Robust Sequential Variational Autoencoder for Multivariate Time Series Anomaly Detection

Yao Yu[1], **Junhyeok Kang**[1], **Jae-Gil Lee**[1*], **Jonghwa Kim**[2] and **Kyungdeok Seo**[2]
[1]Graduate School of Knowledge Service Engineering, KAIST
[2]Hanwha Systems
{yuyao, junhyeok.kang, jaegil}@kaist.ac.kr, {jonghwa3.kim, kyungdeok.seo}@hanwha.com

Abstract

Various practical applications call for an effective anomaly detection method owing to its significance and high demand. However, most time series anomaly detection models lack robustness, suffering from noise and outliers in the training sets and thus always achieve sub-optimal anomaly detection performance in the real world. To tackle this problem, in this paper, we propose *Gated Recurrent Unit–Robust Variational AutoEncoder* (**GRU-RVAE**), an unsupervised anomaly detection model for multivariate time series data. GRU-RVAE applies the bidirectional gated recurrent units to model informative dependencies among time series and the variational autoencoder with a modified loss function to explicitly process noise and outliers in the training stage. We conduct experiments on two representative multivariate time series datasets, and the experimental results show that GRU-RVAE outperforms four state-of-art baselines in both anomaly detection performance and robustness, achieving the improvements of 2.7–6.6% in the best F1-scores for different levels of noise and outliers in the training sets.

1 Introduction

An *anomaly* or *outlier* [Hawkins, 1980] is defined as an observation that deviates so much from the other observations that it can be considered to be generated by a different mechanism. It always contains useful information about the abnormal characteristics of the systems and entities that impact the data generation process [Aggarwal, 2016]. Anomaly detection is of great importance and has been of interest to various research and application domains for ages. With regard to different data types, anomaly detection typically requires dedicated techniques. In this paper, we will focus on anomaly detection for *multivariate time series* data, which can be frequently collected from the Internet-of-Things.

A significant amount of multivariate time series data is generated, collected, and analyzed in diverse practical applications such as health care, system diagnosis, and geoscience. For example, in the health care field, electroencephalograms (EEGs) are recorded to monitor the electrical activities of the brains. In the industrial applications, sensory records are always collected from multiple sensors to support system diagnosis. Anomalous patterns in those records can reflect abnormal activities in human bodies and systems, and by applying anomaly detection, corresponding issues can be timely identified and resolved.

We decide to work on *unsupervised* methods because of the lack of fully-labeled datasets in the real world. Unsupervised anomaly detection is still challenging due to the existence of noise and outliers in the training sets. Most unsupervised models assume "noise-free" datasets to learn the normal patterns, and the deviation between data instances and the normal patterns is used to detect anomalies. However, this assumption on being noise-free does not typically hold in real-world applications, since noise is out of human control. Therefore, in support of robustness, unsupervised anomaly detection methods need to cope with noise and outliers in the training sets. Unfortunately, existing methods do not fulfill the above-mentioned requirement. For example, Donut [Xu *et al.*, 2018] can deal with only missing data explicitly in the training phase, but is not applicable to noise and outliers; the Robust Variational AutoEncoder (RVAE) [Eduardo *et al.*, 2019] does not take advantage of the temporal correlation of time series and thus is not appropriate for time series data.

In this paper, to achieve better robustness and anomaly detection accuracy simultaneously, we propose *Gated Recurrent Unit–Robust Variational AutoEncoder* (**GRU-RVAE**), an unsupervised anomaly detection model for multivariate time series data. GRU-RVAE leverages (i) the bidirectional Gated Recurrent Unit (GRU) [Vukotic *et al.*, 2016] to model informative dependencies among time series and (ii) the Variational AutoEncoder (VAE) [An and Cho, 2015] with a modified loss function to automatically identify, explicitly process the potential anomalies in the training sets, and learn the data distribution of time series in an unsupervised manner. Combining the two components, GRU-RVAE shows great effectiveness in multivariate time series anomaly detection. The experiments on two representative datasets confirm that GRU-RVAE outperforms four state-of-art baselines in both anomaly detection accuracy and robustness, achieving the improvements of 2.7–6.6% in the best F1-scores for various levels of noise and outliers in the training sets.

*Contact Author

2 Related Work

In this section, we review four relevant methods for anomaly detection in multivariate time series.

EncDec-AD [Malhotra *et al.*, 2016] is a reconstruction-based model with an autoencoder. The encoder learns a compressed representation of time series, and the decoder reconstructs the time series from the latent representation. Since the model requires to be trained on a clean dataset, it can learn the patterns of normal time series, but cannot reconstruct the anomalous time series very well. The reconstruction error at each time step is used as an anomaly score, where a larger reconstruction error denotes a higher likelihood to be anomalous. Since EncDec-AD conducts dimensionality reduction in the encoding phase, to some extent, it can resist the noise and outliers in the datasets.

LSTM-VAE [Park *et al.*, 2017] is a stochastic model that combines a VAE and LSTMs to catch the temporal dependencies among time series and learn the data distribution of each time step. In the detection phase, it generates a negative reconstruction probability as the anomaly score for each time step, where a higher anomaly score indicates that the current time step cannot be well reconstructed by LSTM-VAE and is more likely to be an anomaly. Since the objective of LSTM-VAE is to learn the data distribution of normal time series, it also requires to be trained on clean training sets. Therefore, noise and outliers in the training sets should have a negative impact on its performance.

Donut [Xu *et al.*, 2018] is a state-of-art unsupervised VAE-based model for detecting anomalies in univariate time series. Donut recognizes the anomalies depending on their indicators and is able to process missing data, noise, and outliers explicitly in the training stage. By applying a modified Evidence Lower BOund (ELBO), Donut excludes anomalies from loss calculation and computes the loss only for normal data to learn the normal patterns. It detects anomalies in an unsupervised manner as well as in a supervised manner. Regarding missing data, it automatically generates the indicators of missing data without any labels, being fully unsupervised. However, since Donut does not recognize the noise and outliers automatically, it requires the labels that mark noise and outliers to generate the corresponding indicators.

OmniAnomaly [Su *et al.*, 2019] is an unsupervised VAE-based anomaly detection model for multivariate time series. Instead of processing the noise and outliers in the training sets explicitly, OmniAnomaly achieves robustness by dimensionality reduction and improves the model performance by generating more informative representation for time series and modeling the temporal dependencies among stochastic variables to include more valid information among time series. Nevertheless, since it does not process noise and outliers explicitly, the model performance can still be affected by the quality of training sets.

3 Proposed Model: GRU-RVAE

3.1 Problem Definition

A *multivariate time series* is defined as $X = \{x_1, x_2, \ldots, x_N\}$, where N is the length of the time series, each time step $x_t = \{x_t^1, x_t^2, \ldots, x_t^m\}$ is an m-dimensional vector, and x_t^m is the value of the m-th dimension of time series X at time t. The goal of multivariate time series anomaly detection is to determine whether each time step x_t is anomalous or not. An anomaly score is provided for x_t, which will be compared with a predefined threshold to denote whether x_t is an anomaly. *Robust* detection is to be trained in an unsupervised manner resisting the anomalous data in the training sets while maintaining good anomaly detection performance.

3.2 Overview

The main goals of our model are set to be as follows:

1. To model the temporal dependencies among multivariate time series and generate informative representation for time series;
2. To automatically identify and explicitly process the anomalous data in the training sets, where the model can be trained in an unsupervised manner;
3. To learn the data distribution of normal multivariate time series in the training stage and provide an anomaly score for each time step in the anomaly detection phase.

To accomplish these goals, we design GRU-RVAE as shown in Figure 1, which consists of a *bidirectional GRU* component and an *RVAE* component:

- The *bidirectional GRU* models the temporal dependencies among time series to generate informative representation for each time step in the encoding stage and the decoding stage (**Goal 1**);
- The *RVAE* identifies anomalous data in the training stage, alleviates its impacts on parameter optimization, and learns the distribution for each time step. In the anomaly detection stage, it provides an anomaly score to denote whether the current time step is anomalous or not (**Goals 2 & 3**). The RVAE shares the same architecture as a vanilla VAE [An and Cho, 2015] but with an adjusted loss function.

3.3 Model Architecture

The workflow of GRU-RVAE is described with Figure 1. First, a multivariate time series $X = \{x_1, x_2, \ldots, x_N\}$ is fed into the bidirectional GRU generating the corresponding output $h = \{h_1, h_2, \ldots, h_N\}$, where h_t contains contextual information of the input time series. Then, h_t is processed by hidden layers and projected into a mean vector μ_t and a standard deviation vector σ_t forming the normal distribution $N(\mu_t, \sigma_t)$ in the latent space. After that, a latent variable z_t is sampled from this distribution and fed into the bidirectional GRU cells to generate a hidden state h'_t, which is afterward mapped into a reconstructed mean vector μ'_t and a reconstructed standard deviation vector σ'_t that parameterize the reconstructed normal distribution $N(\mu'_t, \sigma'_t)$. By feeding x_t into the reconstructed distribution, a reconstruction probability is calculated, showing how well x_t is reconstructed by each RVAE component: a high reconstruction probability indicates that x_t can be well reconstructed and is more likely to be normal, while x_t with a low reconstruction probability is declared as an anomaly.

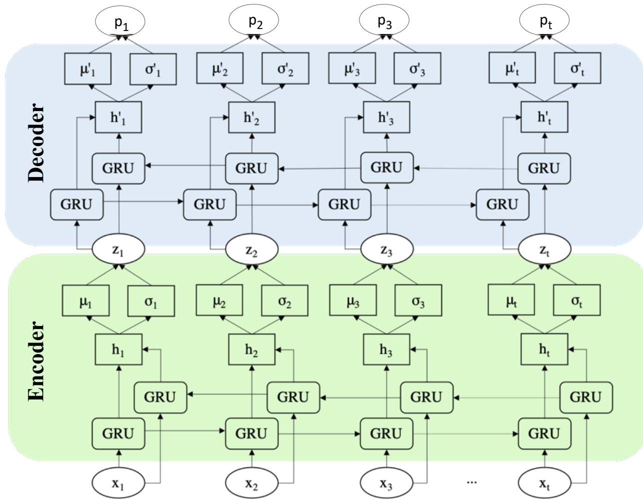

Figure 1: Architecture of GRU-RVAE.

Bidirectional GRU Component

Instead of a basic GRU, we adopt a bidirectional GRU to learn the temporal dependencies since the bidirectional GRU cells allow us to make use of both past and future information to model the temporal correlation for each time step and thus create more informative representation of time series. The internal process in the basic GRU cells is formulated by Equation (1), where W_u, U_u, W_r, U_r, W_h, and U_h are the weights in the GRU cells, u_t and r_t are the update gate and the reset gate, and h_{t-1} and h_t represent the previous hidden state and the current hidden state (output vector), respectively.

$$\begin{aligned} u_t =& \text{sigmoid}(W_u x_t + U_u h_{t-1} + b_u) \\ r_t =& \text{sigmoid}(W_r x_t + U_r h_{t-1} + b_r) \\ h_t =& u_t \odot h_{t-1} + \\ & (1-u_t) \odot \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h) \end{aligned} \tag{1}$$

The bidirectional GRU cells concatenate two hidden states generated from the forward process and the backward process to a single output as in Equation (2), enabling the current time step to be modeled by both past and future states.

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \tag{2}$$

RVAE Component

The RVAE is trained based on the adjusted objective function in Equation (3), where $p_\theta(x_t|z_t)$ is the clean component and p_0 is the outlier component. The *clean* component is a vanilla VAE to learn the patterns of normal time series, while the *outlier* component is to describe the outliers generating high probabilities for outliers and low probabilities for normal inputs. θ and ϕ denote the network parameters of the encoder and the decoder, respectively, of the vanilla VAE. The detailed definitions of q_ϕ and q_π can be found in [Eduardo *et al.*, 2019]. $\pi_t(x_t)$ is treated as an independent parameter of the optimization problem, which should always reach its optimal value while calculating the loss. By taking the derivative of the objective function L with respect to $\pi_t(x_t)$, the optimal $\hat{\pi}_t(x_t)$ is derived as Equation (4), where $\alpha \in [0, 1]$.

$$\begin{aligned} L =& E_{q_\phi(z_t|x_t)}[\pi_t(x_t)\log p_\theta(x_t|z_t) + (1-\pi_t(x_t))\log p_0(x_t)] \\ & - D_{KL}(q_\phi(z_t|x_t)||p(z_t)) - D_{KL}(q_\pi(w_t|x_t)||p(w_t)) \end{aligned} \tag{3}$$

$$\hat{\pi}_t(x_t) = \text{sigmoid}(E_{q_\phi(z_t|x_t)}[\log \frac{p_\theta(x_t|z_t)}{p_0(x_t)} + \log \frac{\alpha}{1-\alpha}]) \tag{4}$$

In Equation (3), $\pi_t(x_t)$ acts as a weight for adjusting the contribution of the clean component and the outlier component to the total loss. When the input tends to be an anomaly, because a vanilla VAE cannot reconstruct it very well, $p_\theta(x_t|z_t)$ is close to 0, but $p_0(x_t)$ is high; thus, $\hat{\pi}_t(x_t)$ will be close to 0. In this case, the outlier component mainly contributes to the total loss. On the other hand, when the input is likely to be normal, $p_\theta(x_t|z_t)$ becomes high, but $p_0(x_t)$ becomes very low; thus, $\hat{\pi}_t(x_t)$ becomes higher, which indicates that the clean component mainly contributes to the total loss. Therefore, the parameters in the clean component (i.e., the vanilla VAE) are optimized based on the normal data, and the RVAE is able to learn the data distribution of normal time series in the existence of noise and outliers.

In order to describe the outliers in the outlier component, we use a multivariate broad Gaussian distribution with mean zero and standard deviation $S\,(> 1)$, because, according to our experience, this distribution outperforms a t-distribution, a gamma distribution, and a Gaussian distribution with other mean and standard variance values.

3.4 Model Training

Algorithm 1 outlines the training process of GRU-RVAE. For a time series $x^{(i)}$ in the training set, each time step $x_t^{(i)}$ is fed into the vanilla VAE as well as the multivariate broad Gaussian distribution, and $p_\theta(x_t^{(i)}|z_t^{(i)})$ as well as $\log p_0(x_t^{(i)})$ are evaluated. Next, the algorithm infers $\hat{\pi}_t(x_t^{(i)})$ by Equation 4, which reflects the predicted probability of $x_t^{(i)}$ being normal with the encoder and decoder networks trained by far. Depending on the value of $x_t^{(i)}$, the algorithm restricts the flow of gradients through the networks to ignore $x_t^{(i)}$ in the optimization the parameters θ and ϕ. Then, the algorithm optimizes these parameters by backpropogation.

Algorithm 1 GRU-RVAE Training

Input: Training dataset:
$X = \{x_1^{(1)}, \ldots, x_N^{(1)}, \ldots, x_1^{(N_{train})}, \ldots, x_N^{(N_{train})}\}$
Output: Optimized parameters: θ and ϕ of the encoder and decoder networks

```
for 1 ≤ i ≤ N_train do
    for 1 ≤ t ≤ N do
        Evaluate p_θ(x_t^(i) | z_t^(i)) and log p_0(x_t^(i)).
        Infer π̂_t(x_t^(i)) by Equation (4).
        Stop gradients depending on π̂_t(x_t^(i)).
        Optimize the network parameters.
    end for
end for
return optimized parameters
```

3.5 Anomaly Detection

Algorithm 2 outlies the anomaly detection process of GRU-RVAE. A unknown time series $x^{(i)} = \{x_1^{(i)}, x_2^{(i)}, \ldots, x_N^{(i)}\}$ is fed into GRU-RVAE, and a *reconstruction probability* $p_t^{(i)}$ is produced for each time step $x_t^{(i)}$. This reconstruction probability, which is regarded as an anomaly score, is compared against a predefined threshold τ. Accordingly, if $p_t^{(i)}$ is lower than τ, $x_t^{(i)}$ is determined to be anomalous; otherwise, it is determined to be normal.

Algorithm 2 GRU-RVAE Anomaly Detection

Input: Trained network parameters: θ and ϕ;
Test dataset: $X = \{\ldots, x_1^{(N_{test})}, \ldots, x_N^{(N_{test})}\}$;
Anomaly threshold: τ
Output: Anomaly label

for $1 \leq i \leq N_{test}$ **do**
  for $1 \leq t \leq N$ **do**
    $\mu_{z_t^{(i)}}, \sigma_{z_t^{(i)}} = f_\theta(z|x_t^{(i)})$.
    Draw L samples from $z \sim N(\mu_{z_t^{(i)}}, \sigma_{z_t^{(i)}})$.
    for $1 \leq l \leq L$ **do**
      $\mu_{\hat{x}_t^{(i,l)}}, \sigma_{\hat{x}_t^{(i,l)}} = g_\phi(x|z_t^{(i,l)})$.
    end for
    Calculate $p_t^{(i)} = \frac{1}{L}\sum_{l=1}^{L} N(x_t^{(i)}|\mu_{\hat{x}_t^{(i)}}, \sigma_{\hat{x}_t^{(i)}})$.
    if $p_t^{(i)} < \tau$ **then**
      $x_t^{(i)}$ is *anomalous*.
    else
      $x_t^{(i)}$ is *normal*.
    end if
  end for
end for
return *normal* or *anomalous*

4 Evaluation

4.1 Setting

Datasets

To verify the effectiveness of GRU-RVAE, we conducted experiments on two representative multivariate time series datasets: Server Machine Dataset (SMD) and Mars Science Laboratory (MSL) Curiosity Rover. SMD [Su *et al.*, 2019] is a five-week dataset collected from an Internet company, which consists server records from different machines; we selected two machine records: SMD-1-1 and SMD-1-4. The labels of point anomalies are provided by domain experts, and each machine record is already divided into the training set and the test set. MSL [Hundman *et al.*, 2018] is an open spacecraft telemetry dataset from NASA. It is also already divided into the training and the test set, and point anomaly labels are provided as well. The details of each dataset are shown in Table 1.

	# of dim.	Training set size	Test set size	#, % of point anomalies
SMD-1-1	38	28,479	28,479	2,694 (9.46%)
SMD-1-4	38	23,706	23,707	720 (3.04%)
MSL	55	58,317	73,729	7,905 (10.72%)

Table 1: Profile of the experimental data.

Data Corruption

To evaluate the robustness of each algorithm, we artificially added different levels of noise and outliers into the training sets by (i) randomly selecting a time series segment and a set of dimensions and (ii) injecting random noise and outliers into the selected dimensions of the time series. For each dataset, we injected 5%, 10%, and 20% *additional* noise and outliers into the training set.

Compared Models

We compared GRU-RVAE with the four models introduced in Section 2: EncDec-AD [Malhotra *et al.*, 2016], LSTM-VAE [Park *et al.*, 2017], Donut [Xu *et al.*, 2018], and OmniAnomaly [Su *et al.*, 2019]. For ablation studies, we also included *GRU-RVAE-b* in which the bidirectional GRU cells were replaced with the basic GRU cells in GRU-RVAE. As for Donut, since it is designed for univariate time series, we applied it on each dimension of multivariate time series separately and summed up the reconstruction probabilities from the multiple dimensions.

Model Configuration

To make comparison fair and intuitive, we replaced LSTM cells in EncDec-AD and LSTM-VAE with GRU cells. After the modification, LSTM-VAE was renamed as GRU-VAE for clarification. Since the model performance is highly sensitive to hyperparameters, for each dataset, we conducted grid search within the range in Table 2 to achieve the best results. The hyperparameters of the existing models were also favorably tuned to achieve the best results. Besides, L2 regularization and early stopping were applied to prevent overfitting.

Parameters	Values
Hidden layers	2
GRU units	10, 50, 100, 500
Sliding window length	50, 100
z dimensions	3
Batch size	50, 100
Optimizer	Adam
Initial learning rate	0.01, 0.001, 0.0005, 0.0001
Input dimensions	38 (SMD), 55 (MSL)

Table 2: Model parameters.

Evaluation Metric

By executing Algorithm 2 on the test sets, we measured the F1-scores for anomaly detection. Since Donut and EncDec-AD do not provide thresholding methods, following the methodologies by [Xu *et al.*, 2018; Su *et al.*, 2019], we chose the best F1-score as the metric to evaluate the performance of each anomaly detection model. The *best F1-score* is the highest score among a set of F1-scores which are generated by enumerating all valid thresholds.

	Donut	EncDec-AD	GRU-VAE	OmniAnomaly	GRU-RVAE-b	**GRU-RVAE**
No corruption	0.9059	0.9394	0.9205	0.9386	0.9569	**0.9589**
5% corruption	0.5847	0.9153	0.8679	0.9272	**0.9685**	0.9669
10% corruption	0.5925	0.9483	0.8970	0.9557	0.9679	**0.9691**
20% corruption	0.7322	0.8760	0.9325	0.9421	0.9684	**0.9686**

Table 3: Experiment results on the SMD-1-1 dataset.

	Donut	EncDec-AD	GRU-VAE	OmniAnomaly	GRU-RVAE-b	**GRU-RVAE**
No corruption	0.5125	0.6004	**0.8473**	0.8377	0.8355	0.8454
5% corruption	0.4352	0.5619	0.5501	0.8417	0.8359	**0.8459**
10% corruption	0.4187	0.5607	0.5341	0.7661	0.7117	**0.8441**
20% corruption	0.2514	0.5502	0.4259	0.7292	0.6237	**0.8344**

Table 4: Experiment results on the SMD-1-4 dataset.

	Donut	EncDec-AD	GRU-VAE	OmniAnomaly	GRU-RVAE-b	**GRU-RVAE**
No corruption	0.5417	0.8183	0.7819	0.7961	0.8207	**0.8223**
5% corruption	**0.8336**	0.7996	0.7667	0.7202	0.8155	0.8254
10% corruption	0.8101	0.7988	0.7844	0.7650	0.8108	**0.8278**
20% corruption	0.8067	0.7957	0.7559	0.7755	**0.8324**	0.8231

Table 5: Experiment results on the MSL dataset.

4.2 Results

SMD-1-1 Dataset (Table 3)

GRU-RVAE-b and *GRU-RVAE* outperformed the other compared methods not only in anomaly detection performance but also in robustness. The performance of Donut was easily affected by the quality of the training set since there was a significant variance in the best F1-scores when the quality of the training set varied. The performances of EncDec-AD and GRU-VAE were relatively stable, but the best F1-scores still fluctuated when the data quality changed. For different levels of noise and outliers in the dataset, the performances of OmniAnomaly, *GRU-RVAE-basic*, and *GRU-RVAE* stayed stable; however, the best F1-scores of *GRU-RVAE-b* and *GRU-RVAE* were higher than that of OmniAnomaly, showing higher anomaly detection capability of our models.

SMD-1-4 Dataset (Table 4)

Donut and EncDec-AD showed relatively poor performance. The performance of GRU-VAE was comparable with those of OmniAnomaly, *GRU-RVAE-b*, and *GRU-RVAE*. However, when there were more noise and outliers in the training set, the best F1-score of GRU-VAE decreased rapidly, indicating its low robustness. There was also a decreasing trend in the best F1-scores of OmniAnomaly and *GRU-RVAE-b* as the quality of the training set degraded. Possibly because SMD-1-4 had fewer point anomalies than SMD-1-1 (see Table 1), the performances of the other models fluctuated more severely in SMD-1-4 than in SMD-1-1, whereas *GRU-RVAE* still maintained good performance even in SMD-1-4, indicating is remarkable robustness.

MSL Dataset (Table 5)

The performance of Donut showed a suspicious trend, where the best F1-score improved as the quality of the training set degraded. We conjecture that this phenomenon was caused by the characteristics of the dataset whose data distribution is severely skewed and mostly distributed around zero. By injecting noise and outliers, the data distribution could become more balanced, and thus noise and outliers started to be differentiated and recognized. The best F1-scores of EncDec-AD, GRU-VAE, and OmniAnomaly showed a similar trend in both SMD and MSL datasets: they fluctuated and decreased when more noise and outliers were injected into the training set. In contrast, the performances of *GRU-RVAE-b* and *GRU-RVAE* stayed relatively stable.

Average of Three Datasets (Table 6) and Summary

The best F1-scores of EncDec-AD, GRU-VAE, and OmniAnomaly showed a decreasing trend when the quality of training sets degraded, showing that these models are sensitive to noise and outliers. On the other hand, the best F1-scores of *GRU-RVAE-b* and *GRU-RVAE* stayed stable and were always higher than those of the other models. For different levels (0%, 5%, 10%, and 20%) of noise and outliers in the training sets, *GRU-RVAE* improved the other models *at least* by 2.7%, 6.6%, 6.1%, and 5.4%, respectively. The higher performance of *GRU-RVAE* over *GRU-RVAE-b* verifies the necessity of using bidirectional GRU. Overall, these results clearly demonstrate the superiority of *GRU-RVAE*, which outperforms the baselines in both anomaly detection performance and robustness.

	Donut	EncDec-AD	GRU-VAE	OmniAnomaly	GRU-RVAE-b	**GRU-RVAE**
No corruption	0.6397	0.8289	0.8201	0.8326	0.8534	**0.8552**
5% corruption	0.7394	0.8062	0.7791	0.7782	0.8524	**0.8593**
10% corruption	0.7208	0.8122	0.7981	0.8089	0.8409	**0.8616**
20% corruption	0.7305	0.7925	0.7780	0.8127	0.8559	**0.8565**

Table 6: Averages of the three datasets (Tables 3, 4, and 5).

5 Conclusion

In this paper, we proposed GRU-RVAE, a novel unsupervised anomaly detection model for multivariate time series. By applying the bidirectional GRU and VAE with an adjusted loss function, GRU-RVAE is able to model the temporal dependencies among time series, automatically identify and explicitly process noise and outliers in the training stage, and learn the data distribution of each time step even for noisy training sets. To demonstrate the performance of GRU-RVAE, we experimented on two representative multivariate time series datasets. Moreover, we injected different levels of noise and outliers into the datasets to verify its robustness. The experimental results showed that GRU-RVAE outperforms four state-of-art baselines in both anomaly detection performance and robustness, achieving the improvements of 2.7–6.6% in the best F1-scores for different levels of noise and outliers in the training sets.

Acknowledgments

This work was supported by a grant-in-aid of HANWHA SYSTEMS.

References

[Aggarwal, 2016] Charu C. Aggarwal. *Outlier Analysis*. Springer, 2nd edition, 2016.

[An and Cho, 2015] Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. Technical Report SNUDM-TR-2015-03, Seoul National University, December 2015.

[Eduardo *et al.*, 2019] Simão Eduardo, Alfredo Nazábal, Christopher K. I. Williams, and Charles A. Sutton. Robust variational autoencoders for outlier detection in mixed-type data. *CoRR*, abs/1907.06671, 2019.

[Hawkins, 1980] D. M. Hawkins. *Identification of Outliers*. Monographs on applied probability and statistics. Chapman and Hall, 1980.

[Hundman *et al.*, 2018] Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Söderström. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 387–395. ACM, 2018.

[Malhotra *et al.*, 2016] Pankaj Malhotra, Anusha Ramakrishnan, Gaurangi Anand, Lovekesh Vig, Puneet Agarwal, and Gautam M. Shroff. LSTM-based encoder-decoder for multi-sensor anomaly detection. *CoRR*, abs/1607.00148, 2016.

[Park *et al.*, 2017] Daehyung Park, Yuuna Hoshi, and Charles C. Kemp. A multimodal anomaly detector for robot-assisted feeding using an LSTM-based variational autoencoder. *CoRR*, abs/1711.00614, 2017.

[Su *et al.*, 2019] Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 2828–2837. ACM, 2019.

[Vukotic *et al.*, 2016] Vedran Vukotic, Christian Raymond, and Guillaume Gravier. A step beyond local observations with a dialog aware bidirectional GRU network for spoken language understanding. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association, Interspeech 2016, San Francisco, CA, USA, September 8-12, 2016*, pages 3241–3244. ISCA, 2016.

[Xu *et al.*, 2018] Haowen Xu, Wenxiao Chen, Nengwen Zhao, Zeyan Li, Jiahao Bu, Zhihan Li, Ying Liu, Youjian Zhao, Dan Pei, Yang Feng, Jie Chen, Zhaogang Wang, and Honglin Qiao. Unsupervised anomaly detection via variational auto-encoder for seasonal KPIs in web applications. *CoRR*, abs/1802.03903, 2018.