IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 17, 2021, accepted December 14, 2021, date of publication December 20, 2021,
date of current version December 28, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3136505

Robust Anomaly Detection for Multivariate Data of Spacecraft Through Recurrent Neural Networks and Extreme Value Theory

GANG XIANG[1,2] AND RUISHI LIN[2]
[1]School of Automation and Electrical Engineering, Beihang University, Beijing 100191, China
[2]Beijing Aerospace Automatic Control Institute, Beijing 100854, China

Corresponding author: Gang Xiang (xianggang@buaa.edu.cn)

This work was supported in part by the National Basic Scientific Research under Grant JCKY2016203A003, and in part by the Equipment Pre-Research Key Laboratory Funds under Grant 61425010102.

ABSTRACT Spacecraft anomaly detection which could find anomalies in the telemetry or test data in advance and avoid the occurrence of catastrophic failures after taking corresponding measures has elicited the attention of researchers both in academia and aerospace industry. Current spacecraft anomaly detection systems require costly knowledge and human expertise to identify a true anomaly. Moreover, some new problems and challenges such as large volume of test data, imbalanced data distribution and the scarcity of faulty labeled samples have emerged. In this work, we propose an unsupervised anomaly detection algorithm combining Gated Recurrent Unit (GRU) based Recurrent Neural Network (RNN) and Extreme Value Theory (EVT). First, we develop a two-layer ensemble learning based predictor framework which stacks three GRU-based networks with different architectures to learn and capture the normal behavior of multiple channels of data. Then, the prediction errors are calculated and smoothed using Exponentially Weighted Moving Average (EWMA) algorithm. Next, we propose a detection rule setting anomaly threshold automatically through EVT which does not assume any parent distribution on the prediction errors. To the best of our knowledge, it is the first attempt that stacked GRU-based predictors with EVT has been employed into the spacecraft anomaly detection. Through extensive experiments conducted on public datasets as well as real data sampled from a launch vehicle, we show that the proposed detection algorithm is superior to other state-of-the-art anomaly detection approaches in terms of model performance and robustness.

INDEX TERMS Anomaly detection, GRU, extreme value theory, RNN, spacecraft.

I. INTRODUCTION

Spacecraft is a very sophisticated and complicated system which consists of many subsystems and components such as structure system, engine, control system and telemetry system. It is essential to monitor the health status and strengthen the stability and reliability of spacecraft because it often operates in the harshest outer space environment. Anomaly detection strives to discovery patterns and data that do not adapt to expected behaviors [1], [2]. Anomalies which are unexpected instance data greatly deviating from normal behaviors defined by the most of a dataset usually do not occur abruptly and there exists subtle and imperceptible changes in telemetry data of spacecraft [3]. Anomaly detection technology should detect these changes and anomalies in advance to further prevent the potential cascading downtime which may result in catastrophic and unforeseen damages to spacecraft. Therefore, anomaly detection plays a significant role in real-time managing health status and promoting reliability of spacecraft [4], [5].

At present, one of the most used anomaly detection methods is Out-Of-Limit (OOL) [6], [7], where a pre-defined threshold is set and once a monitored value strays outside of this threshold, and then an anomaly is detected. Although OOL methods are simple and easy to implement, they require large amount of domain knowledge and human expertise to set thresholds and are difficult to identify contextual anomalies when the values are less than the pre-defined threshold. Therefore, in the recent years there have been many anomaly detection algorithms which make great improvements over

The associate editor coordinating the review of this manuscript and approving it for publication was Yuan Zhuang.

OOL, such as density-based approaches [8], [9], statistical methods [10], [11], k-nearest neighbor methods [12], [13] and Support Vector Machine (SVM) based methods [14], [15]. However, these methods perform well only when the size of data is small. With the missions of spacecraft are becoming more and more complex, the volumes of telemetry data increase rapidly. For instance, the Synthetic Aperture Radar satellite generates nearly 85 terabytes of data per day [16]. Therefore, it becomes almost unlikely for the aforementioned anomaly detection methods to adapt to such big telemetry data to identify anomalies in the real or near real time.

Another challenge for spacecraft anomaly detection is that the telemetry data are quite imbalanced [4], which means that there are large number of normal samples but insufficient or even no labeled abnormal samples in telemetry dataset, resulting from the fact that spacecraft usually works in normal condition and failures seldom occur. It is time-consuming and labor-intense to label thousands of telemetry data, which can lead traditional approaches to overfit and not be able to detect anomalies accurately. Moreover, most of these anomaly detection methods are single channel models designed to only be able to detect anomalies for each telemetry channel individually. However, a single model may not perform well for all the telemetry channels, and it is much more reasonable and preferred to detect anomalies using multivariate channels rather than using univariate channel for several reasons. First, it needs more hardware resources and time to train and maintain models for each telemetry channel. Second, in real application scenarios, engineers are encouraged to concern more about the status of spacecraft as a whole than each channel.

In this research, in order to mitigate and balance the challenges mentioned above, we propose an unsupervised deep learning-based anomaly detection model using Gated Recurrent Units (GRU) [17] based Recurrent Neural Networks (RNNs) [18], [19] and Extreme Value Theory (EVT) [20], [21] to identify anomalies in telemetry data of spacecraft. Our core idea is to forecast the time series of spacecraft by designing a two-layer ensemble learning based predictor framework with GRU, and subsequently apply EVT rule on prediction errors. Our approach is an unsupervised algorithm which tries to learn and capture the normal behavior of multiple channels data by leveraging superior performance of GRU. The bigger a prediction error is, the more likely it will be regarded as an anomaly. The main contributions of this paper are summarized as follows:

1) We propose an unsupervised anomaly detection algorithm. To the best of our knowledge, this algorithm is the first multivariate time series anomaly detection method which combines GRU based neural networks and EVT in spacecraft.

2) We propose an ensemble learning based predictor framework based on diverse GRUs and is robust to multiple channels of telemetry data.

3) We introduce EVT to process the prediction errors and propose a dynamic threshold setting method without making any critic assumptions about the prediction error distribution.

4) We conduct extensive experiments and compare our approach with other state-of-the-art anomaly detection methods on different dataset. Experimental results demonstrate that our method achieve the best performance among all the baseline methods.

The remainder of this paper is organized as follows. Related researches are reviewed in Section 2. Section 3 details the proposed anomaly detection methods. Experimental results and discussions are conducted in Section 4. Finally, conclusions are drawn in Section 5.

II. RELATED WORK

The telemetry data of spacecraft are time series, and the anomalies could be divided into three categories – point, contextual, and collective [22]. Point anomalies are observations that differ significantly from the majority of data; contextual anomalies are regarded as anomalous only in a specific context and collective anomalies demonstrate that a set of observations are anomalies as a whole, but any single value in that sequence may be not by itself. Since contextual and collective are much more difficult to detect and both require temporal information, we merge them into the contextual category in this article.

Several anomaly detection approaches have been explored and exploited for spacecraft, among which expert-systems [6], [7], [23], [24] are the simplest and most widely used methods. The notable expert-system called Intelligent Satellite Control Software DOCtor (ISACS-DOC) is successfully implemented in Hayabusa, Geotail, and Nozomi missions [24]. The Out-Of-Limits (OOL) [6], [7] method using a pre-defined threshold to detect anomalies is a simple form of expert-system and still be adopted by a large number of spacecraft due to low computation expense, ease of interpretability, and easy applicability. Envelop Learning and Monitoring using Error Relaxation (ELMER) [25] is an improved OOL method which uses neural network to periodically set new threshold bounds. However, it is usually hard for expert-system to detect contextual anomalies and it requires sufficient and accurate knowledge of spacecraft [16]. A myriad of machine learning methods which automatically learn and extract knowledge from telemetry data have been introduced to spacecraft anomaly detection, such as k nearest neighbors-based methods [12], [13], cluster-based methods [26], [27], Support Vector Machine (SVM) based methods [14], [15], [28], and Artificial Neural Networks (ANN) based methods [4], [16], [29], [30], among others. The k nearest neighbor-based method is used by XMM-Newton satellite [6] as well as the Space and the International Space Station [31]. Cluster based methods which identify values falling outside of well-defined clusters as anomalous are adopted by NASA. The Centre National d'Etudes Spatiales (CNES) uses One-Class Support Vector Machine (OS-SVM) to separate anomalies from nominal data [32]. Although machine learning methods show great improvement over expert-system based approaches, each has its own shortcomings

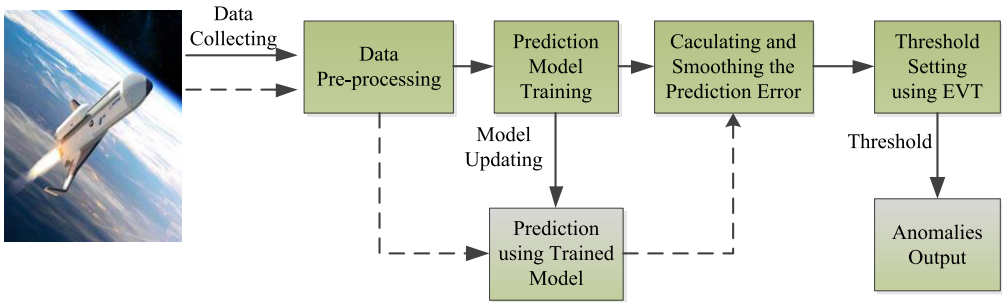


FIGURE 1. The overall structure of the proposed method.

related to computational expense, model complexity, or generalizability.

Recently, given the rapid advancement of computing capacity and neural network architecture, deep learning has begun to be applied in anomaly detection in spacecraft with high-dimension and large volumes of data. Su *et al.* [30] proposes a model called DAGMM which reduces the dimension of input data to extract hidden characteristics using Autoencoder (AE) and estimates the density of the characteristics using Gaussian Mixture Model (GMM). Recurrent neural networks as well as its variants Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU) have shown powerful capability to maintain memory of long-term dependencies and model complex nonlinear feature interactions. Therefore, LSTM and GRU are especially suitable for anomaly detection with time-series [16]. Malhotra *et al.* [33] used LSTM to construct a predictor to model normal behavior of time series, and subsequently use the prediction error to detect anomalous behaviors. The LSTM-based anomaly detection method in [33] gave better detection results than other methods, but this method assumed that the prediction errors must fit a normal distribution. In [16], Hundman *et al.* proposed a LSTM-based predictor and a nonparametric anomaly thresholding approach without assuming the distribution of prediction errors to detect telemetry data of the Soil Moisture Active Passive (SMAP) satellite and the Mars Science Laboratory (MSL) rover. The main disadvantage of this paper is that it is a single channel model and we should train many models for multiple telemetry channels which will cost more training time and require more computation expense. Tariq *et al.* proposed an anomaly detection algorithm using a multivariate convolution LSTM with mixtures of probabilistic principal component analyzers and demonstrated the effectiveness of the method with the real Korea Multi-Purpose Satellite 2 (KOMPSAT-2) data in [34].

Literatures in [4], [16] are prediction-based anomaly detection algorithms which usually apply several rules on the prediction errors to identify anomalies. In [33], the prediction errors from the training data were assumed to fit a Gaussian distribution of which the mean, μ, and variance, σ^2 are computed using the Maximum Likelihood Estimation (MLF). Then an anomaly score is calculated and a low score indicates that the corresponding observation may be an anomaly with relatively high probability. In literature [35], it models the prediction error distribution as a normal distribution, and applies a threshold to the Gaussian tail probability (Q-function) to decide whether or not to declare an anomaly. However, the assumption that prediction errors follow Gaussian distributions does not always hold. Therefore, in [16], a novel dynamic anomaly threshold is set that, when all the observations above the thresholds are removed, would lead to the largest percent decrease in mean and standard deviation, but this method would cause many false positives.

III. METHOD

A. OVERALL STRUCTION

Spacecraft usually consists of many subsystems, each of which generates several channels of telemetry data. Therefore, it is very difficult to detect the root causes of anomalies due to the correlation among these channels. Moreover, a single predictor may not perform well for other channels. Hence, in this section, we design a multivariate telemetry channels anomaly detection model.

The overall structure of our proposed anomaly detection algorithm is illustrated in Fig.1, and it includes two steps: model training and detection. The model training procedure is often conducted offline at an acceptable frequency, such as weekly or monthly. First, we leverage the Z-score normalization to process the raw telemetry dataset and then divide the dataset into training data and test data in the preprocessing module. Second, the training data are used to train the ensemble learning based prediction framework which learn and capture the normal behaviors of multivariate time series. Third, the prediction errors are calculated and smoothed. Then, the threshold setting module automatically sets thresholds for prediction errors using EVT. Once the model training procedure is completed, the well-trained model is stored for the detection procedure. The test data could be fed into the model online and outputs prediction errors. We use the EVT method to decide whether a test observation is anomalous or not.

B. DATA PREPROCESSING

In this study, we mainly observe multivariate time series $X = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$, where each step $x^{(t)} \in R^m$ in the time t ($1 \leq t \leq N$) is an m-dimensional vector $\{x_1^{(t)}, x_2^{(t)}, \ldots, x_m^{(t)}\}$, in which each element corresponds to a specific output channel of spacecraft, and the total number of output channels is m. The time series is shown in Fig. 2, where each row in the matric represents the data of a single channel.

The time series X is transformed into a sequence $X' = \{X_{l_b}, X_{l_b+1}, \ldots, X_{N-l_b-l_a}\}$, where l_b denotes the length of historical values used to predict the next l_a values in the time t. The formats of inputs and outputs are shown in Fig. 2. Then, we split the data into training and test datasets after conducting the Z-score normalization to process the input data.

$$X_t = \left\{ \begin{bmatrix} x_1^{(t-l_b+1)} \\ x_2^{(t-l_b+1)} \\ x_3^{(t-l_b+1)} \\ \vdots \\ x_m^{(t-l_b+1)} \end{bmatrix}, \ldots, \begin{bmatrix} x_1^{(t-1)} \\ x_2^{(t-1)} \\ x_3^{(t-1)} \\ \vdots \\ x_m^{(t-1)} \end{bmatrix}, \begin{bmatrix} x_1^{(t)} \\ x_2^{(t)} \\ x_3^{(t)} \\ \vdots \\ x_m^{(t)} \end{bmatrix} \right\} y_t = \left\{ \begin{bmatrix} x_1^{(t+1)} \\ x_2^{(t+1)} \\ x_3^{(t+1)} \\ \vdots \\ x_m^{(t+1)} \end{bmatrix}, \begin{bmatrix} x_1^{(t+2)} \\ x_2^{(t+2)} \\ x_3^{(t+2)} \\ \vdots \\ x_m^{(t+2)} \end{bmatrix}, \ldots, \begin{bmatrix} x_1^{(t+l_a)} \\ x_2^{(t+l_a)} \\ x_3^{(t+l_a)} \\ \vdots \\ x_m^{(t+l_a)} \end{bmatrix} \right\}$$

FIGURE 2. The input and output of data.

C. GRU-BASED PREDICTOR

LSTMs first proposed in 1990s [18] can learn long-dependencies and contextual reasoning in time series and overcome the vanishing gradients problem by replacing an ordinary neuron by the LSTM unit or block; hence, LSTMs have become popular in forecasting problems in time series [4], [16], [21]. However, it is quite difficult to train LSTMs due to their complex architecture. In order to address this shortcoming, the LSTM is improved, and the input gate and the forgetting gate in LSTM are replaced by an update gate to form a new GRU [17], [36]. The performance of GRU is as good as LSTM while the structure of GRU is simpler than that of LSTM, which is shown in Fig. 3. Therefore, we apply GRU in our predictor model to learn the temporal dependence in multivariate telemetry data.

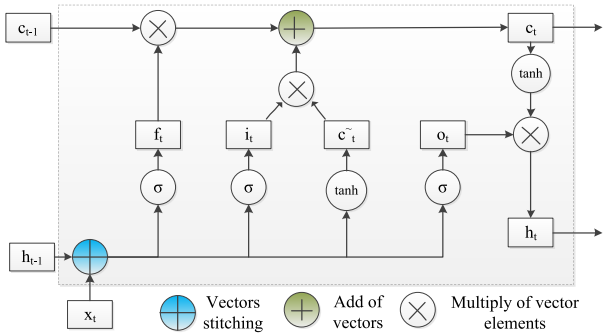
In Fig. 3b), the x_t and h_{t-1} are the inputs while the hidden variable h_t is the output, and the GRU can be formulated as follows:

$$h_t = z_t \odot h_{t-1} + (1 - z_t) tanh \odot (W_h x_t + U_h (r_t \odot h_{t-1}) + b_h) \quad (1)$$

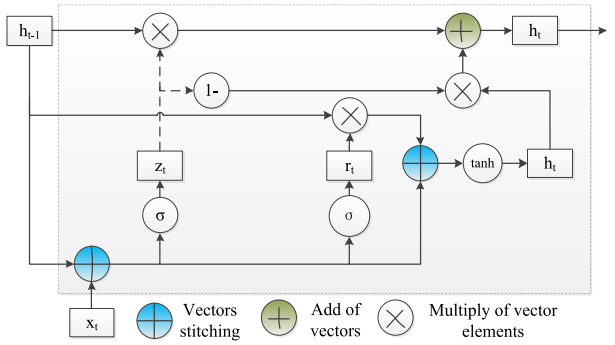
$$r_t = sigmoid(W_r x_t + U_r h_{t-1} + b_r) \quad (2)$$

$$z_t = sigmoid(W_z x_t + U_z h_{t-1} + b_z) \quad (3)$$

where r_t, z_t represents the reset gate and the update gate, respectively. $W_*, U_*, b_*, * \in \{h, r, z\}$ denote the parameters to learn.



a) Architecture of LSTM unit



b) Architecture of GRU unit

FIGURE 3. Architecture of LSTM and GRU.

In the former applications, a specific predictor is trained for each channel and each predictor is used to forecast values for that corresponding channel. This is feasible when the number of channels is small, or channels with analogous data distribution. However, in spacecraft, there usually have thousands of telemetry channels, so it is impractical and infeasible to train a prediction model for each channel due to the computation, storage and time constraints. Therefore, inspired by ensemble learning, we propose a GRU-based stacked predictor in this section, and the framework is illustrated in Fig. 4.

The proposed stacked predictor consists of two layers: a prediction layer and a decision layer. The prediction layer contains 3 GRU-based predictors with various GRU-based

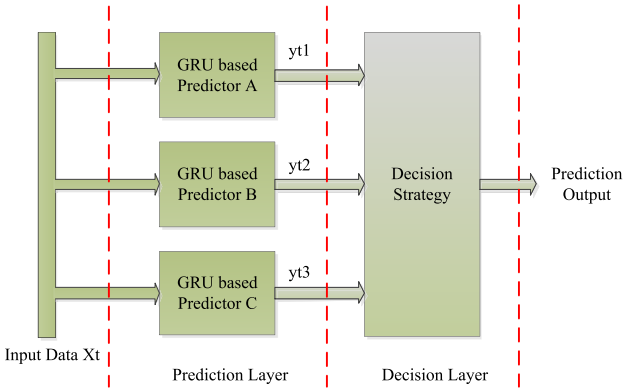


FIGURE 4. Architecture of stacked predictor.

neural network structures and aims to learn the normal behaviors of telemetry data. The input data are fed into three GRU-based predictors simultaneously, and the outputs of each predictor are what the values are expected to be at the next several timestamps.

The decision layer combines the predictions of the multiple GRU models in order to get a more accurate prediction for a dataset. In this paper, the decision strategy gets the means of all the predicted values produced by each individual model as final prediction values. This layer is essential as it enables the proposed predictor to be a generic and robust model in the sense that it does not rely on a specific prediction model. Our proposed model employs various different models to promote the prediction accuracy. Through this way, the stacked predictor is robust to multivariate channels of data.

D. AUTOMATIC PREDICTION ERROR THRESHOULD SETTING USING EVT

The stacked predictor is trained on the training dataset without any abnormal samples, so it learns and captures the normal behavior of the telemetry data. The prediction errors $e(t)$ are defined as the absolute differences between the ground truth observations y_t and their corresponding predicted values $\hat{y}_t$ at time t.

$$e(t) = \left| y_t - \hat{y}_t \right| \tag{4}$$

All the prediction errors in training dataset form a vector $\mathbf{e} = [e(1), e(2), \ldots, e(t), \ldots, e(n)]$. Since there are abrupt changes in the output channels of spacecraft, there exist spikes in prediction errors; therefore, it is necessary and useful to dampen these spikes using the Exponentially Weighted Moving Average (EWMA) method [37]. The smoothed prediction errors are calculated in (5).

$$e_s(t) = \beta e_s(t-1) + (1-\beta)\mathbf{e}(t) \tag{5}$$

In order to find the anomalies, we set the anomaly threshold using the EVT. The goal of EVT is to find the law of extreme values in $e_s(t)$ rather than the data distribution of $e_s(t)$, so it makes no assumptions about the data distribution of prediction errors when finding extreme values. These extreme laws have the following mathematical form

$$G_\gamma : x \longmapsto \exp\left(-(1+\gamma x)^{-\frac{1}{\gamma}}\right), \quad \gamma \in \mathbb{R}, \quad 1+\gamma x > 0. \tag{6}$$

The above equation is called Extreme Value Distribution (EVD), in which the parameter γ is the extreme value index determined by original distribution. In order to evaluate the probability of anomalies using EVD in telemetry data, the parameter γ should be estimated. The Peaks-Over-Threshold (POT) algorithm is an efficient and effective method to compute γ.

Assuming that the cumulative distribution function of the random variable X is F(x), if and only if a function σ exists, for all $x \in R$ s.t. $1 + \gamma x > 0$:

$$\begin{aligned} \bar{F}_\theta(x) &= 1 - F_\theta(x) \\ &= P(X - \theta > x | X > \theta) \sim (1 + \frac{\gamma x}{\sigma(\theta)})^{-\frac{1}{\gamma}} \end{aligned} \tag{7}$$

where θ is the initial threshold and selected as the 98% quantile, γ and σ are shape and scale parameters of a Generalized Pareto Distribution (GPD). This result shows that the excesses over a threshold θ, denoted as $X - \theta$, are likely to follow GPD. The Peaks-Over-Threshold (POT) approach attempts to fit a GPD to the excesses $X - \theta$. Similar to [20], the value of parameters $\hat{\gamma}$ and $\hat{\sigma}$ could be estimated by Maximum Likelihood Estimation (MLE). Therefore, the final threshold ε is computed by:

$$\varepsilon \simeq \theta + \frac{\hat{\gamma}}{\hat{\sigma}} \left[\left(\frac{qn}{N_\theta} \right)^{-\hat{\gamma}} - 1 \right] \tag{8}$$

where q represents the desired probability and usually lies between 10^{-5} and 10^{-3}, n denotes the total number of values, N_θ is the number of peaks, i.e., the number of X_i s.t. $X_i > \theta$.

The EVT, by using the POT algorithm, enables us set dynamic threshold such that P(X $>$ ε) $<$ q without any strong assumption on the original distribution of X and without any clear knowledge about this distribution. We apply this method to detect anomalies in the telemetry data of spacecraft. We compute a first threshold ε_{tr} in the smoothed prediction errors from training dataset. This training procedure is illustrated in the algorithm 1.

Algorithm 1 Set a First Threshold $\boldsymbol{\varepsilon_{tr}}$ in Training Dataset

Input: The smoothed prediction errors of training dataset $\mathbf{e_{s-tr}}$, the risk probability $\boldsymbol{q}$.

Begin:

1: Set the initial threshold $\boldsymbol{\theta}$;

2: Calculate the number of values N_θ excess over $\boldsymbol{\theta}$;

3: Calculate the estimated values of $\hat{\gamma}$ and $\hat{\sigma}$ using MLE;

4: Calculate threshold $\boldsymbol{\varepsilon_{tr}}$ using equation 4;

5: Return $\boldsymbol{\varepsilon_{tr}}$ and $\boldsymbol{\theta}$

End

Then for the prediction errors of the test dataset, we could find whether the values are anomalous or not. If a value in $\mathbf{e_{s-te}}$ exceeds the threshold $\boldsymbol{\varepsilon_{tr}}$, then we consider the corresponding input data as abnormal. The anomalies are not used to update the model. In other cases, either the value exceeds the initial threshold (peak case) either it is a normal value (normal case). In the peak case, we update the threshold. This procedure is shown in Algorithm 2.

IV. EXPERIMENTS

In this work, we are the first to introduce GRU-based stacked predictor as well as extreme value theory into the anomaly detection of spacecraft telemetry data. All experiments are based on the environment of TensorFlow 2.0, python 3.6+.

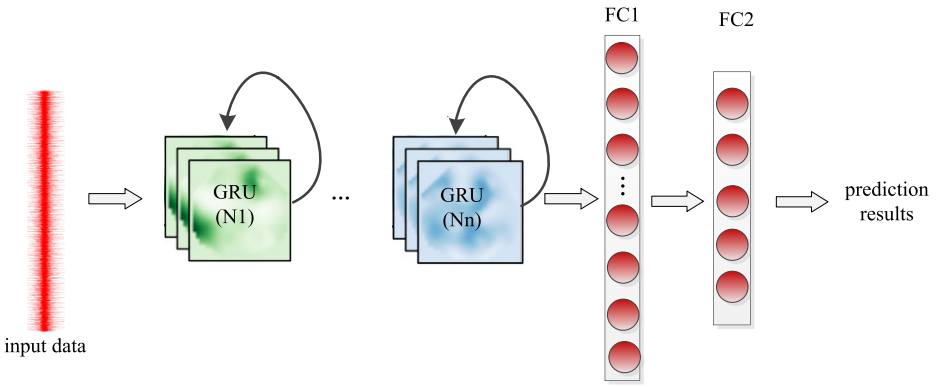


FIGURE 5. The general architecture of GRU-based predictor.

Algorithm 2 Anomaly Detection Method

Input: The smoothed prediction errors of test dataset $\boldsymbol{e_{s-te}}$, $\boldsymbol{\varepsilon_{tr}}$, $\boldsymbol{q}$, $\boldsymbol{\theta}$

Begin:

1: Initialize the anomalous set $\boldsymbol{A} \in \emptyset$;
2: for $\boldsymbol{k} > \mathbf{0}$ do
3: $\boldsymbol{\varepsilon_{te}} = \boldsymbol{\varepsilon_{tr}}$
4: if $\mathbf{e_{s-te}}(\boldsymbol{k}) > \boldsymbol{\varepsilon_{te}}$ then
5: the corresponding input data are anomalous and add them into set A
6: else if $\mathbf{e_{s-te}}(\boldsymbol{k}) > \boldsymbol{\theta}$ then
7: update $\boldsymbol{\varepsilon_{te}}$ using Algorithm1
8: else
9: $\boldsymbol{k} = \boldsymbol{k} + \mathbf{1}$;
10: end if
11: end for
12: Return $\boldsymbol{A}$

End

TABLE 1. The hyper parameters setting of proposed model.

No.	Hyper Parameters	Value
1	Units in FC1, FC2	64, 5
2	l_b, l_a	200, 5
3	dropout	0.5
4	Batch size	64
5	training epochs	48
6	optimizer	Adam
7	loss function	MSE
8	Activation function of FC	linear
9	Learning rate	10^{-3}

A. DATASETS AND EVALUATION METRICS

In order to compare other anomaly detection methods with our proposed model and demonstrate the overall effectiveness of our proposed model, we first conduct several experiments on two public datasets released by NASA Jet Propulsion Laboratory [16]. These two public datasets consist of telemetry data with labels from the Soil Moisture Active Passive (SMAP) satellite and Mars Science Laboratory (MSL) rover, Curiosity. There are 55 and 27 channels in SMAP and MSL datasets, respectively. Then we implement experiments on three public datasets to show the superior prediction performance of stacked GRU models. Finally, a dataset collected from a real launch vehicle is used to elaborate verify the effectiveness of EVT-based detection rule.

Similar to [3], [16], we select Precision, Recall, F1-Score (denoted as F1) as metrics to evaluate the performance of our proposed method and other baseline approaches. The Precision, Recall and F1 are defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

where TP is the true positive values, FP represents false positive values, and FN denotes false negative samples.

B. EXPERIMENTAL SETTINGS

As described in section 2.3, a stacked predictor which contains 3 different architectures of GRU-based neural networks is used to forecast the telemetry time series data of spacecraft. The general architecture is deep which consists of several GRU layers and two fully-connected lays, as shown in Fig. 5. Before training, we employ the Tree-structured Parzen Estimator (TPE) Bayesian Optimization [38] to select the hyper parameters of our proposed model. The GRU A network contains 2 hidden GRU layers with 32, 64 neurons, respectively, the GRU B network contains 3 hidden GRU layers with 80, 80, 64 neurons, respectively, and the GRU C network has 4 hidden GRU layers with 48, 80, 80, 64 neurons, respectively. The rest of hyper parameters are listed in Table 1.

The output of the second fully connected layer is the prediction of the next l_a timestamps and in our model we set l_a to 5 considering balance between training efficiency and

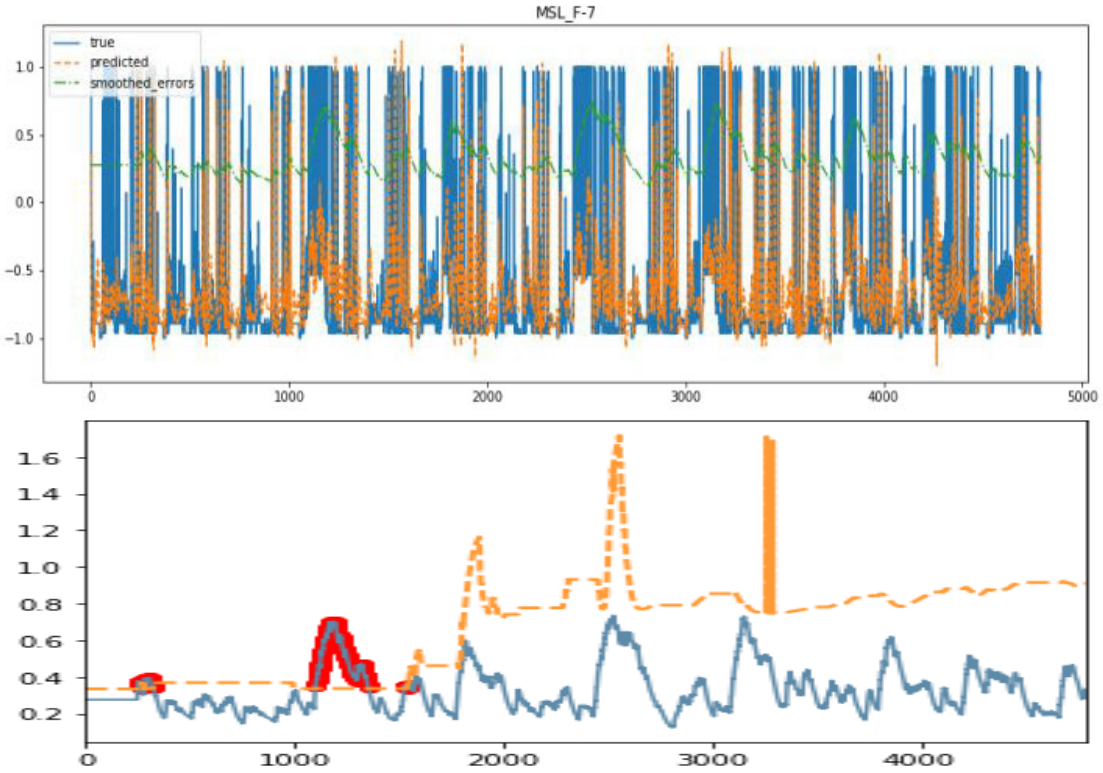


FIGURE 6. Anomaly detection results for channel F-7.

model performance. The number of units in the first fully connected layer is set to 64. Early stopping and dropout strategy are used to avoid over-fitting. The parameters of the model are trained by adopting Adam optimizer [39] for stochastic gradient descent with an initial learning rate of 10^{-3} to minimize the Mean Square Error (MSE) loss function.

For the POT parameters, we follow the recommendation in [20] that the range of q is $\left[10^{-5}, 10^{-3}\right]$ and typically the initial threshold θ is the 98% quantile.

C. RESULTS ON SMAP AND MSL

1) THE OVERALL RESULTS

Our proposed anomaly detection method is a prediction-based algorithm which combines GRU-based predictors and EVT to identify anomalies. Fig. 6 and Fig. 7 illustrate the detection results for channel F-7 from MSL dataset and channel S-1 from SMAP dataset, respectively. The first plots in Fig. 6 and Fig. 7 contain the true data, the predicted data and the smoothed errors. The dash lines in the second plots of Fig. 6 and Fig. 7 represent the dynamic thresholds while the blue lines denote the smoothed errors, and once the smoothed error exceeds the thresholds an anomalous point is labeled. Our proposed approach could detect all the 423 anomalies with 34 false positives in channel F-7 and all the 448 anomalies without any false positive in channel S-1. Therefore, the precision and recall in channel F-7 are 92.26% and 100% while the precision and recall in channel S-1 are both 100%.

In order to demonstrate the effectiveness and superiority of our proposed algorithm, we compare it with three state-of-the-art methods for time series anomaly detection: Hierarchical Temporal Memory (HTM) model [35], LSTM-based model with nonparametric anomaly thresholding approach (LSTM-NAT for abbreviation) [16], Stacked predictor with dynamic thresholding algorithm (SP-DTA for short) [3]. The results of precision, recall, F1 of these baseline models and the proposed method on SMAP and MSL are showed in Table 2 and Fig. 8. The precision and F1 metrics of our method outperform all baseline methods, and the recall metrics are only lower than the best baseline approach on SMAP dataset. Our approach gets the highest F1 scores with 87.4% which exceeds the best baseline approach by 3.6% on MSL dataset. All these results show the superiority of our proposed method compared to the state-of-the-art methods. Then, we analyze the performance of these approaches in detail.

HTM [35] which operates in real-time is a theoretical framework for time series learning in the cortex. HTM outputs the value of current input and computes the prediction error and likelihood. Then this likelihood of an anomaly L which is defined using the tail probability Q determines whether an anomaly is identified. $L = 1 - Q(\frac{\mu_s - \mu_W}{\sigma_W^2})$ where μ_s denotes a short time average, μ_W and σ_W^2 represents mean and variance of a time window W, respectively. This method usually imposes a strong assumption on the prediction errors

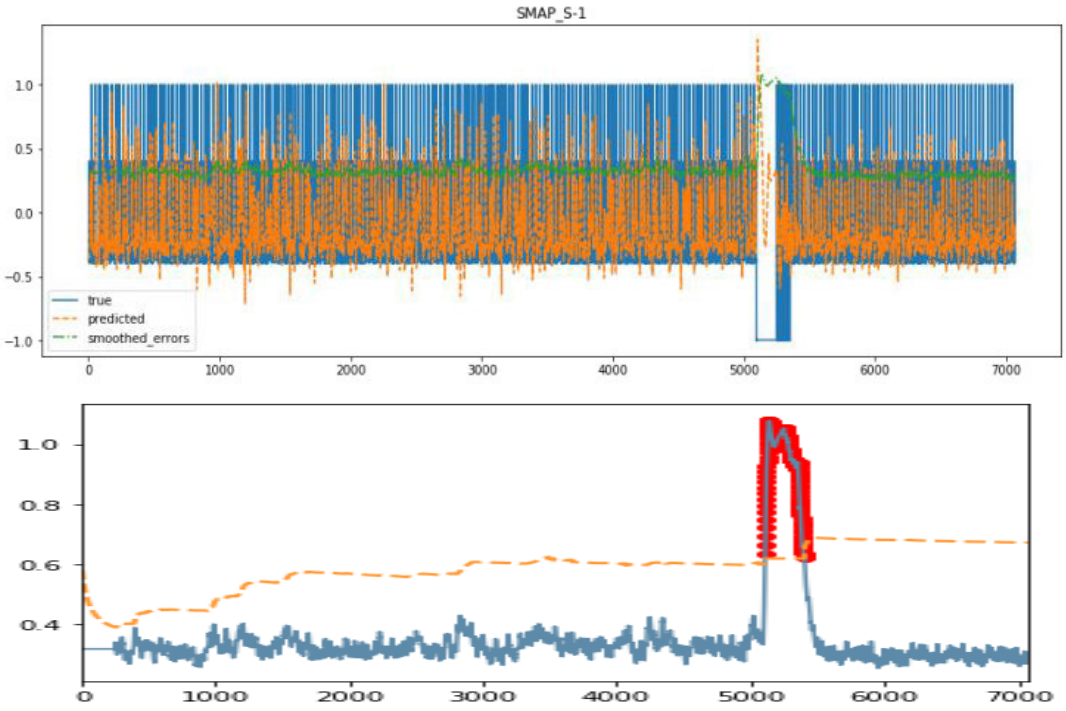


FIGURE 7. Anomaly detection results for channel S-1.

TABLE 2. Performance comparison of baseline models with proposed method.

Method	SMAP			MSL		
	Precision	Recall	F1 score	Precision	Recall	F1 score
HTM[35]	92.7%	73.9%	82.2%	88.2%	41.7%	56.6%
LSTM-NAT[16]	85.5%	85.5%	85.5%	**92.6%**	69.4%	79.3%
SP-DTA[3]	90.0%	91.3%	90.6%	81.6%	**86.1%**	83.8%
Our propose method	88.2%	**94.4%**	**91.2%**	90.2%	84.8%	**87.4%**

that they should follow a normal distribution which does not always hold in real application scenarios. As a result, HTM method does not perform well on these two datasets with the lowest F1 scores (56.6% for MSL) among all the methods.

Hundman *et al.* in [16] proposed a LSTM-based network to detect anomalies with a complementary unsupervised and nonparametric anomaly thresholding approach in telemetry data from SMAP and MSL spacecraft. One of disadvantages of this method is that it could not detect contextual anomalies well, which explains why this method gets the lowest precision 85.5% for SMAP. Another limitation is that a specific model needs to be created for each telemetry channel, thus it would need to train many models for various telemetry channels, which is infeasible and impractical in real application due to computation constrains.

Stacked predictor in [3] stacks three LSTM-based predictors with different architectures and a Support Vector Machine (SVM) based predictor to forecast the input values, then a dynamic thresholding algorithm is applied to optimally extract anomalous data. This method could enhance generalization and adaptability, so it is suitable for multivariate telemetry channels. However, this method will need more time to search the best thresholds for the prediction error sequences than other approaches and it is difficult to determine the optimal bandwidth of Gaussian kernel. This explains its worse performance than our proposed method.

Our proposed method is a prediction model which promotes the forecasting performance by adopting stacked GRU-based predictor. Unlike HTM and SP-DTA, our method applies no assumption on the probability distribution of prediction errors and employs EVT to automatically set threshold. Compared with LSTM, the architecture of GRU is simpler while the performance of GRU is as good as LSTM; therefore, the GRU-based model proposed is more efficient and effective than other baseline models. Moreover, our model trains on the entire dataset instead of training a specific model for each channel, which great promotes efficiency and reduce training time.

2) EFFECT OF GRU

The prediction-based anomaly detection algorithms usually consist of two steps. First, a well-designed prediction model

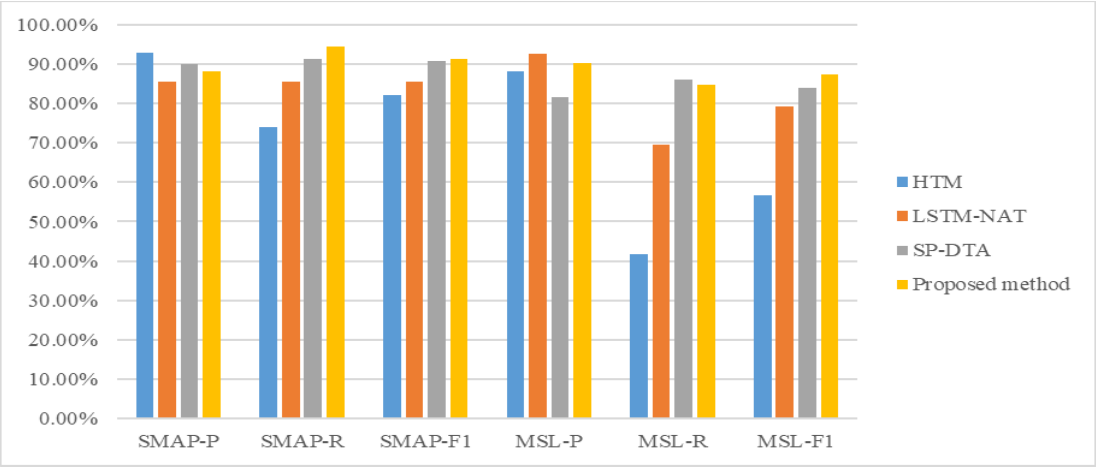


FIGURE 8. Metrics of proposed method and all baseline models.

is constructed to learn the normal representatives of time series data and forecast the future time series. Then a specific detective rule is performed to identify anomalous time series on the prediction errors. Thus, training an effective and efficient predictive model is vital because it directly determines the prediction errors. Considering the superior performance of RNN in processing temporal or sequential data, in this section, the prediction performances of GRU and LSTM which are two main variants of RNN are compared in two dimensions of prediction accuracy on the test set and training time on the training set. We select commonly used datasets in the field of anomaly detection for comparative analysis: Numenta Machine Temperature dataset [40], Power Demand dataset [41], and ECG dataset [42]. For fairness, the GRU and LSTM based networks have the same architecture and hyper parameters. The experimental results are shown in Table 3 below.

From the experimental results in Table 3, we can conclude that the prediction accuracy of the prediction model implemented with GRU is slightly higher than that of LSTM. When using a predictive model for anomaly detection, it is usually necessary to compromise between prediction accuracy and detection accuracy. A high prediction accuracy on the test set often results in low detection accuracy. This is because the prediction-based anomaly detection algorithm determines whether the corresponding data is abnormal through the magnitude of the prediction error.

In terms of training efficiency, the training time of each epoch for the GRU model is shorter than that of the LSTM model, with a maximum reduction of 12.2% in ECG dataset. From the perspective of processing the large-scale test data of a launch vehicle, the shorter the training times are, the faster the data will be processed.

Combined with the above analysis, the GRU-based prediction model has a faster running rate. Although the prediction accuracy is slightly higher than that of LSTM, it does not impact the results of anomaly detection. Therefore, we choose GRU to construct the prediction network model.

TABLE 3. Comparison between GRU and LSTM.

No.	dataset	GRU		LSTM	
		time	Accuracy	time	Accuracy
1	Numenta	90s	2.82	99s	2.38
2	Power Demand	70s	0.09	75s	0.08
3	ECG	130s	349.01	148s	346.11

3) EFFECT OF EVT

In addition to prediction model, another aspect that significantly influences the final performance of anomaly detection is the detection rule. Motivated by the fact that the tails distributions of the prediction errors follow a Generalized Pareto Distribution (GPD), we propose a detection rule based on Extreme Value Theory (EVT) which is able to automatically set thresholds for prediction errors.

In order to assess the performance of our EVT-based detection rule, we also compare the EVT-based methodology to a complementary unsupervised and nonparametric anomaly thresholding approach [16], a method leveraging Chebyshev's inequality [43] and a Gaussian-based thresholding rule [44]. The former three detection rules are not bound to the underlying distribution of the prediction errors; while the Gaussian-based rule assume that the prediction errors should follow a Gaussian distribution.

The experiments are conducted on a real dataset containing voltage values sampled from a power system of a launch vehicle. The dataset contains a total of 11 abnormal points, and values below 34V are defined as abnormal points and emphasized as red dots in Fig. 8. The abnormal points are caused by detonating related initiating explosive devices or controlling the swing of the nozzles during flight, which

TABLE 4. Comparisons of different detection rules.

No.	Method	Precision	recall	F1
1	EVT-based method	**100%**	**90.9%**	**95.3%**
2	LSTM-NAT	88.9%	72.3%	79.7%
3	Chebyshev's inequality	72.7%	66.7%	69.6%
4	Gaussian-based rule	70.0%	63.6%	66.6%

leads the voltage values to drop at the corresponding time. This dataset is fed into the GRU-based stacked predictor to get prediction errors on which these four detection rules are implemented, respectively. The results are shown in Table 4.

It can be seen from Table 3 that our EVT-based detection method could identify 10 anomalies and none false positive, which outperforms three other baseline detection rules. Although the Chebyshev's inequality method could detect most of anomalies, it also results in many false positives, which lead to a low precision. The Gaussian-base method also results in many false positives indicating that presuming a Gaussian distribution on the prediction errors is a very strong assumption and might not always hold for many application scenarios. The precision of LSTM-NAT method is 88.9%, but the recall is 72.3% showing that it would miss many real anomalies.

V. CONCLUSION

Spacecraft anomaly detection can help to discover and identify abnormal behaviors in advance and avoid potential cascading downtime. In this paper, to address the problems of large amount of telemetry data and imbalanced samples in spacecraft, we propose an unsupervised deep learning-based anomaly detection methodology using stacked Gated Recurrent unit (GRU) based recurrent neural networks and Extreme Value Theory (EVT). Inspired by ensemble learning, we proposed a stacked GRU-based network which could not only promote the prediction performance, but also be robust to multivariate telemetry data. Moreover, we devise an EVT-based detection rule to set dynamic thresholds for prediction errors without making any critic assumptions about the distribution. Through extensive experiments, our proposed approach exceeds state-of-the-art methods on many large datasets.

To the best of our knowledge, this is the initial exploration of the GRU based neural networks and EVT in the field of spacecraft telemetry data anomaly detection. In the future, we will focus on how to improve the efficiency and robustness across multiply related spacecraft and machines by using transfer learning.

REFERENCES

[1] P. Zhao, X. Chang, and M. Wang, "A novel multivariate time-series anomaly detection approach using an unsupervised deep neural network," *IEEE Access*, vol. 9, pp. 109025–109041, 2021.

[2] M. Munir, S. A. Siddiqui, A. Dengel, and S. Ahmed, "DeepAnT: A deep learning approach for unsupervised anomaly detection in time series," *IEEE Access*, vol. 7, pp. 1991–2005, 2019.

[3] T. Li, M. Comer, E. Delp, S. R. Desai, J. L. Mathieson, R. H. Foster, and M. W. Chan, "A stacked predictor and dynamic thresholding algorithm for anomaly detection in spacecraft," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, Norfolk, VA, USA, Nov. 2019, pp. 165–170.

[4] J. Chen, D. Pi, Z. Wu, X. Zhao, Y. Pan, and Q. Zhang, "Imbalanced satellite telemetry data anomaly detection model based on Bayesian LSTM," *Acta Astronautica*, vol. 180, pp. 232–242, Mar. 2021.

[5] H. M. Hashemian and W. C. Bean, "State-of-the-art predictive maintenance techniques," *IEEE Trans. Instrum. Meas.*, vol. 60, no. 10, pp. 3480–3492, Oct. 2011.

[6] J.-A. Martínez-Heras and A. Donati, "Enhanced telemetry monitoring with novelty detection," *AI Mag.*, vol. 35, no. 4, pp. 37–46, Dec. 2014.

[7] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, p. 15, Sep. 2009.

[8] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 93–104, 2000.

[9] L. Zhang, J. Lin, and R. Karim, "Adaptive kernel density-based anomaly detection for nonlinear systems," *Knowl.-Based Syst.*, vol. 139, pp. 50–63, Jan. 2018.

[10] J. Lei, T. Jiang, K. Wu, H. Du, and L. Zhu, "Robust local outlier detection with statistical parameters for big data," *Comput. Syst. Sci. Eng.*, vol. 30, no. 5, pp. 411–419, 2015.

[11] A. Goldman and I. Cohen, "Anomaly detection based on an iterative local statistics approach," *Signal Process.*, vol. 84, no. 7, pp. 1225–1229, Jul. 2004.

[12] Y. Chen, Q. Zhao, and L. Lu, "Combining the outputs of various K-nearest neighbor anomaly detectors to form a robust ensemble model for high-dimensional geochemical anomaly detection," *J. Geochem. Explor.*, vol. 231, Dec. 2021, Art. no. 106875.

[13] M. Shen, X. Jiang, and T. Sun, "Anomaly detection based on nearest neighbor search with locality-sensitive B-tree," *Neurocomputing*, vol. 289, pp. 55–67, May 2018.

[14] S. M. Erfani, S. Rajasegarar, S. Karunasekera, and C. Leckie, "High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning," *Pattern Recognit.*, vol. 58, pp. 121–134, Oct. 2016.

[15] T. Yairi, N. Takeishi, T. Oda, Y. Nakajima, N. Nishimura, and N. Takata, "A data-driven health monitoring method for satellite housekeeping data based on probabilistic clustering and dimensionality reduction," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 53, no. 3, pp. 1384–1401, Jun. 2017.

[16] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Soderstrom, "Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding," 2018, *arXiv:1802.04431*.

[17] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] J. L. Elman, "Finding structure in time," *Cognit. Sci.*, vol. 14, no. 2, pp. 179–211, Mar. 1990.

[20] A. Siffer, P.-A. Fouque, A. Termier, and C. Largouet, "Anomaly detection in streams with extreme value theory," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 1067–1075.

[21] D. Toshkova, M. Asher, P. Hutchinson, and N. Lieven, "Automatic alarm setup using extreme value theory," *Mech. Syst. Signal Process.*, vol. 139, May 2020, Art. no. 106417.

[22] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," 2019, *arXiv:1901.03407*.

[23] P. M. Hughes and E. C. Luczak, "The generic spacecraft analyst assistant (GenSAA): A tool for automating spacecraft monitoring with expert systems," *Telematics Informat.*, vol. 8, no. 4, pp. 339–351, Jan. 1991.
[24] N. Nishigori, "Fully automatic and operator-less anomaly detecting ground support system for mars probe'nozomi," in *Proc. 6th Int. Symp. Artif. Intell. Robot. Autom. Space*, 2001, pp. 1–8.
[25] D. Bernard, R. Doyle, and E. Riedel, "Autonomy and software technology on NASA's deep space one," *IEEE Intell. Syst. Appl.*, vol. 14, no. 3, pp. 10–15, May 1999, doi: 10.1109/5254.769876.
[26] J. Li, H. Izakian, W. Pedrycz, and I. Jamal, "Clustering-based anomaly detection in multivariate time series data," *Appl. Soft Comput.*, vol. 100, Mar. 2021, Art. no. 106919.
[27] H. B. Gunay and Z. Shi, "Cluster analysis-based anomaly detection in building automation systems," *Energy Buildings*, vol. 228, Dec. 2020, Art. no. 110445.
[28] W. Khreich, B. Khosravifar, A. Hamou-Lhadj, and C. Talhi, "An anomaly detection system based on variable N-gram features and one-class SVM," *Inf. Softw. Technol.*, vol. 91, pp. 186–197, Nov. 2017.
[29] S. Kanarachos, S.-R. G. Christopoulos, A. Chroneos, and M. E. Fitzpatrick, "Detecting anomalies in time series data via a deep learning algorithm combining wavelets, neural networks and Hilbert transform," *Expert Syst. Appl.*, vol. 85, pp. 292–304, Nov. 2017.
[30] Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, and D. Pei, "Robust anomaly detection for multivariate time series through stochastic recurrent neural network," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Anchorage, AK, USA, Jul. 2019, pp. 1–15.
[31] D. Iverson, "Data mining applications for space mission operations system health monitoring," in *Proc. SpaceOps Conf.*, May 2008, pp. 1–6, doi: 10.2514/6.2008-3212.
[32] S. Fuertes, G. Picart, J.-Y. Tourneret, L. Chaari, A. Ferrari, and C. Richard, "Improving spacecraft health monitoring with automatic anomaly detection techniques," in *Proc. SpaceOps Conf.*, May 2016, p. 2430.
[33] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, "Long short term memory networks for anomaly detection in time series," in *Proc. Eur. Symp. Artif. Neural Netw. (ESANN)*, 2015, pp. 89–94.
[34] S. Tariq, S. Lee, Y. Shin, M. S. Lee, O. Jung, D. Chung, and S. S. Woo, "Detecting anomalies in space using multivariate convolutional LSTM with mixtures of probabilistic PCA," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Anchorage, AK, USA, Jul. 2019, pp. 2123–2133.
[35] J. Wu, W. Zeng, and F. Yan, "Hierarchical Temporal Memory method for time-series-based anomaly detection," *Neurocomputing*, vol. 273, pp. 535–546, Mar. 2018.
[36] J. Chung, C. Gulcehre, and K. Cho, "Gated feedback recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2067–2075.
[37] S. V. Crowder, "Design of exponentially weighted moving average schemes," *J. Quality Technol.*, vol. 21, no. 3, pp. 155–162, Jul. 1989.
[38] J. Bergstra, R. Bardenet, and Y. Bengio, "Algorithms for hyper-parameter optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 2546–2554.
[39] D. Kingma and J. Ba Adam, "A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.
[40] S. Ahmad, A. Lavin, S. Purdy, and Z. Agha, "Unsupervised real-time anomaly detection for streaming data," *Neurocomputing*, vol. 262, pp. 134–147, Nov. 2017.
[41] M. Jones, D. Nikovski, M. Imamura, and T. Hirata, "Anomaly detection in real-valued multidimensional time series," in *Proc. Int. Conf. Big-data/Socialcom/Cybersecur.*, 2014, pp. 1–11.
[42] M. C. Chuah, *ECG Anomaly Detection via Time Series Analysis*. Berlin, Germany: Springer, 2007, pp. 123–135.
[43] G. Shevlyakov and M. Kan, "Stream data preprocessing: Outlier detection based on the Chebyshev inequality with applications," in *Proc. 26th Conf. Open Innov. Assoc. (FRUCT)*, Apr. 2020, pp. 402–407.
[44] A. Singh, "Anomaly detection for temporal data using long short-term memory," M.S. thesis, School Inf. Commun. Technol., KTH Roy. Inst. Technol., Stockholm, Sweden, 2017.

GANG XIANG received the B.Eng. and M.S. degrees in automatic test and control from the Harbin Institute of Technology, Harbin, China, in 2009 and 2011, respectively. He is currently pursuing the Ph.D. degree with Beihang University. He has been a Senior Engineer with the Beijing Aerospace Automatic Control Institute, since 2011. His current research interests include deep learning, intelligent control, fault diagnosis, and PHM.

RUISHI LIN received the B.Eng. and M.S. degrees in automation from Harbin Engineering University, Harbin, China, in 2005 and 2008, respectively. He has been a Senior Engineer with the Beijing Aerospace Automatic Control Institute, since 2008. His current research interests include machine learning and its applications in complex machines health management.