Resilient VAE: Unsupervised Anomaly Detection at the SLAC Linac Coherent Light Source

Ryan Humble¹,* William Colocho², Finn O’Shea², Daniel Ratner², and Eric Darve¹,³

¹Institute for Computational and Mathematical Engineering, Stanford University
²SLAC National Laboratory
³Department of Mechanical Engineering, Stanford University

Abstract. Significant advances in utilizing deep learning for anomaly detection have been made in recent years. However, these methods largely assume the existence of a normal training set (i.e., uncontaminated by anomalies) or even a completely labeled training set. In many complex engineering systems, such as particle accelerators, labels are sparse and expensive; in order to perform anomaly detection in these cases, we must drop these assumptions and utilize a completely unsupervised method. This paper introduces the Resilient Variational Autoencoder (ResV AE), a deep generative model specifically designed for anomaly detection. ResV AE exhibits resilience to anomalies present in the training data and provides feature-level anomaly attribution. During the training process, ResV AE learns the anomaly probability for each sample as well as each individual feature, utilizing these probabilities to effectively disregard anomalous examples in the training data. We apply our proposed method to detect anomalies in the accelerator status at the SLAC Linac Coherent Light Source (LCLS). By utilizing shot-to-shot data from the beam position monitoring system, we demonstrate the exceptional capability of ResV AE in identifying various types of anomalies that are visible in the accelerator.

1 Introduction

Anomaly detection (AD), which is the task of finding abnormal data or events, is a critical task for nearly all data-heavy, complex systems, such as industrial facilities, manufacturing, and large-scale science experiments [1–4]. These systems can produce thousands of real-time signals, which quickly overwhelm human operators that seek to monitor system performance. Identifying failures in these systems is a critical task, as failures can result in faulty outputs or cause damage to components. Unfortunately, the same complexity that overwhelms operators ensures that labeled data is rare or nonexistent and expensive to acquire.

In this work, we focus on detecting anomalies at SLAC’s Linac Coherent Light Source (LCLS), which is a free-electron laser (FEL) system that enables users to take X-ray snapshots of microscopic phenomena. Since LCLS loses approximately 3% of availability annually to unplanned downtime and experiences additional beam degradation without downtime, the task of identifying failures is critical to maximizing the amount of stable X-ray beam delivered to user experiments and thereby maximizing the scientific contribution of LCLS. Unfortunately,
detecting anomalies at LCLS is challenging because of the data volume and lack of labeled data. This combination makes most traditional AD methods [5–8] challenging to apply. Deep learning approaches [9–15] are often quite effective in this high-volume, high-dimensional setting; however, they nearly always assume a normal-only training set. Since we lack labels, our training data will inevitably be contaminated by anomalies.

Our work’s main contributions are: (i) introduction of a new deep generative model for anomaly detection, called Resilient VAE, to cope with training data contamination (Section 2); and (ii) its application to identifying anomalies in accelerator status at LCLS (Section 3).

2 Resilient VAE

We consider a particular type of anomaly detection task where the data is high-dimensional and completely unlabeled. We suppose a dataset \(D = \{x\}\), where each sample is \(x \in \mathbb{R}^D\) and \(D\) is the number of features. We suppose an (unknown) anomaly process that contaminates samples and features there-within, in our dataset. In particular, we assume that any training dataset has already been contaminated by anomalies, so there is no normal-only training set.

2.1 Background: Variational autoencoders for anomaly detection

Variational autoencoders (VAEs) [16] are probabilistic generative models designed for efficient inference and learning. The data is modeled as \(p(x) = \mathbb{E}_{p(z)}[p_\theta(x|z)]\) where \(p(z)\) is a prior over the latent representation \(z \in \mathbb{R}^K\) and \(p_\theta(x|z)\) is the decoder model. Typically \(K < D\), so the latent space acts as an information bottleneck, thereby implicitly assuming our data has a low-dimension representation. When \(p_\theta(x|z)\) is parameterized as a deep neural network, both the marginal likelihood \(p(x)\) and posterior likelihood \(p(z|x)\) are typically intractable [16]; thus, approximate variational inference is required, via an encoder model \(q_\phi(z|x)\). The resulting variational bound on the marginal log-likelihood, called the evidence lower bound (ELBO), is

\[
\log p(x) \geq \mathcal{L}_{\text{VAE}}(x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{\text{KL}}(q_\phi(z|x)||p(z)),
\]

and the lower bound \(\mathcal{L}_{\text{VAE}}(x)\) is maximized w.r.t. to the parameters \(\theta, \phi\). We refer to the first term, \(\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]\), as the reconstruction likelihood, and the second term, \(D_{\text{KL}}(q_\phi(z|x)||p(z))\), as the prior-matching penalty.

To apply the VAE model to anomaly detection, prior work most commonly uses the reconstruction likelihood to identify anomalies, where anomalies are assumed to be poorly reconstructed [17]. Furthermore, previous studies usually train the VAE model exclusively on normal data, as otherwise, the VAE might learn to reconstruct anomalous examples present in the training data. In fact, the VAE loss is weighted towards the anomalous examples, as both the reconstruction likelihood and prior-matching penalties are analogous to minimizing an \(L_2\) norm, which is not robust to outliers.

2.2 Mixture model

Since the standard VAE is not robust to anomalies and we assume our training data is contaminated by anomalies, we propose an extension of the VAE model, which we call Resilient VAE (ResVAE), that aims to distinguish between normal and anomalous samples during training using a mixture model approach.

We begin by adding several binary latent variables to the standard VAE: a sample mixing variable \(v \in \{0, 1\}\) and feature mixing variables \(w_d \in \{0, 1\}\), where 1 means inlier and 0 outlier.
These mixing variables control whether an inlier or outlier path through the ResVAE model is taken, as shown in Figure 1. With the new latent variables, our data is now modeled as 

\[ p(x) = \mathbb{E}_{p(v,w,z)}[p(x|v,w,z)]; \]

we use the variational posterior \( q(v,w,z|x) \) for inference. Our model supposes Bernoulli distributions for the mixing variables, a mixture prior on the \( z \) latent variable, and encoder and decoder mixture models:

\[
\begin{align*}
p(v) &= \text{Ber}(v; \rho), & q_\gamma(v|x) &= \text{Ber}(v; \gamma(x)), \\
p(w_d|v) &= \text{Ber}(w_d; \alpha_d)^v \text{Ber}(w_d; 0)^{1-v}, & q_\epsilon(w_d|v,x) &= \text{Ber}(w_d; \pi_d(x))^v \text{Ber}(w_d; 0)^{1-v}, \\
p(z|v) &= p(z)^v p_\phi(z)^{1-v}, & q(z|v,x) &= q_\phi(z|x)^v q_\omega(z|x)^{1-v}, \\
p(x|v,w,z) &= \prod_{d=1}^{D} p_\theta(x_d|z)^{w_d} p_\phi(x_d|z)^{1-w_d},
\end{align*}
\]

where \( \rho, \alpha_d \in [0,1] \) can be used to capture prior knowledge on the degree of contamination in the training data. We revisit these prior parameters in Section 2.4. The functions \( \gamma(x) \) and \( \pi_d(x) \) yield estimates of the mixing variables \( v \) and \( w_d \). As we show in Section 2.3, they have simple closed-form expressions as functions of the encoder and decoder parameters; this avoids the need for \( \gamma \) and \( \pi_d \) to be parameterized as separate neural networks.

Our mixture model introduced several outlier distributions—\( p_o(z), p_\phi(x|z), \) and \( q_\omega(z|x) \)—whose purpose is to provide an alternate model for outlier data, thereby allowing the inlier encoder \( q_\phi(z|x) \) and decoder \( p_\theta(x|z) \) to focus on representing only the inlier data in our training data. Concretely, we define the inlier and outlier distributions as

\[
\begin{align*}
p(z) &= \mathcal{N}(z; 0, I), & p_\theta(x_d|z) &= p(x_d; \eta_d(z; \theta)), & q_o(z|x) &= \mathcal{N}(z; \mu(x; \phi), \Sigma(x; \phi)), \\
p_o(z) &= \mathcal{N}(z; 0, \delta_z^2 I), & p_\phi(x_d|z) &= p(x_d; \eta_d(z; \theta_0)/\delta_z^2), & q_\omega(z|x) &= q_{\theta_0}(z|x),
\end{align*}
\]

where \( \delta_x, \delta_z \) are hyperparameters of the method, \( p_\theta(x_d|z) \) is taken to be an exponential family distribution with natural parameters \( \eta_d(z; \theta), \theta_0 \) is a copy of \( \theta \), and \( q_{\theta_0} \) is a copy of \( \phi \) (where the copy prevents gradients via automatic differentiation). (We refer the reader to a longer version of the paper [18] for an explanation of these modeling choices and subsequent derivations.)

2.3 ResVAE loss and coordinate ascent

Instead of maximizing the ELBO, we adopt the constrained-optimization approach of \( \beta \)-VAE [19], which introduces several coefficients regularizing the prior-matching penalties. Our objective is then to maximize the following loss function:

\[
\max_{\phi,\theta,\omega,\theta_0,\delta_\theta,\delta_z} \mathcal{L}(x) = \mathbb{E}_{q_{\theta}(v|x)}[\mathbb{E}_{q(z|v)}[\log p(x|v,w,z)]] - \beta_1 \mathbb{E}_{q_v(x)}[D_{KL}(q(z|x,v)||p(z|v))] \\
- \sum_{d=1}^{D} \frac{1}{\beta_{2,d}} \mathbb{E}_{q_{\epsilon}(d|x)}[D_{KL}(q_{\epsilon_d}(w_d|v,x)||p(w_d|v))] - \frac{1}{\beta_3} D_{KL}(q_\phi(v|x)||p(v)).
\]

Figure 1. Network diagram for our proposed ResVAE. ResVAE adds a sample mixing variable \( v \) and feature mixing variables \( w_d \) and splits the encoder and decoders into an inlier (solid path) and outlier (dotted path) version.
\( \beta_1 \) plays an identical role to that in \( \beta \)-VAE \([19]\), controlling the latent information bottleneck against the reconstruction objective. We discuss the other regularization coefficients \( \beta_{2,d}, \beta_3 \) in Section \( 2.4 \).

We can simplify this loss function by performing coordinate ascent (as in \([20, 21]\)) on \( \gamma(x) \) and \( \pi(x) \) since \( L(x) \) is convex in these variables for each \( x \). Their optimal values are:

\[
\pi_d(x) = \sigma(\beta_{2,d} r_d(x) + \logit(\alpha_d)), \quad \gamma(x) = \sigma(\beta_3 g(x) + \logit(\rho)),
\]

where \( \sigma(\cdot) \) is the sigmoid function, \( \logit(\cdot) \) is the logit function, and \( r_d(x) \) and \( g(x) \) can be interpreted as likelihood ratio tests \([22]\) (i.e., \( r_d(x) > 0 \) suggests \( x \) is drawn from the inlier model instead of the outlier model and the opposite for \( r_d(x) < 0 \)).

Using the closed form for \( \pi_d(x) \) and \( \gamma(x) \) and the given outlier distributions, and removing untrainable terms, our ResVAE loss function becomes:

\[
\max_{\phi, \beta} L_{\text{ResVAE}}(x) = \gamma(x) \left[ \sum_d \pi_d(x) \mathbb{E}_{q_\phi(x|z)} \left[ \log p_\theta(x_d|z) \right] - \beta_1 D_{KL}(q_\phi(z|x)p(z)) \right].
\]

This is nearly the form of the standard VAE loss (or \( \beta \)-VAE loss \([19]\)), with the addition of several weighting factors that either preserve the sample (or features) in the loss function or remove it. \( \gamma(x) \) weights entire samples and \( \pi_d(x) \) weights individually features.

### 2.4 Learnable logistic parameters

We can interpret the form of \( \pi_d(x) \) and \( \gamma(x) \) as a logistic regression, where the input variables are the likelihood ratios \( r_d(x) \) and \( g(x) \) respectively and the parameters are the regularization coefficients \( \beta_{2,d}, \beta_3 \) and contamination prior parameters \( \alpha_d, \rho \). Ideally, \( \pi_d(x) \) and \( \gamma(x) \) should be nearly 1 for inliers and 0 for outliers. Achieving this requires setting the logistic parameters appropriately.

To ensure at all stages of training that \( \pi_d(x) \) and \( \gamma(x) \) are expressive, we propose to learn the logistic parameters. Since we can interpret \( \pi_d(x) \) and \( \gamma(x) \) as discriminators (as in generative adversarial nets \([23]\)), we train them as binary classifiers with the binary cross entropy loss:

\[
\max_{\beta_{2,d}, \alpha_d} L_{\pi_d}(x) = \mathbb{E}_{p(w_d|v=1,x)} \left[ \log q_\pi(w_d|v=1,x) \right], \quad \max_{\beta_3, \rho} L_{\gamma}(x) = \mathbb{E}_{p(v|x)} \left[ \log q_\gamma(v|x) \right],
\]

where \( q_\pi(w_d|v=1,x) \) depends on the logistic parameters \( \beta_{2,d}, \alpha_d \) through the definition of \( \pi_d(x) \) and \( q_\gamma(v|x) \) similarly depends on \( \beta_3, \rho \) through \( \gamma(x) \). The label distributions of the binary cross entropy terms are the posterior distributions \( p(w_d|v=1,x) = \text{Ber}(w_d; \tilde{w}_d(x)) \) and \( p(v|x) = \text{Ber}(v; \tilde{v}(x)) \). However, since these posteriors are intractable (which originally motivated the use of approximate posteriors), the labels \( \tilde{w}_d(x) \) and \( \tilde{v}(x) \) must be approximated.

In defining our label approximations, we want to achieve two goals. First, we want the empirical distributions of \( \pi_d(x) \) and \( \gamma(x) \) to match our prior knowledge of the dataset’s contamination. Since \( \rho, \alpha_d \) are now learned as logistic parameters instead of specified as contamination priors, we re-introduce the idea of a contamination prior through new prior label distributions \( f_{w_d}(w_d) \) and \( f_v(v) \) over the entire dataset. Second, since we assume that inlier examples have higher likelihood ratios \( r_d(x) \) and \( g(x) \) than outlier examples, we want \( \pi_d(x), \gamma(x) \) to be increasing in the likelihood ratios \( r_d(x), g(x) \) respectively. Therefore, we approximate the labels as:

\[
\tilde{w}_d(x) \approx F_{w_d} \left( \mathbb{P}_{y \sim \mathcal{D}} (\pi_d(x) \geq \pi_d(y)) \right), \quad \tilde{v}(x) \approx F_v \left( \mathbb{P}_{y \sim \mathcal{D}} (\gamma(x) \geq \gamma(y)) \right),
\]

where the inner rank statistic terms ensure the property of increasing in likelihood ratio and the cumulative distribution functions (CDFs) \( F_{w_d}(w_d) \) and \( F_v(v) \) (of \( f_{w_d}(w_d) \) and \( f_v(v) \) respectively) ensure the prior-empirical label distribution matching.
3 Application to LCLS

We now apply our ResVAE method to detecting anomalies at LCLS. We consider pulse-by-pulse measurements of the beam, as measured by stripline beam position monitors (BPMs) [24] and recorded by the beam-synchronous acquisition system [25]. Each of the 150 BPMs records three signals (X, Y, and TMIT) for each electron pulse, where X and Y measure transverse position and TMIT (transmitted intensity) measures the passing beam charge, for a total of 450 signals. Since each pulse should be i.i.d. under normal operation, we currently consider each pulse to be a separate input; using windows of consecutive pulses is left for future work.

Since the operators of LCLS are constantly changing the machine settings, we limit our study to periods of “steady” operation at 120 Hz (as detailed in [26]) and train on a single day of data at a time. Nonetheless, due to the contamination in the training data, the standard VAE [16], RVAE [21], and DSVDD [15] methods, as well as ResVAE without logistic learning, fail to train; they experience disrupting gradients due to the anomalous examples. Gradient clipping does not fix the training failures, as this solely moderates the magnitudes of applied gradients instead of ignoring the gradient contribution of anomalous examples entirely. We train our ResVAE model using the PyTorch framework [27] and the Adam optimizer [28] using mini-batches. (Full details on the training setting and architecture are found in the longer version of our paper [18].)

To explore the anomalies detected by our method, we extract all anomalous pulses (where \( \gamma(x) \leq 0.5 \)) and then perform clustering on the vector \( \pi(x) \) (which provides feature-level attribution) using UMAP [29]. We show the existence of several anomaly clusters in Figure 2. We also dive into several clusters to determine the nature of the anomaly. Figure 3 shows an anomaly where the X-ray beam is partially lost, observed in the beam’s TMIT signals. Figure 4 displays the onset of an anomaly observed in the beam’s X signals. ResVAE is capable of providing operators with information regarding both the timing and location of the anomaly.

Figure 2. UMAP [29] clustering of the feature-level anomaly attribution vector \( \pi(x) \) for anomalous pulses (where \( \gamma(x) \leq 0.5 \)).

4 Related Work

There are numerous prior works studying the problem of unsupervised anomaly detection. Classical methods [5–8], typically rely on density or distance and therefore struggle with high-dimensional inputs, especially when anomalies are present in the training set. The main alternative methods rely on compression techniques, such as autoencoders [9, 10], VAEs [11],
and generative adversarial nets (GANs) [12, 13], or combine deep compression with classical methods [14, 15]. Nonetheless, these methods almost always assume a normal training set to avoid learning to compress anomalous examples.

Several methods attempt to add robustness to outliers to existing methods, including robust PCA [30, 31], $\beta$-robust VAE [32], and RVAE [21]. However, all of these approaches have various drawbacks. Although robust PCA (and its deep counterpart) can effectively identify both sample and feature anomalies in contaminated training data, its main drawbacks are the relative difficulty of interpreting the hyperparameter controlling the outlier fraction and the need to store the outlier matrix $S$ of size equal to the dataset. $\beta$-robust VAE [32] replaces the reconstruction likelihood term of the traditional ELBO with a more robust $\beta$-divergence term; however, this does not address robustness concerns in the latent space nor does this enable the model to identify feature outliers. Our approach is most similar to and extends the RVAE [21] work, following its general structure and derivation but making several key changes to add robustness to the latent space and ensure expressive weights.

5 Conclusion

This study introduces the ResVAE anomaly detection method, an innovative extension of the VAE model [16] designed to detect anomalies even when trained on contaminated data. We
provide a theoretical explanation for the limitations of VAE when faced with contaminated training data and propose a solution using a mixture model that assigns weights to both samples and features during training, effectively mitigating the impact of outliers. Subsequently, we apply this method to anomaly detection at LCLS, successfully identifying anomalous accelerator status and illustrating various types of anomalies.

References


