the investigator to be included in the investigation report.

VII. CONCLUSION AND FUTURE WORK

In this paper, we demonstrate how to train a log-based anomaly detection model that can prioritise the prediction of higher severity anomalies on drone flight logs while increasing the prediction confidence score at the same time. Considering the nature of the dataset, where samples that belong to different severity levels share common features, a multitask vector label representation along with severity-oriented decoding is proposed.

An extensive experiment proved that the proposed approach is better than the previous work baselines while being inferior to our baseline scenario based on the accuracy and F1 scores. From the anomaly detection perspective, the proposed method achieves higher prediction confidence and can prioritise higher severity levels during the inference on the test dataset. Despite the promising results, the proposed model is tested only on messages acquired from DJI-made devices, leading to untested generality. Moreover, the testing set does not reflect an incident scenario, making the resulting model untested in a real-case environment.

Further future studies may include exploring the possibility of doing high-level oversampling to introduce the models with more log message patterns, producing a dataset with incident scenarios to perform case studies to verify the performance of the proposed method, and making it publicly available.

DATA AVAILABILITY

The dataset used in the experiment is available on a reasonable request. The code for the experiment, the resulting performance evaluation, the scripts for data analysis, and the figures in this paper are made publicly available on GitHub (https://github.com/swardiantara/DroLoVe) to promote transparent, reproducible, and verifiable research.

REFERENCES

[1] W. Van Der Aalst, A. Adriansyah, A. K. A. De Medeiros, F. Arcieri, T. Baier, T. Blickle, J. C. Bose, P. van den Brand, R. Brandtjen, and J. Buijs, "Process mining manifesto," in *Proc. Bus. Process Manage. Workshops*, in Lecture Notes in Bus. Information Processing, 2011, pp. 169–194.

[2] E. Mantas and C. Patsakis, "Who watches the new watchmen? The challenges for drone digital forensics investigations," *Array*, vol. 14, Jul. 2022, Art. no. 100135.

[3] A. B. Nassif, M. A. Talib, Q. Nasir, and F. M. Dakalbab, "Machine learning for anomaly detection: A systematic review," *IEEE Access*, vol. 9, pp. 78658–78700, 2021.

[4] X. Zhao, Z. Jiang, and J. Ma, "A survey of deep anomaly detection for system logs," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2022, pp. 1–8.

[5] M. Landauer, S. Onder, F. Skopik, and M. Wurzenberger, "Deep learning for anomaly detection in log data: A survey," *Mach. Learn. Appl.*, vol. 12, Jun. 2023, Art. no. 100470.

[6] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review," *ACM Comput. Surv.*, vol. 54, no. 2, pp. 1–38, 2021.

[7] X. Ma, J. Wu, S. Xue, J. Yang, C. Zhou, Q. Z. Sheng, H. Xiong, and L. Akoglu, "A comprehensive survey on graph anomaly detection with deep learning," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 12, pp. 12012–12038, Jun. 2023.

[8] J. Li, H. He, S. Chen, and D. Jin, "LogGraph: Log event graph learning aided robust fine-grained anomaly diagnosis," *IEEE Trans. Dependable Secure Comput.*, pp. 1–15, Jul. 2023, doi: 10.1109/TDSC.2023.3293111.

[9] H. Studiawan, F. Sohel, and C. Payne, "Anomaly detection in operating system logs with deep learning-based sentiment analysis," *IEEE Trans. Dependable Secure Comput.*, vol. 18, no. 5, pp. 2136–2148, Sep. 2021.

[10] T.-A. Pham and J.-H. Lee, "TransSentLog: Interpretable anomaly detection using transformer and sentiment analysis on individual log event," *IEEE Access*, vol. 11, pp. 96272–96282, 2023.

[11] S. Silalahi, T. Ahmad, and H. Studiawan, "Transformer-based sentiment analysis for anomaly detection on drone forensic timeline," in *Proc. 11th Int. Symp. Digit. Forensics Secur. (ISDFS)*, May 2023, pp. 1–6.

[12] Z. Zhao, C. Xu, and B. Li, "A LSTM-based anomaly detection model for log analysis," *J. Signal Process. Syst.*, vol. 93, no. 7, pp. 745–751, Jul. 2021.

[13] M. Memarzadeh, B. Matthews, and T. Templin, "Multiclass anomaly detection in flight data using semi-supervised explainable deep learning model," *J. Aerosp. Inf. Syst.*, vol. 19, no. 2, pp. 83–97, Feb. 2022.

[14] F. Shahzad, A. Mannan, A. R. Javed, A. S. Almadhor, T. Baker, and D. Al-Jumeily, "Cloud-based multiclass anomaly detection and categorization using ensemble learning," *J. Cloud Comput.*, vol. 11, no. 1, p. 74, Nov. 2022.

[15] G. O. Anyanwu, C. I. Nwakanma, J. M. Lee, and D.-S. Kim, "Novel hyper-tuned ensemble random forest algorithm for the detection of false basic safety messages in Internet of Vehicles," *ICT Exp.*, vol. 9, no. 1, pp. 122–129, Feb. 2023.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[17] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[18] S. Silalahi, T. Ahmad, and H. Studiawan, "Drone flight log anomaly severity classification via sentence embedding," in *Proc. Int. Conf. Artif. Intell., Blockchain, Cloud Comput., Data Anal. (ICoABCD)*, Nov. 2023, pp. 100–105.

[19] J. Zhou, Y. Qian, Q. Zou, P. Liu, and J. Xiang, "DeepSyslog: Deep anomaly detection on syslog using sentence embedding and metadata," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 3051–3061, 2022.

[20] L. K. Shar, W. Minn, N. B. D. Ta, J. Fan, L. Jiang, and D. L. W. Kiat, "DronLomaly: Runtime detection of anomalous drone behaviors via log analysis and deep learning," in *Proc. 29th Asia–Pacific Softw. Eng. Conf. (APSEC)*, Dec. 2022, pp. 119–128.

[21] C. Zhang, X. Wang, H. Zhang, J. Zhang, H. Zhang, C. Liu, and P. Han, "LayerLog: Log sequence anomaly detection based on hierarchical semantics," *Appl. Soft Comput.*, vol. 132, Jan. 2023, Art. no. 109860.

[22] J. Qi, Z. Luan, S. Huang, C. Fung, H. Yang, H. Li, D. Zhu, and D. Qian, "LogEncoder: Log-based contrastive representation learning for anomaly detection," *IEEE Trans. Netw. Service Manage.*, vol. 20, no. 2, pp. 1378–1391, Jul. 2023.

[23] X. Li, P. Chen, L. Jing, Z. He, and G. Yu, "SwissLog: Robust anomaly detection and localization for interleaved unstructured logs," *IEEE Trans. Dependable Secure Comput.*, vol. 20, no. 4, pp. 2762–2780, Oct. 2023.

[24] T. Xiao, Z. Quan, Z.-J. Wang, Y. Le, Y. Du, X. Liao, K. Li, and K. Li, "Loader: A log anomaly detector based on transformer," *IEEE Trans. Services Comput.*, vol. 16, no. 5, pp. 3479–3492, Sep. 2023.

[25] V.-H. Le and H. Zhang, "Log-based anomaly detection without log parsing," in *Proc. 36th IEEE/ACM Int. Conf. Automated Softw. Eng. (ASE)*, Nov. 2021, pp. 492–504.

[26] D. Zhang, D. Dai, R. Han, and M. Zheng, "SentiLog: Anomaly detecting on parallel file systems via log-based sentiment analysis," in *Proc. 13th ACM Workshop Hot Topics Storage File Syst.*, Jul. 2021, pp. 86–93.

[27] S. Lupton, H. Washizaki, N. Yoshioka, and Y. Fukazawa, "Literature review on log anomaly detection approaches utilizing online parsing methodology," in *Proc. 28th Asia–Pacific Softw. Eng. Conf. (APSEC)*, Dec. 2021, pp. 559–563.

[28] J. Singh and S. Gupta, "Evaluating the impact of local data imbalance on federated learning performance for IoT anomaly detection," in *Proc. 14th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)*, Jul. 2023, pp. 1–7.

[29] F. Kong, J. Li, B. Jiang, H. Wang, and H. Song, "Integrated generative model for industrial anomaly detection via bidirectional LSTM and attention mechanism," *IEEE Trans. Ind. Informat.*, vol. 19, no. 1, pp. 541–550, Jan. 2023.

[30] S. Das, S. S. Mullick, and I. Zelinka, "On supervised class-imbalanced learning: An updated perspective and some key challenges," *IEEE Trans. Artif. Intell.*, vol. 3, no. 6, pp. 973–993, Dec. 2022.

[31] N. T. Anh, L. H. Hoang, V. D. Minh, and T. H. Hai, "BKIDset—A new intrusion detection dataset to mitigate the class imbalance problem," in *Proc. 15th Int. Conf. Adv. Comput. Appl. (ACOMP)*, Nov. 2021, pp. 106–111.

[32] H. Studiawan and F. Sohel, "Performance evaluation of anomaly detection in imbalanced system log data," in *Proc. 4th World Conf. Smart Trends Syst., Secur. Sustainability*, Jul. 2020, pp. 239–246.

[33] X. Ma and W. Shi, "AESMOTE: Adversarial reinforcement learning with SMOTE for anomaly detection," *IEEE Trans. Netw. Sci. Eng.*, vol. 8, no. 2, pp. 943–956, Apr. 2021.

[34] M. M. Rashid, F. Sabrina, B. Ray, A. Morshed, S. Gordon, and S. Wibowo, "Anomaly detection in IoT applications using deep learning with class balancing," in *Proc. IEEE Asia–Pacific Conf. Comput. Sci. Data Eng. (CSDE)*, Dec. 2022, pp. 1–6.

[35] T. Sutthipanyo, T. Lamsan, W. Thawornsusin, and W. Susutti, "Log-based anomaly detection using CNN model with parameter entity labeling for improving log preprocessing approach," in *Proc. IEEE Region 10 Conf. (TENCON)*, Oct. 2023, pp. 914–919.

[36] O. Elghalhoud, K. Naik, and M. Zaman, "Data balancing and CNN based network intrusion detection system," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2023, pp. 1–6.

[37] J. Qi, Z. Luan, S. Huang, Y. Wang, C. Fung, H. Yang, and D. Qian, "Adanomaly: Adaptive anomaly detection for system logs with adversarial learning," in *Proc. IEEE/IFIP Netw. Oper. Manage. Symp.*, Apr. 2022, pp. 1–5.

[38] T. Al-Shehari, M. Al-Razgan, T. Alfakih, R. A. Alsowail, and S. Pandiaraj, "Insider threat detection model using anomaly-based isolation forest algorithm," *IEEE Access*, vol. 11, pp. 118170–118185, 2023.

[39] P. Chand, M. Moh, and T.-S. Moh, "An approach to improving anomaly detection using multiple detectors," in *Proc. 16th Int. Conf. Ubiquitous Inf. Manage. Commun. (IMCOM)*, Jan. 2022, pp. 1–8.

[40] S. Yan, S. Wang, Z. Chen, X. Jiang, and X. Cao, "CSLog: Anomaly detection for syslog based on contrastive self-supervised representation learning," in *Proc. 24th Asia–Pacific Netw. Oper. Manage. Symp. (APNOMS)*, 2023, pp. 165–170.

[41] X. Ma, J. Keung, P. He, Y. Xiao, X. Yu, and Y. Li, "A semisupervised approach for industrial anomaly detection via self-adaptive clustering," *IEEE Trans. Ind. Informat.*, vol. 20, no. 2, pp. 1687–1697, Feb. 2024.
[42] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.
[43] P. Cerda, G. Varoquaux, and B. Kégl, "Similarity encoding for learning with dirty categorical variables," *Mach. Learn.*, vol. 107, nos. 8–10, pp. 1477–1494, Sep. 2018.
[44] X. Zhou, Y. Gao, C. Li, and Z. Huang, "A multiple gradient descent design for multi-task learning on edge computing: Multi-objective machine learning approach," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 1, pp. 121–133, Jan. 2022.
[45] S. Silalahi, T. Ahmad, and H. Studiawan, "DroNER: Dataset for drone named entity recognition," *Data Brief*, vol. 48, Jun. 2023, Art. no. 109179.
[46] (2023). *Drone Error and Warning Codes—World's Most Comprehensive List*. [Online]. Available: https://app.airdata.com/wiki/Notifications/
[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6000–6010.
[48] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[49] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1724–1734.
[50] C. Wang, P. Nulty, and D. Lillis, "A comparative study on word embeddings in deep learning for text classification," in *Proc. 4th Int. Conf. Natural Lang. Process. Inf. Retr.*, Dec. 2020, pp. 37–46.
[51] H. Studiawan, G. Grispos, and K.-K.-R. Choo, "Unmanned aerial vehicle (UAV) forensics: The good, the bad, and the unaddressed," *Comput. Secur.*, vol. 132, Sep. 2023, Art. no. 103340.

SWARDIANTARA SILALAHI (Graduate Student Member, IEEE) received the bachelor's degree in informatics education, in 2021, and the master's degree in informatics, in 2023. He is currently pursuing the Ph.D. degree in computer science with the Institut Teknologi Sepuluh Nopember (ITS), Indonesia. His research interests include digital forensics, log mining, natural language processing, and deep learning.

TOHARI AHMAD (Member, IEEE) received the bachelor's degree in computer science from Institut Teknologi Sepuluh Nopember (ITS), Indonesia, the master's degree in information technology from Monash University, Australia, and the Ph.D. degree in computer science from RMIT University, Australia, in 2012.

From 2001 to 2003, he was a Consultant with some international companies. In 2003, he moved to ITS, where he is currently a Professor. His research interests include network security, information security, data hiding, and computer networks.

Prof. Ahmad is a member of ACM. His awards and honors include the Hitachi Research Fellowship and the JICA Research Program to conduct research in Japan. He is a reviewer of a number of journals.

HUDAN STUDIAWAN (Member, IEEE) received the bachelor's and master's degrees from the Institut Teknologi Sepuluh Nopember, Indonesia, in 2009 and 2011, respectively, and the Ph.D. degree from Murdoch University, Australia, in 2021. He is currently a Lecturer with the Institut Teknologi Sepuluh Nopember. His current research interests include digital forensics and natural language processing.

EIRINI ANTHI is currently a Lecturer in cybersecurity with the School of Computer Science and Informatics, Cardiff University. She teaches operating systems security and cybersecurity operations. As a part of the Ph.D. degree, she developed state-of-the-art tools to detect and defend against network-based cyber-attacks in such infrastructures. In addition, her research interests include the security of the Internet of Things (IoT) and industrial control systems (ICS). More particularly, her research examines the security issues that come along with these devices/systems and focuses on developing intelligent and more robust cyber-attack detection mechanisms for such networks using machine learning and adversarial machine learning techniques.

LOWRI WILLIAMS is currently a Lecturer with the School of Computer Science and Informatics, Cardiff University. Her work focuses on the development of novel approaches toward automated cyber defense. In particular, her interest is in how to apply text mining and machine learning techniques in defense methodologies and different cybersecurity contexts.