Unsupervised Anomaly Detection in High-Dimensional Flight Data Using Convolutional Variational Auto-Encoder

Milad Memarzadeh*
milad.memarzadeh@nasa.gov
USRA, Data Sciences Group, NASA Ames Research Center
Moffett Field, California

Bryan Matthews
Ilya Avrekh
Daniel Weckler
bryan.l.matthews@nasa.gov
Ilya.avrekh-1@nasa.gov
daniel.i.weckler@nasa.gov
KBR Inc., Data Sciences Group, NASA Ames Research Center
Moffett Field, California

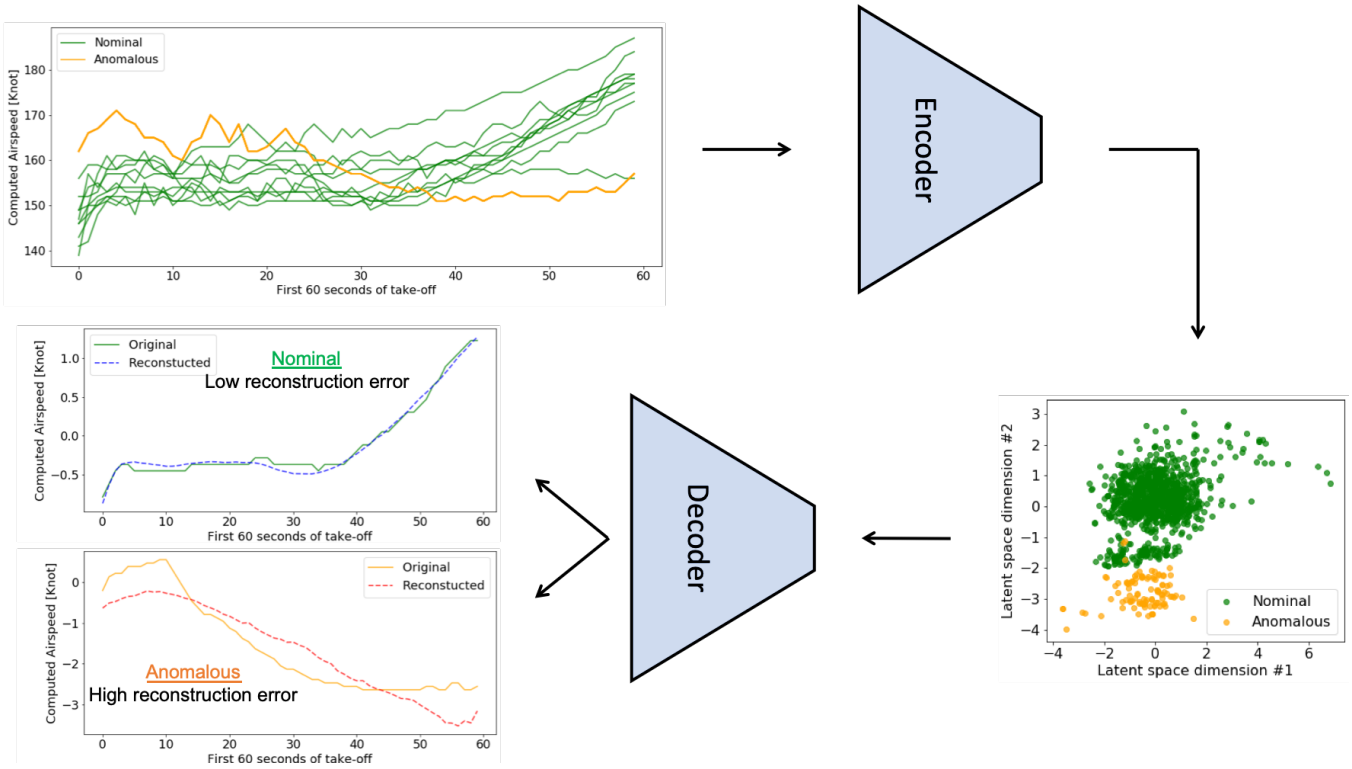

Figure 1: Overview of the proposed unsupervised anomaly detection approach.

ABSTRACT

The modern National Airspace System (NAS) is an extremely safe system. The industry has experienced a steady decrease in fatalities over the years. This can be contributed to both improved flight critical systems with redundant hardware and software protections as well as an increased focus on active monitoring and response to real time and historically identified vulnerabilities by implementing more resilient procedures and protocols. The main approach for identifying vulnerabilities in operations leverages domain expertise using knowledge about how the system should behave within the expected tolerances to known safety margins. This approach works well when the system has a well-defined operating condition. However, the operations in the NAS can be highly complex with various nuances that render it difficult to clearly pre-define all known safety vulnerabilities. With the advancement of data science and machine learning techniques, the potential to automatically identify emerging vulnerabilities in the observed operations has become more practical in recent years. The state-of-the-art

*Corresponding author.

KDD'20, August 22–27, San Diego, CA, USA
2020. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

anomaly detection approaches in aerospace data usually rely on supervised or semi-supervised learning. However, in many real-world problems such as flight safety creating labels for the data requires huge amount of efforts and is largely impractical. To address this challenge, we develop a Convolutional Variational Auto-Encoder (CVAE), an unsupervised learning approach for anomaly detection in high-dimensional heterogeneous time-series data. We validate performance of CVAE compared to the state-of-the-art supervised learning approach as well as unsupervised clustering-based approach using KMeans++ and kernel-based approach using One-Class Support Vector Machine (OC-SVM) on Yahoo!'s benchmark time series anomaly detection data. Finally, we showcase performance of CVAE on a case study of identifying anomalies in the first 60 seconds of commercial flights' take-offs using Flight Operational Quality Assurance (FOQA) data.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; Neural networks; • **Applied computing** → **Aerospace**.

KEYWORDS

anomaly detection, variational autoencoder, flight safety, time series

ACM Reference Format:
Milad Memarzadeh, Bryan Matthews, Ilya Avrekh, and Daniel Weckler. 2020. Unsupervised Anomaly Detection in High-Dimensional Flight Data Using Convolutional Variational Auto-Encoder. In *Proceedings of ACM SIGKDD (KDD'20).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

As the National Airspace System (NAS) has evolved over the years it has been able to accommodate commercial passenger demand while maintaining exceptional levels of safety. According to the National Transportation Safety Board (NTSB): since 2000 the accident rate per 100,000 flight hours has been cut in half, from 0.306 to 0.156 in 2018 [8]. While the number of passenger enplanements have increased 20% from 706 million in 2009 to 851 million in 2017, the number of departures has decreased 5% from 9.7 million to 9.3 million over the same period [7]. This trend has resulted in a historically high passenger load factor 82.3% in 2017 [31]. While the number of flights have remained relatively flat, the passenger load factor is approaching saturation and will result in more departing flights in the future. To remain at this historically low level of accidents per year the NAS will need to innovate and proactively identify operationally significant safety events that are currently not being tracked.

Identifying situations where unknown risk or vulnerabilities exists is not a trivial problem. Much of the knowledge of adverse events comes from after-the-fact analysis using forensic investigations to determine the root cause of an incident or accident such as the manual process that NTSB uses when investigating accidents [6]. In 2007 the Federal Aviation Administration (FAA) partnered with the MITRE Corp to develop the Aviation Safety Information and Sharing (ASIAS) system to archive air carrier flight data and promote proactive analytics that could identify safety risks in the NAS before they lead to a significant incident or accident. One aspect of the program acts as a repository for Flight Operational Quality Assurance (FOQA) data. These data are comprised primarily of 1 Hz recording for each flight, covering a variety of systems including: the state and orientation of the aircraft, positions and inputs of the control surfaces, engine parameters, and auto pilot modes and corresponding states. The data is acquired in real time on-board the aircraft and downloaded by the airline once the aircraft has reached the destination gate. These time series are analyzed by domain expert derived threshold based algorithms post-flight to flag known events that are deemed to be of operational significance by each of the airlines. These events are monitored over time to determine emerging trends or quantify safety improvements. The ASIAS program acts as an independent broker that does not have regulatory authority and can provide each airline a centralized assessment of their safety performance compared to other similar airlines in a de-identified context. However, in 2013 an Inspector General's (IG) report [26] found that the "system lacked advanced analytical capabilities" and tasked the FAA to further improve the system. In October of 2019 the IG began a follow-up review to assess the progress of ASIAS in addressing the IG's 2013 recommendations [27].

Improving the ability to identify emerging vulnerabilities in current operations helps to increase awareness of new threats. Proactively addressing safety requires developing, testing, and validating new approaches that can process and model large amounts of historically recorded heterogeneous data that describe the operations of millions of flights over multiple years and covering various diverse regions in the NAS. Data science and machine learning approaches have the potential to automatically identify anomalous events in these observed data. The events identified will still need to be reviewed and assessed by the subject matter experts familiar with the procedures to help better understand how the operations are carried out and the safety implications. Highlighting these possible vulnerabilities can be addressed by mitigating the contributing factors with countermeasures, such as improved pilot/controller training, or developing automation safety processes that, when in place, help to avoid states that result in an increased likelihood of an incident or accident that may result in damage to the aircraft, injury, or loss of life. It is important that any decision support tool has both low false positive and false negative rates, to ensure the user has trust in the system and takes appropriate actions.

In order to improve and automate identification of these vulnerabilities, we have developed an unsupervised machine learning algorithm that constructs models based on the observed operations and identifies operationally significant safety anomalies. This algorithm is demonstrated to have improved performance as compared to existing anomaly detection methods used in this domain. The paper is organized as follows: We cover related work in Section 2. In Section 3, a description of the proposed method is discussed with a background on the existing concepts used to construct the method and the innovative contribution we have made. In Section 4, we demonstrate the performance on the publicly available Yahoo! benchmark time series data and real world FOQA data against three existing methods: Kmeans++ [3], One-Class SVM [12], and ADOPT [20]. Finally in Section 5 we will discuss our conclusions and future work.

2 RELATED WORK

The standard anomaly detection technique in aerospace data is exceedance detection, in which specific parameters are compared with pre-defined thresholds identified based on the domain knowledge. The exceedance analysis is described in FAA document on FOQA program [1] and implemented in flight data monitoring software used by airlines and aviation equipment manufacturers (e.g. eFOQA from GE or AirFase from Teledyne). The exceedance detection method performs well on known issues, but is incapable of identifying unknown risks and vulnerabilities.

In order to be able to identify unknown risks and vulnerabilities, we need to go beyond simplistic rule-based thresholding approaches. Recent advancements in the field of machine learning have shed light on their application for identifying anomalies in aviation data. Generally, machine learning approaches used for anomaly detection can be categorized into supervised and unsupervised methods. The presence of the labels is a key differentiator between the two, and with the difficulties in obtaining labels even for known anomalies in aerospace data, the unsupervised approach often becomes the only feasible one. Unsupervised machine learning algorithms used for anomaly detection in aerospace data include proximity-based methods (nearest neighbors and clustering-based), support vector machines (SVM) and, more recently, deep learning methods.

Bay and Schwabacher [4] is among proximity-based approaches that develop an algorithm defining an anomaly as a point in the feature space whose nearest neighbors are far from it. This algorithm was applied to detect anomalies in Space Shuttle main engines. Another line of work rely on clustering methods, such as the Sequence Miner algorithm for discrete flight parameters (cockpit switch flips) [9] and Inductive Monitoring System (IMS) [19] for continuous parameters. These studies rely on identifying "normal" regions in the feature space, and then computing an anomaly score by measuring the distance between the observed data and these regions. In the investigation of the Space Shuttle Columbia disaster, IMS has been applied to the data from temperature sensors on the Shuttle's left wing, detecting in retrospect the damage after the foam impact [25]. The ClusterAD-Flight method [24] transforms FOQA time series data into high-dimensional vectors, making different flights comparable by sampling each flight parameter at fixed temporal or distance-based intervals starting from the anchoring event (e.g. time from takeoff or distance from touchdown) with subsequent clustering using the density-based spatial clustering algorithm.

One-class SVM (OC-SVM) is among other popular approaches that is an unsupervised approach developed for anomaly detection. OC-SVM constructs an optimal hyperplane separating normal data in the high dimensional kernel space by maximizing the margin between the origin and the hyperplane. This approach has been developed for anomaly detection in aviation data as well [13]. A major challenge in implementation of OC-SVM is the computational complexity of the kernel building step, which is quadratic with respect to the number of training examples. Moreover, they usually perform poorly in detecting short-duration anomalies.

Anomaly detection using deep neural networks has caught most of the attention recently reflecting rising trend in popularity of deep learning due to their flexibility and scalability. One of these approaches is Autoencoder (AE), which is a feed-forward multilayer neural network, trained to copy its input to its output by minimizing the reconstruction error. It could be viewed as a nonlinear generalization of Principal Component Analysis (PCA) that uses a multi-layer encoder network to transform the high-dimensional data into a low-dimensional latent space and a decoder network to recover the input data from the latent space [18]. Anomaly detection with AE uses the reconstruction error as an anomaly score. Reddy et al. [29] applied AE to raw time series data from multiple flight sensors by using sliding overlapping time windows to form input vectors (much earlier example of applying AE to spacecraft data can be found in [16]). Zhou et al. [38] implements AE with regularization term (called "robust AE") to eliminate outliers in case of lacking clean training data. The main difficulty of applying autoencoders is the choice of the right "degree of compression", i.e. dimensionality of the latent space and finding its right trade-off with over-fitting.

Work by Kingma and Welling [23] and Rezende et al. [30] bridged recent advancements in deep learning with variational inference by introducing Variational Auto-Encoder (VAE) (see details in section 3 of this paper). VAEs have been used for various applications, with anomaly detection becoming increasingly popular one. An and Cho [2] proposed anomaly detection method based on VAE with anomaly score as Monte Carlo estimate of the reconstruction log-likelihood, which they called "reconstruction probability". Haowen Xu et al [35] used this approach for detecting anomalies in univariate time series representing seasonal key performance indicators in web applications, with the input vector formed by applying sliding time window. Following success of using deep Recurrent Neural Networks (RNN) for machine learning applications with sequential data, some approaches of using VAE with RNN for anomaly detection in time series have been actively explored. They usually use long short-term memory (LSTM) constructs for encoder and decoder networks in VAE to handle temporal dependencies in data [10, 28, 33, 36, 37]. LSTM-VAE approach has been also applied for anomaly detection in telemetry data from Soil Moisture Active Passive (SMAP) satellite and Mars Curiosity rover [32]. Another example is ADOPT algorithm for mining precursors to "outlier flight events" [20]. However, VAE based on the RNN architecture are computationally expensive to train for high-dimensional time series and may overlook local temporal dependencies.

3 METHOD

Unsupervised detection of anomalous patterns in high-dimensional heterogeneous time series such as flight operational quality assurance (FOQA) data is an extremely challenging task. The model trained for this task must be able to capture the complex patterns in correlated heterogeneous data in order to identify anomalous trends. If there exist significant amount of labelled anomalous patterns within the data, then the problem can be approached using an appropriate supervised machine learning model such as ADOPT [20]. However, in many real-world problems such as flight safety, creating labels for the data requires huge amount of efforts and is largely expensive.

In this paper, we develop a Convolutional Variational Auto-Encoder (CVAE), specifically designed for anomaly detection in

high-dimensional heterogeneous time series data. VAEs [23] are a family of machine learning models combining deep neural networks with variational inference, where the model is comprised of two main parts: (1) an encoder, which maps the original data space, $x \in X$, into a compressed low-dimensional latent space, $z \in Z$, and (2) a decoder, which reconstructs the original data by sampling from the low-dimensional latent space. As illustrated in Figure 1, given all data entries, if we over-fit the model to the unbalanced training data, which contains significantly lower number of anomalous examples compared to the nominal ones, then CVAE is able to successfully learn the optimal mapping of the nominal data to the latent space, and reconstruct them with small error. However, for the anomalous data, this mapping to the latent space is not optimized and hence would result in significantly higher reconstruction error, which can be used as a metric to identify anomalous patterns. It should be noted that the level of over-fitting to the training data needs careful consideration, as the CVAE might be able to also over-fit to the anomalous patterns, if anomalies are present in the training data. This means that the reconstruction error for the anomalies would be as low as the nominal data, which is an undesirable outcome. We take inspiration from [17] to control the level of over-fitting by introducing a hyper-parameter in the loss function. We start with summarizing a basic understanding of variational inference and VAEs, and then explain the proposed model in detail.

3.1 Variational Inference

Let us assume this problem: given the original data, $x \in X$ and the latent variables $z \in Z$, the goal is to estimate the conditional density of the posterior of the latent variables, i.e., $p(z \mid x)$, which can be computed using the Bayes rule,

$$p(z \mid x) = \frac{p(z, x)}{p(x)} \tag{1}$$

The denominator in the above equation is called the evidence. In order to calculate this evidence, one needs to compute the following integral,

$$p(x) = \int_z p(z, x) dz \tag{2}$$

However, computing this integral is usually intractable. In order to approximate the posterior of the latent variables, i.e., $p(z \mid x)$, two paradigms are used: (1) Markov chain Monte Carlo (MCMC) [15], which uses sampling across an ergodic Markov chain on the latent variable z whose stationary distribution is the posterior $p(z \mid x)$; and (2) variational inference (VI) [5, 21], which uses optimization instead of sampling to approximate the posterior by minimizing the Kullback-Leibler (KL) divergence between the estimated posterior and the exact one,

$$q^*(z) = \text{argmin}_{q \in Q} \text{KL}\left(q(z) || p(z \mid x)\right) \tag{3}$$

While MCMC provides guarantees of producing samples from the exact posterior distribution, they are computationally intensive, specially when data sets are large and models are very complex. On the other hand, VI is faster and applicable to complex problems, while sacrificing the guarantee of convergence to the exact posterior.

Moreover, the objective defined in Eq. (3) is not tractable to compute as it requires computing the log of evidence, i.e., $\log p(x)$. To see this, we need to extend the definition of the KL divergence,

$$\begin{aligned} \text{KL}\left(q(z) || p(z \mid x)\right) &= \int q(z) \log \left(\frac{q(z)}{p(z \mid x)} \right) dz \\ &= \mathbb{E}_{q(z)}[\log q(z)] - \mathbb{E}_{q(z)}[\log p(z \mid x)] \\ &= \mathbb{E}_{q(z)}[\log q(z)] - \mathbb{E}_{q(z)}[\log p(z, x)] + \log p(x) \end{aligned} \tag{4}$$

Due to the dependence of KL divergence to the evidence, we cannot compute it. As a result, VI relies on optimizing an alternative objective, which is called the evidence lower bound (ELBO),

$$\text{ELBO}(q) = \mathbb{E}_{q(z)}[\log p(z, x)] - \mathbb{E}_{q(z)}[\log q(z)] \tag{5}$$

As noted, ELBO is the negative KL divergence (defined in Eq. (4)) plus $\log p(x)$, which is a constant when we take expectation with respect to $q(z)$. As a result, maximizing the ELBO is equivalent to minimizing the KL divergence, which is the main objective of VI's optimization, i.e., Eq. (3).

3.2 Variational Auto-Encoder

Variational auto-encoder (VAE) approximately optimizes the evidence defined in Eq. (2). It should be noted that VAEs are called autoencoders because their training objective resembles an encoder and a decoder [14], as we discuss later. Re-organizing the definition of the KL divergence in Eq. (4) we have,

$$\begin{aligned} \text{KL}\left(q_\phi(z \mid x) || p(z \mid x)\right) = \mathbb{E}_{q_\phi(z|x)}[&\log q_\phi(z \mid x) - \log p_\theta(x \mid z) \\ &- \log p(z)] + \log p_\theta(x) \end{aligned} \tag{6}$$

Where ϕ and θ are parameters of functions q and p that map X to Z (i.e., encoder) and Z to X (i.e., decoder), respectively. Using the KL divergence definition again Eq. (6) turns into,

$$\begin{aligned} \log p_\theta(x) - \text{KL}\left(q_\phi(z \mid x) || p(z \mid x)\right) = &\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] - \\ &\text{KL}\left(q_\phi(z \mid x) || p_\theta(z)\right) \end{aligned} \tag{7}$$

Eq. (7) is the key equation in VAEs: The left hand side is the term that we would like to optimize, which is the sum of the log-likelihood of the data X minus the error in approximating the true posterior $p_\theta(z \mid x)$ with the approximate one $q_\phi(z \mid x)$. The right hand side of the equation is equivalent to the definition of the ELBO in Eq. (5) and is an objective that we can optimize using stochastic gradient descent given the right choice of q (refer to [14] for further details). Hence, the objective function of the VAE is defined as follows,

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] - \text{KL}\left(q_\phi(z \mid x) || p_\theta(z)\right) \tag{8}$$

However, taking the gradient of $\mathcal{L}(\theta, \phi; x)$ with respect to ϕ is problematic specially for the first term. Kingma and Welling [23] propose a solution called reparameterization trick, which introduces variable $\epsilon \sim \mathcal{N}(0, I)$, and reformulates the objective function so that

the expectation is only with respect to fixed X and ϵ. This ensures the objective function to be deterministic and continuous in θ and ϕ, which makes the backpropagation with stochastic gradient descent possible.

Let the prior over the latent variables z be a standard Gaussian, i.e., $p_\theta(z) = \mathcal{N}(z; 0, I)$, and the variational approximate posterior also a multivariate Gaussian with a diagonal covariance function, $q_\phi(z \mid x) = \mathcal{N}(z; \mu, \sigma^2 I)$. Then the objective function in Eq. (8) becomes,

$$\begin{aligned}\mathcal{L}(\theta, \phi; x) = &\frac{1}{L}\sum_{l=1}^{L} \log p_\theta\left(x \mid z^{(l)}\right) + \\ &\frac{1}{2}\sum_{j=1}^{J}\left(1 + \log(\sigma_j^2) - \mu_j^2 - \sigma_j^2\right)\end{aligned} \tag{9}$$

Where, $z^{(l)} = \mu + \sigma\epsilon^{(l)}$, and $\epsilon^{(l)} \sim \mathcal{N}(0, I)$. It should be noted that the first term is a negative reconstruction error in the autoencoder terminology and the second term is the KL divergence of the multivariate Gaussian posterior from the standard Gaussian prior.

3.3 Convolutional Variational Auto-Encoder (CVAE)

Recently, there has been a growing interest in modifying the loss function of VAEs to improve the disentanglement of different dimensions of the latent space with the goal of having each latent space dimension correspond to a continuum of a meaningful domain specific attribute. Higgins et al. [17] formulates this problem as a constrained optimization problem,

$$\begin{aligned}&\max_{\theta,\phi} \mathbb{E}_{x\sim D}[\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] \\ &\text{s.t.} \quad \mathrm{KL}\left(q_\phi(z \mid x)||p(z)\right) < \epsilon\end{aligned} \tag{10}$$

And then uses Lagrangian KKT conditions to define,

$$\begin{aligned}\mathcal{F}(\theta, \phi, \beta; x, z) = &\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] - \\ &\beta\left(\mathrm{KL}\left(q_\phi(z \mid x)||p(z)\right) - \epsilon\right)\end{aligned} \tag{11}$$

Higgins et al. [17] finds that increasing β improves the disentanglement in the latent space dimensions, however, it decreases the reconstruction quality. More recent works [11, 22] have introduced extra terms to factorize latent space and improve the total correlation of the latent space dimensions, which has shown to improve the disentanglement of the latent space. Although, this disentanglement is easy to quantify and validate when dealing with imagery data, it turns out that such disentanglement is not quite clear when it comes to time series data. Inspired by the work of [17], we define the CVAE loss function as follows,

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] - \beta\mathrm{KL}\left(q_\phi(z \mid x)||p_\theta(z)\right) \tag{12}$$

Since β and ϵ in Eq. (11) are both positives, then $\mathcal{L}$ is the lower bound for $\mathcal{F}$: $\mathcal{F}(\theta, \phi, \beta; x, z) \geq \mathcal{L}(\theta, \phi; x, z, \beta)$. It should be noted that we do not introduce β to improve disentanglement in the latent space, rather we use it as a regularization hyper-parameter. Recall that the KL divergence term in the loss function penalizes posteriors of the latent variables that are far from the prior (which is standard normal distribution). As a result, one can imagine that hyper-parameter β serves as a metric on how much we want CVAE to over-fit on the training data. Given that we are dealing with a completely unsupervised approach in this article, it means our training data consist of both nominal and anomalous time series and there should be a trade-off in how much we want CVAE to over-fit the mapping to the latent space for this data. As a result, we treat β as a regularization hyper-parameter that needs tuning.

CVAE uses windowed time-series data as an input and applies series of convolutional operations with different filter sizes to incorporate local and global temporal dependence into account. Then the results of each series of convolution is concatenated together before going into the latent space. CVAE uses a similar architecture for both encoder and decoder. As a result, the decoder is comprised of series of deconvolution and up-sampling with different filter sizes. Figure 2 shows the general architecture used for all of the results presented in the paper.

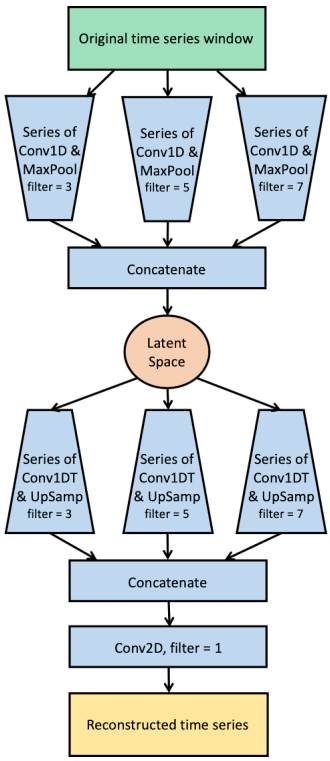

Figure 2: Network architecture of the Convolutional Variational Auto-Encoder (CVAE).

4 RESULTS AND DISCUSSION

We first comprehensively validate the performance of CVAE on the recently published benchmark data set of Yahoo!'s Data for time series anomaly detection [34]. This dataset is comprised of four different time series, A1-4, where A1 and A2 are univariate

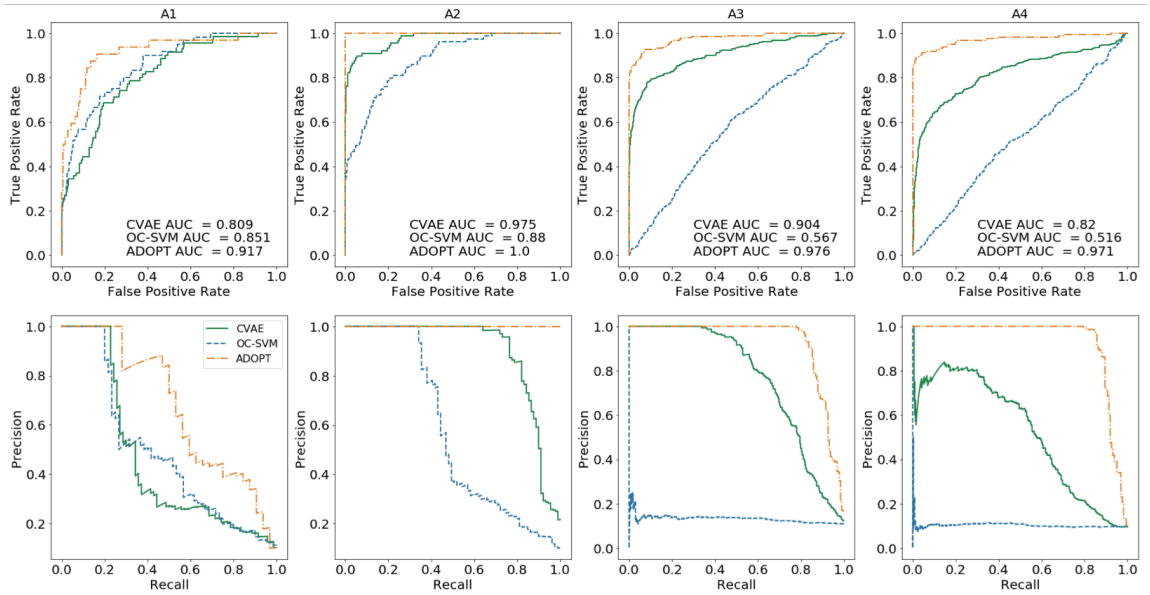

Figure 3: Performance comparison on Yahoo!'s benchmark data for CVAE, ADOPT and OC-SVM.

real and synthetic production traffic to some of the Yahoo! properties respectively, and A3 and A4 are synthetic multivariate time series with outlier and change-point anamalies, respectively. The multivariate data has additive noise and 12-hour, daily, and weekly seasonality associated with the actual values of the traffic.

Then, we evaluate the effect of the hyper-parameter β on the interpretability and the anomaly detection performance of CVAE. Finally, we validate CVAE's performance in identifying operationally significant anomalies in FOQA data, specifically on a case study of identifying anomalies in the first 60 seconds of commercial flight's take-offs.

We compare performance of the proposed CVAE to three alternatives: (a) ADOPT [20], which is a supervised anomaly detection approach based on RNN and can serve as a upper bound for the CVAE's performance (since CVAE is unsupervised, we expect to perform worse than ADOPT), and (b) two very well-known unsupervised anomaly detection methods, i.e. One-Class SVM (OC-SVM) [12], which is a unsupervised kernel-based classification algorithm, and Kmeans++ [3], which is an unsupervised clustering algorithm.

4.1 Validation on Yahoo!'s Data

We evaluated the best window size and dimension of the latent space for all of A1-4 benchmark time series and found the best performance using a 50-step window for A1-2 with 2-dimensional latent space, and a 20-step window for A3-4 with 100-dimensional latent space (we used the same window size for all approaches). Figure 3 shows the ROC curves (top panels) and precision-recall curves (bottom panels) for the performance of CVAE, ADOPT and OC-SVM across the four benchmark time series dataset. Since the training and testing data are imbalanced, meaning that there are significantly less samples of anomalous patterns compared to nominal ones, then the precision-recall curve is a better choice of comparison, however, for the sake of clarity, we have included the ROC curves with reported AUCs as well.

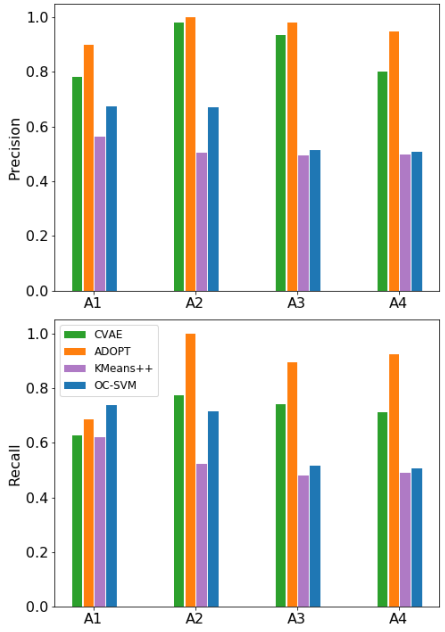

Figure 4: Anomaly detection in Yahoo's data using CVAE, ADOPT, OC-SVM and KMeans++.

As it can be seen, CVAE performs significantly good, compared to the ADOPT, which is a supervised approach, and outperforms OC-SVM significantly. The difference between the performance

of CVAE and OC-SVM is more significant in multi-variate data (A3-4), which kernel-based approaches such as OC-SVM struggle to identify anamolies accurately.

Figure 4 reports average precision and recall for CVAE, ADOPT, OC-SVM and KMeans++ on A1-4 datasets. As shown in the figure, CVAE performs significantly better than OC-SVM ($\sim$ 51% higher precision and $\sim$ 21% higher recall) and KMeans++ ($\sim$ 72% higher precision and $\sim$ 38% higher recall), while at the same time performing close to the supervised ADOPT ($\sim$ 10% lower precision and $\sim$ 17% lower recall).

4.2 Effect of Regularization Parameter

In this section, we evaluate the effect of hyper-parameter β on performance of CVAE. We use the case study of identifying anomalies in the first 60 seconds of commercial flight's take-off for the validation. Our database for this case study consists of 30,000 nominal take-offs and 1000 anomalous ones identified by the subject matter experts. The surrogate labels for operationally significant anomalies in the take-off are created based on the drop in the computed airspeed. For the purpose of illustrations and simplicity in this section, we evaluate the anomaly detection performance based on using only the computed airspeed variable from the FOQA data (using only computer airspeed variable allows us to set the dimension of the latent space as 2D, which is ideal for illustration). We will extend this to anomaly detection using the entire FOQA data later on.

Figure 5 visualizes the effect of β on the performance of anomaly detection (A-B) as well as the distribution of the nominal and anomalous data in the latent space (C-D). We have tested different values of $\beta \in [0, 1]$ and have shown only two values for the sake of illustration here: (a) $\beta = 0.001$: as an example of a model with a high variance, and (b) $\beta = 0.5$: as an example of a model with a high bias. As it is clear, the model with a high bias significantly outperforms the model with a high variance. The reason for this is the existence of anomalous examples in the training data; a model with a high variance (i.e., $\beta = 0.001$) is able to learn the optimal mapping from the original data space to the latent space for both nominal and anomalous examples and as a result, reconstructs the anomalous data with low error as well (Figure 6). On the other hand, the model with a high bias, only optimizes this mapping for those data that represent the popular trend of the nominal examples (Figure 6), and hence result in a high reconstruction error for the anomalous examples.

It should be noted that this is not the case for all datasets and that is the reason we treat β as a hyper-parameter that needs tuning. For example, in the case of Yahoo!'s data presented in Section 4.1, the best performance was obtained by the model with a high variance ($\beta = 0.001$). The reason is that there are significant differences in the patterns of the time series in the Yahoo!'s data, which requires a model with a high variance to capture them. In Figure 6, True Negative and True Positive are the nominal and anomalous data that are labelled correctly, respectively.

On the other hand, although the model with a high bias performs better in anomaly detection, the interpretability of the distributions of anomalous and nominal examples in the latent space completely vanishes (Figure 5 C-D). The model with high variance is able to distinguish the anomalous and normal data very good and as a result, there is a clear separation between the samples from the posterior distribution of the normal and anomalous data in the latent space. This separation easily disappears as we increase the hyper-parameter β, which forces the posterior of the latent space to be very close the prior (standard Normal distribution).

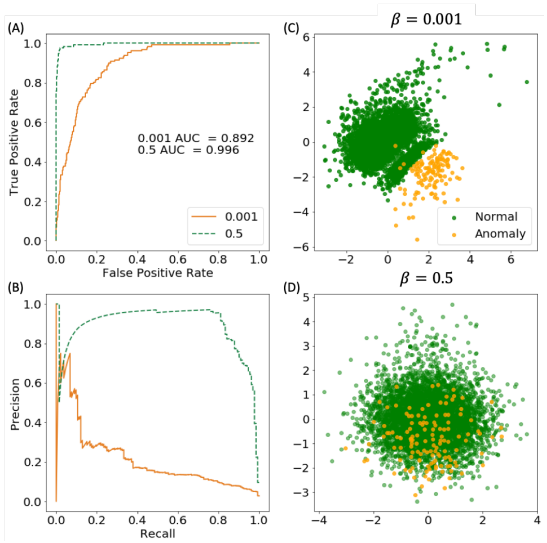

Figure 5: This figure illustrates the effect of hyper-parameter β on the interpretability and performance of the CVAE.

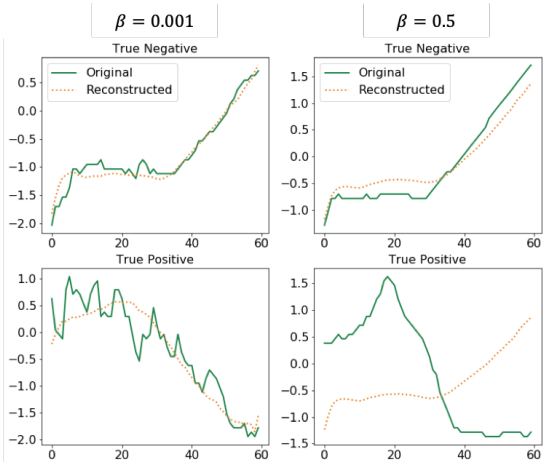

Figure 6: This figure illustrates the effect of hyper-parameter β on the reconstruction quality for two example data.

4.3 Anomaly Detection in Flight's Data

As discussed before, we use the case study of identifying anomalies in the first 60 seconds of commercial flight's take-off for the

validation. We group the FOQA data of these flight records into five groups based on the domain expert opinion, where groups represent roll attitude, altitude information, pitch attitude, speed information, and yaw attitude. CVAE sends each group into its own encoder (as illustrated in Figure 2) and concatenates the outcome of each encoding before passing the information to the shared latent space. After that, each group have its own decoder to reconstruct the inputs by sampling from the shared latent space. This grouping is done purely based on the domain expert knowledge without any correlation analysis, so that there is no supervision from data involved in the process.

In addition to ADOPT, OC-SVM and KMeans++, in Figure 7 we compare the performance of unsupervised CVAE to the semi-supervised version as well. In the semi-supervised approach, we only train CVAE on the nominal data (meaning that our training data does no include any anomalies), which is an approach that most of literature pursue [2, 10, 28, 32]. As it can be seen, unsupervised CVAE (CVAE-Un) outperforms both KMeans++ and OC-SVM and achieves higher precision, recall and F1-score. A more significant difference is achieved when the semi-supervised CVAE (CVAE-Semi) is used, where it achieves ∼ 25% higher precision and recall compared to the unsupervised CVAE and performs significantly close to the results obtained by supervised ADOPT (∼ 10% lower precision and recall).

These results show a promise in first steps of developing and deploying an unsupervised and scalable machine learning algorithm basing upon recent advancements in VAEs, that is able to identify anomalies in high-dimensional flight time series with reasonable accuracies.

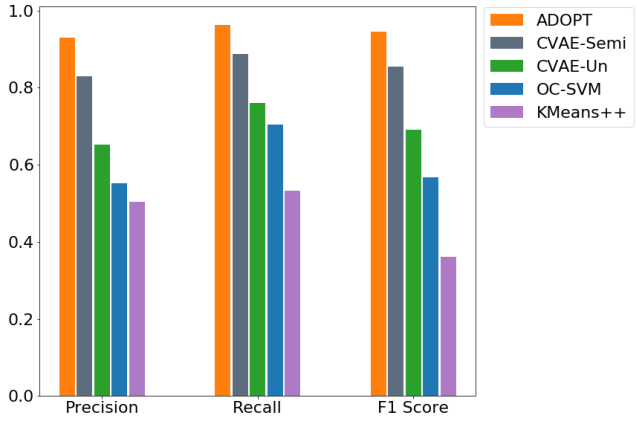

Figure 7: Validating the performance of CVAE on the FOQA data.

5 CONCLUSIONS

In order to improve and automate identification of unknown vulnerabilities in flight's operations, we have developed an unsupervised machine learning approach for identifying operationally significant anomalies in high-dimensional heterogeneous flight's time-series. The proposed approach constructs models based on the observed operations and identifies operationally significant safety anomalies. This algorithm is demonstrated to have improved performance as compared to existing anomaly detection methods used in the aviation domain. Majority of approaches in the aerospace literature either rely on rule-based thresholding or supervised learning approach. Although the supervised approaches show a good performance, creating labels for the data in aviation requires huge amount of efforts and is largely impractical. Our approach bases upon recent advancements in combination of deep learning and variational inference to develop the Convolutional Variational Auto-Encoder (CVAE), an unsupervised approach for anomaly detection in high-dimensional heterogeneous time-series data (Figure 1).

We validate CVAE compared to the several supervised and unsupervised approaches used in the literature for anomaly detection on several datasets. Validating on Yahoo!'s benchmark time series anomaly detection database, we show that (Figure 4) CVAE performs significantly close to the supervised approaches (∼ 10% lower precision, and ∼ 17% lower recall) and outperforms unsupervised clustering and kernel-based approaches (on average ∼ 62% higher precision and ∼ 30% higher recall). Moreover, we illustrates the effect of hyper-parameters in the CVAE model on the interpretability of the findings and performance of anomlay detection (Figures 5 and 6).

Application of CVAE to anomaly detection in Flight Operational Quality Assurance (FOQA) shows promise in further development of this line of work for anomaly detection in high-dimensional heterogeneous time series. Specifically by designing a case study of anomaly detection in the first 60 seconds of commercial flight's take-off (Figure 7), we show that CVAE outperforms clustering and kernel-based anomaly detection approaches (on average ∼ 24% higher precision and ∼ 26% higher recall). The performance significantly improves when the semi-supervised CVAE approach is used (on average ∼ 58% higher precision and ∼ 46% higher recall) and it performs significantly close to the supervised approach (∼ 11% lower precision and ∼ 7% lower recall).

Future Work: potential next steps will focus on developing an architecture to handle different heterogeneous time series data such as binary channels or categorical inputs. One possible approach may be to map multiple inputs into a state space representation to capture the changes in the time series modes. Additionally, evaluating the method for scalability and practical deployment of the algorithm on more complex operationally significant real-world data sets is something that would need to be tested and validated before the algorithm can be adopted into an existing vulnerability discovery program.

6 ACKNOWLEDGMENTS

This research is supported by the NASA Airspace Operation and Safety Program and the NASA System-wide Safety Project. We would also like to thank Dr. Hamed Valizadegan, and Marc-Henri Bleu-Laine for their insights and comments in developing and testing the algorithm.

REFERENCES

[1] Federal Aviation Administration. 2004. Flight Operational Quality Assurance. *Technical Report* 120-82 (2004). https://www.faa.gov/documentLibrary/media/Advisory_Circular/AC_120-82.pdf

[2] Jinwon An and Sungzoon Cho. 2015. Variational Autoencoder based Anomaly Detection using Reconstruction Probability. *SNU Data Mining Center, Technical*

Report (2015).
[3] David Arthur and Sergei Vassilvitskii. 2007. K-Means++: The advantages of careful seeding. *Proc. Symp. Discrete Algorithms* (2007), 1027–1035.
[4] Stephen D. Bay and Mark Schwabacher. 2003. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (2003), 29–38. https://doi.org/10.1145/956750.956758
[5] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. 2017. Variational Inference: A Review for Statisticians. *J. Amer. Statist. Assoc.* 112 (2017), 859–877.
[6] National Transportation Safety Board. 2002. National Transportation Safety Board Aviation Investigation Manual Major Team Investigations. https://www.ntsb.gov/investigations/process/Documents/MajorInvestigationsManual.pdf
[7] National Transportation Safety Board. 2017. US Transportation Fatality Statistics. https://www.ntsb.gov/investigations/data/Pages/AviationDataStats2017.aspx
[8] National Transportation Safety Board. 2019. Annual Summaries of US Civil Aviation Accidents. https://www.ntsb.gov/investigations/data/Documents/AviationAccidentStatistics_1999-2018_20191101.xlsx
[9] Suratna Budalakoti, Ashok N. Srivastava, and Matthew E. Otey. 2009. Anomaly Detection and Diagnosis Algorithms for Discrete Symbol Sequences with Applications to Airline Safety. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* 39, 1 (2009), 101–113.
[10] Run-Qing Chen, Guang-Hui Shi, Wan-Lei Zhao, and Chang-Hui Liang. 2019. Sequential VAE-LSTM for Anomaly Detection on Time Series. *arXiv:1910.03818* (2019).
[11] Tian Q. Chen, Xuechen Li, Roger B. Grosse, and David K. Duvenaud. 2018. Isolating Sources of Disentanglement in Variational Autoencoders. *Advances in Neural Information Processing Systems (NeurIPS)* 31 (2018), 2610–2620.
[12] Yunqiang Chen, Xiang Sean Zhou, and T.S. Huang. 2001. One-class SVM for learning in image retrieval. *Proceedings 2001 International Conference on Image Processing* (2001), 34–37. https://doi.org/10.1109/ICIP.2001.958946
[13] Santanu Das, Bryan Matthews, Ashok N. Srivastava, and Nikunj Oza. 2010. Multiple Kernel Learning for Heterogeneous Anomaly Detection: Algorithm and Aviation Safety Case Study. *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2010), 47–56.
[14] Carl Doersch. 2016. Tutorial on Variational Autoencoders. *arXiv* (2016). https://doi.org/arXiv:1606.05908
[15] Alan E. Gelfand and Adrian F.M. Smith. 1990. Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* 85 (1990), 398–409.
[16] T-H Guo and J. Musgrave. 1995. Neural network based sensor validation for reusable rocket engines. *Proceedings of 1995 American Control Conference - ACC'95* (1995). https://doi.org/10.1109/ACC.1995.520974
[17] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. β-VAE: Learning Basic Visual Concepts With A Constrained Variational Framework. *International Conference on Learning Representations (ICLR)* (2017).
[18] G.E. Hinton and R.R. SAlakhutdinov. 2006. Reducing the Dimensionality of Data with Neural Networks. *Science* 313, 5786 (2006), 504–407.
[19] David L. Iverson. 2004. Inductive System Health Monitoring. *Proceedings of the International Conference on Artificial Intelligence* (2004).
[20] Vijay Manikandan Janakiraman. 2018. Explaining Aviation Safety Incidents Using Deep Temporal Multiple Instance Learning. *KDD '18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2018), 406–415.
[21] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. 1999. Introduction to variational methods for graphical models. *Machine Learning* 37 (1999), 183–223.
[22] Hyunjik Kim and Andriy Mnih. 2018. Disentangling by Factorising. *International Conference on Machine Learning (ICML)* (2018).
[23] Diederik P. Kingma and Max Welling. 2013. Auto-Encoding Variational Bayes. *International Conference on Learning Representations (ICLR)* (2013).
[24] Lishuai Li, Santanu Das, R. John Hansman, Rafael Palacios, and Ashok N. Srivastava. 2015. Analysis of Flight Data Using Clustering Techniques for Detecting Abnormal Operations. *Journal of Aerospace Information Systems* 12, 9 (2015), 587–598.
[25] Bryan Matthews, Ashok N. Srivastava, John Schade, Dave Schleicher, Kennis Chan, Richard Gutterud, and Mike Kiniry. 2013. Discovery of Abnormal Flight Patterns in Flight Track Data. *Proceedings of 2013 Aviation Technology, Integration, and Operations Conference* (2013), 4386.
[26] Office of Inspector General Audit Report. 2014. FAA's Safety Data Analysis and Sharing System Shows Progress, but More Advanced Capabilities and Inspector Access Remain Limited. https://www.oig.dot.gov/sites/default/files/FAA%20ASIAS%20System%20Report%5E12-18-13.pdf
[27] Office of Inspector General Audit Report. 2019. INFORMATION: Audit Announcement | FAA's Implementation of the Aviation Safety Information Analysis and Sharing (ASIAS) System. https://www.oig.dot.gov/sites/default/files/Audit%20Announcement%20-%20FAA%20ASIAS.pdf
[28] Daehyung Park, Yuuna Hoshi, and Charles C. Kemp. 2018. A Multimodal Anomaly Detector for Robot-Assisted Feeding Using an LSTM-Based Variational Autoencoder. *IEEE Robotics and Automation Letters* 3, 3 (2018), 1544–1551.
[29] Kishore K. Reddy, Soumalya Sarkar, Vivek Venugopalan, and Michael Giering. 2016. Anomaly Detection and Fault Disambiguation in Large Flight Data: A Multi-modal Deep Auto-encoder Approach. *Annual Conference of the Prognostics and Health Monitoring Society* 7 (2016).
[30] Danilo J. Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. *Proceedings of the 31st International Conference on Machine Learning (ICML-14)* (2014), 1278–1286.
[31] Michael J. Sprung, Matthew Chambers, and Sonya Smith-Pickel. 2018. Transportation Statistics Annual Report 2018. https://rosap.ntl.bts.gov/view/dot/37861 Technical Report.
[32] Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. 2019. Robust Anomaly Detection for Multivariate Time Series through Stochastic Recurrent Neural Network. *KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2019), 2828–2837.
[33] Xuhong Wang, Ying Du, Shijie Lin, Ping Cui, Yuntian Shen, and Yupu Yang. 2019. adVAE: A self-adversarial variational autoencoder with Gaussian anomaly prior knowledge for anomaly detection. *Knowledge-Based Systems* (2019), 105187. https://doi.org/10.1016/j.knosys.2019.105187
[34] Yahoo! Webscope. 2019. dataset ydata-labeled-time-series-anomalies-v10. (2019). https://webscope.sandbox.yahoo.com/catalog.php?datatype=s&did=70
[35] Haowen Xu, wenxiao Chen, Nengwen Zhao, Zeyan Li, Jiahao Bu, Zhihan Li, Ying Liu, Youjian Zhao, Dan Peii, Yang Feng, Jie Chen, Zhaogang Wang, and Honglin Qiao. 2018. Unsupervised Anomaly Detection via Variational Auto-Encoder for Seasonal KPIs in Web Applications. *Proceedings of the 2018 World Wide Web Conference on World Wide Web* (2018), 187–196.
[36] Chunkai Zhang and Yingyang Chen. 2019. Time Series Anomaly Detection with Variational Autoencoders. *arXiv:1907.01702* (2019).
[37] Chuxu Zhang, Dongjin Song, Yucong Chen, Xinyang Feng, Cristian Lumerzanu, Wei Cheng, Jingchao Ni, Bo Zhang, Haifeng Chen, and Nitesh V. Chawla. 2019. A Deep Neural Network for Unsupervised Anomaly Detection and Diagnosis in Multivariate Time Series Data. *Proceedings of AAAI-19* (2019), 1409–1416.
[38] Chong Zhou and Randy C. Paffenroth. 2017. Anomaly Detection with Robust Deep Autoencoders. *KDD '17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2017), 665–674.