Disentangled Conditional Variational Autoencoder for Unsupervised Anomaly Detection

1st Asif Ahmed Neloy
Department of Computing Studies and Information Systems
Douglas College
700 Royal Ave, New Westminster,
British Columbia, V3M 5Z5, Canada
neloya@douglascollege.ca

2nd Maxime Turgeon
Department of Statistics
University of Manitoba
50 Sifton Rd, Winnipeg, Manitoba, R3T 2N2, Canada
max.turgeon@umanitoba.ca

***Abstract*—Recently, generative models have shown promising performance in anomaly detection tasks. Specifically, auto-encoders learn representations of high-dimensional data, and their reconstruction ability can be used to assess whether a new instance is likely to be anomalous. However, the primary challenge of unsupervised anomaly detection (UAD) is learning the appropriate disentangled features and avoiding information loss while incorporating known sources of variation to improve reconstruction. In this paper, we propose a novel generative auto-encoder architecture by combining the frameworks of β-VAE, conditional variational auto-encoder (CVAE) and the principle of total correlation (TC). We show that our architecture improves the disentanglement of latent features, optimizes TC loss more efficiently, and improves the ability to detect anomalies unsupervised with respect to high-dimensional instances, such as in imaging datasets. Through both qualitative and quantitative experiments on several benchmark datasets, we demonstrate that our proposed method excels in terms of anomaly detection and capture of disentangled features. Our analysis underlines the importance of learning disentangled features for UAD tasks.**

***Index Terms*—Unsupervised anomaly detection (UAD), Autoencoder, Disentangled Features, Principle of total correlation (TC), Generative Models.**

I. Introduction

Unsupervised anomaly detection (UAD) has been an abundant ground for methodological research for several decades. Recently, generative models, such as Variational Autoencoders (VAEs) [1] and Generative Adversarial Networks (GANs) [2, 3], have shown exceptional performance in UAD tasks. By learning the distribution of normal data, generative models can naturally score new data as anomalous based on how well they can be reconstructed. For a recent review of deep learning for anomaly detection, see [4].

In a complex task like UAD, disentanglement as a meta-prior encourages latent factors to be captured by different independent variables in the low-dimensional representation. This phenomenon has been shown in recent work that has used representation learning as a backbone for developing new VAE architectures. Some of the methods proposed new objective functions [5, 6], efficient decomposition of the evidence lower bound (ELBO) [7], partitioning of the latent space by adding a regularization term to the mutual information function [8], introducing disentanglement metrics [9], and penalizing total correlation (TC) loss [10]. Penalized TC efficiently learns disentangled features and minimizes the dependence across the dimension of the latent space. However, it often leads to a loss of information, which leads to a lower quality of reconstruction. For example, methods such as β-VAE, Disentangling by Factorising (FactorVAE) [9], and Relevance FactorVAE (RFVAE) [11] encourage more factorized representations with the cost of either losing reconstruction quality or losing considerable information about the data and dropping in disentanglement performance. To draw clear boundaries between an anomalous sample and a normal sample, we must minimize information loss.

To address these limitations, we present Disentangled Conditional Variational Autoencoder (dCVAE). Our approach is based on multivariate mutual information theory. Our main contribution is a generative modeling architecture which learns disentangled representations of the data while minimizing the loss of information and thus maintaining good reconstruction capabilities. We achieve this by modeling known sources of variation in a similar fashion as Conditional VAE [12].

Our contributions are outlined as follows:

- **Efficient Encoding:** We propose a novel method that integrates a generative modeling framework combining Conditional Variational Autoencoders (CVAE) with the principle of Total Correlation Explanation (CorEx). This integration aims to enhance disentanglement capabilities specifically tailored for Unsupervised Anomaly Detection (UAD).
- **Controlled Generation:** Leveraging insights from β-VAE and CVAE architectures, we implement controlled generation techniques to minimize information loss in latent variables, thereby improving the fidelity of reconstructed data.
- **Trade-offs.** Empirical evaluations demonstrate that our Disentangled Conditional Variational Autoencoder (dCVAE) effectively manages the trade-offs between reconstruction loss and quality. By incorporating total correlation (TC) and conditional variables, our approach achieves robust performance in UAD scenarios.
- **Comparable Evaluation:** Instead of introducing novel evaluation metrics, our proposed method modifies the

objective function to leverage traditional metrics such as reconstruction loss, anomaly score, and ROC-AUC. Through extensive empirical studies on multiple benchmark datasets, we establish the method's comparative performance against existing baseline methods.

Our paper is structured as follows. We first briefly discuss related methods (Section II), draw connection between them, and present our proposed method dCVAE (Section III). In Section IV, we discuss our experimental design including competing methods, datasets, and model configuration. Finally, experimental results with an ablation study are presented in Section V, Section VI and Section VII concludes this paper.

II. Related Work

In this section, we discuss the autoencoder methods focusing on two types of architecture: extensions of VAE enforcing disentanglement, and architectures based on mutual information theory.

A. β-VAE

β-VAE and its extensions proposed by Higgins et al. [5] is an augmentation of the original VAE with learning constraints of β applied to the objective function of the VAE. As a result, β-VAE is capable of discovering the disentangled latent factors and generating more realistic samples while retaining the small distance between the actual and estimated distributions.

Recall the objective function of VAE proposed by Kingma and Welling [1]:

$$L_{\mathrm{V\Lambda E}}(\theta,\phi) = -\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x} \mid \mathbf{z}) + D_{\mathrm{KL}}\left(q_\phi(\mathbf{z} \mid \mathbf{x}) \| p_\theta(\mathbf{z})\right).$$

Here, $p_\theta(\mathbf{x} \mid \mathbf{z})$ is the probabilistic decoder, $q_\phi(\mathbf{z} \mid \mathbf{x})$ is the recognition model, KLD is denoted by $D_{\mathrm{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \| p_\theta(\mathbf{z} \mid \mathbf{x}))$ parameterized by the weights (θ) and bias (ϕ) of inference and generative models. As the incentive of β-VAE is to introduce the disentangling property, maximizing the probability of generating original data, and minimizing the distance between them, a constant δ is introduced in the objective VAE to formulate the approximate posterior distributions as below:

$$\begin{aligned} &\max_{\phi,\theta} \mathbb{E}_{\boldsymbol{x}\sim \mathrm{X}}\left[\mathbb{E}_{q_\phi(z|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x} \mid z)\right]\right] \\ &\text{such that } \ D_{\mathrm{KL}}\left(q_\phi(z \mid \boldsymbol{x}) || p(z)\right) < \delta. \end{aligned} \tag{1}$$

Rewriting the Equation in Lagrangian form and using the KKT conditions, Higgins et al. [5] derive the following objective function:

$$\mathcal{L}_{\beta VAE}(\theta,\phi) = \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x \mid z)\right] - \beta D_{\mathrm{KL}}\left(q_\phi(z \mid x) \| p(z)\right), \tag{2}$$

Here, β is the regularization coefficient that enforces the constraints to limit the capacity of the latent information z. When $\beta = 1$, we recover the original VAE. Increasing the value of $\beta > 1$ enforces the constraints to capture disentanglement. However, Hoffman et al. [13] argue that with an implicit prior, optimizing the regularized ELBO is equivalent to performing variational expectation maximization (EM).

B. FactorVAE

Disentangling by Factorising or FactorVAE is another modification of β-VAE proposed by Kim and Mnih [9]. FactorVAE emphasizes the trade-off between disentanglement and reconstruction quality. The authors primarily focused on the objective function of the VAE and β-VAE. The authors propose a new loss function to mitigate the loss of information that arise while penalizing both the mutual information and the KLD to enforce disentangled latent factors.

According to Hoffman and Johnson [14] and Makhzani and Frey [15], the objective function of β-VAE can be further extended into:

$$\mathbb{E}_{p_{\text{data}}(x)}[KL(q(z \mid x) \| p(z))] = I(x;z) + KL(q(z) \| p(z)), \tag{3}$$

Here, $I(x;z)$ is the mutual information between x and z under the joint distribution $p_{\text{data}}(x)q(z \mid x)$. FactorVAE learns the second term of $KL(q(z) \| p(z))$ and resolved the aforementioned issues by introducing total correlation penalty and density-ratio trick to approximate the distribution $\bar{q}(z)$ generated by d samples from $q(z)$. The loss function of the FactorVAE is as follows:

$$\begin{aligned} \mathbb{E}_{q\left(z|x^{(i)}\right)}\left[\log p\left(x^{(i)} \mid z\right)\right] - KL\left(q\left(z \mid x^{(i)}\right) \| p(z)\right) \\ - \gamma KL(q(z) \| q(z)) \end{aligned} \tag{4}$$

C. The principle of total Correlation Explanation (CorEx)

Gao et al. [10] introduced CorEx to mitigate the problem of learning disentangled and interpretable representations in a purely information-theoretic way. In general, for VAE, we assume a generative model where $\mathbf{x}$ is a function of a latent variable $\mathbf{z}$, and afterward maximize the log likelihood of $\mathbf{x}$. On the other hand, CorEx follows the reverse process where $\mathbf{z}$ is a stochastic function of $\mathbf{x}$ parameterized by θ, i.e., $p_\theta(\mathbf{z} \mid \mathbf{x})$, and seek to estimate the joint distribution $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{z} \mid \mathbf{x})p(\mathbf{x})$. The underlying true data distribution maximizes the following objective:

$$\mathcal{L}(\theta;\mathrm{x}) = \underbrace{TC(\mathrm{x}) - TC_\theta(\mathrm{x} \mid \mathrm{z})}_{\text{informativeness}} - \underbrace{TC_\theta(\mathbf{z})}_{\text{(dis)entanglement}} \tag{5}$$

Recall the definition of the total correlation (TC) in terms of entropy $H(\mathbf{x})$ [16]:

$$TC(\mathbf{x}) = \sum_{i=1}^{d} H(\mathbf{x}_i) - H(\mathbf{x}) = D_{KL}\left(p(\mathbf{x}) \| \prod_{i=1}^{d} p(\mathbf{x}_i)\right). \tag{6}$$

By non-negativity of TC, Equation 5 naturally forms variational lower bound $TC(\mathrm{x})$ to the CorEx objective, i.e., $TC(\mathrm{x}) \geq \mathcal{L}(\theta;\mathrm{x})$ for any θ. Equation 5 can be rewritten in terms of mutual information $I(\mathrm{x} : \mathrm{z}) = H(\mathrm{x}) - H(\mathrm{x} \mid \mathrm{z}) = H(\mathrm{z}) - H(\mathrm{z} \mid \mathrm{x})$. Further constraining the search space

$p_\theta(\mathbf{z} \mid \mathbf{x})$ to have the factorized form $p_\theta(\mathbf{z} \mid \mathbf{x}) = \prod_{i=1}^{m} p_\theta(\mathbf{z}_i \mid \mathbf{x})$ and the mutual information terms can be bounded by approximating the conditional distributions $p_\theta(\mathbf{x}_i \mid \mathbf{z})$ and $p_\theta(\mathbf{z}_i \mid \mathbf{x})$. Finally, we can further rewrite and derive the lower bound for the objective function:

$$\begin{aligned}\mathcal{L}(\theta; \mathrm{x}) &= \sum_{i=1}^{d} I_\theta\left(\mathrm{x}_i : \mathrm{z}\right) - \sum_{i=1}^{m} I_\theta\left(\mathrm{z}_i : \mathrm{x}\right) \\ &\geq \left(\sum_{i=1}^{d} H\left(\mathrm{x}_i\right)\right) + E_{p_\theta(\mathrm{x},\mathrm{z})}\left(\log \underbrace{q_\phi(\mathrm{x} \mid \mathrm{z})}_{\text{decoder}}\right) \\ &\quad - D_{KL}(\underbrace{p_\theta(\mathrm{z} \mid \mathrm{x})}_{\text{encoder}} \| r_\alpha(\mathrm{z})).\end{aligned} \tag{7}$$

D. Total Correlation Variational Autoencoder (β-TCVAE)

Chen et al. [7] proposed disentanglement in their learned representations by adjusting the functional structure of the ELBO objective. The authors argued that each dimension of a disentangled representation should be able to represent a different factor of variation in the data and be changed independently of the other dimensions. β-TCVAE modifies the originally proposed ELBO objective by Higgins et al. [5] forcing the algorithm to learn representations without explicitly making restrictions or reduction to the latent space. To introduce TC and disentanglement into the original β-VAE, Chen et al. [7] decomposed the original KLD into **Index-Code MI, Total Correlation** and **Dimension-wise KL** terms. Furthermore, in the ELBO TC-Decomposition, each training samples are identified with a unique index $\mathbf{n}$ and a uniform random variable that refers to the aggregated posterior as $q(z) = \sum_{n=1}^{N} q(z \mid n)p(n)$ and can be denoted as:

$$\begin{aligned}\mathbb{E}_{p(n)}[\mathrm{KL}(q(z \mid n)\|p(z))] &= \mathrm{KL}(q(z,n)\|q(z)p(n)) + \\ &\mathrm{KL}\left(q(z)\|\prod_i q\left(z_i\right)\right) + \sum_i \mathrm{KL}\left(q\left(z_i\right)\|p\left(z_i\right)\right)\end{aligned} \tag{8}$$

Finally, with a set of latent variables z_i, with known factors v_k, the authors introduced a disentanglement measuring metric called mutual information gap (MIG) and defined in terms of empirical mutual information $I_n\left(z_i; v_k\right)$:

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{H\left(v_k\right)}\left(I_n\left(z_{i(k)}; v_k\right) - \max_{i \neq i(k)} I_n\left(z_i; v_k\right)\right) \tag{9}$$

Here, $i^{(k)} = \mathrm{argmax}_j\, I_n\left(z_i; v_k\right)$ and K is the number of known factors under v_k.

III. Disentangled Conditional Variational Autoencoder (dCVAE)

Our approach builds on CorEx and models known sources of variation in the data, in a manner similar to Conditional Variational Autoencoder (CVAE) [12]. In what follows, we will represent this known source of variation using the variable C. In the experiment below, C is discrete and represents the class of each image. Modifying Equation 5 to incorporate C, we get

$$\mathcal{L}(\theta; x, c) = TC_\theta(x \mid c) - TC_\theta(x \mid z, c) - TC_\theta(z \mid c). \tag{10}$$

The first two terms measure the amount of correlation explained by z, and by maximizing it, we maximize the informativeness of the latent representation. The third term measures the correlation between the components of z, and by minimizing it, we maximize the disentanglement between the latent dimensions. Using Mutual Information Theory [16], we can define the conditional differential entropy of $H(x)$ given c and interpret mutual information as a reduction in uncertainty after conditioning:

$$\begin{aligned}I(x; z \mid c) &= H(x \mid c) + H(z \mid c) - H(x, z \mid c) \\ I(x; z \mid c) &= H(x \mid c) - H(x \mid z, c) = H(z \mid c) - H(z \mid x, c).\end{aligned} \tag{11}$$

We can now rewrite Equation 10 using derived mutual information theory from Equation 11:

$$\mathcal{L}(\theta; x, c) = \sum_{i=1}^{n} I\left(x_i; z \mid c\right) - \sum_{i=1}^{m} I\left(z_i : x \mid c\right). \tag{12}$$

Now, consider the KLD between $p_\theta(\mathbf{x} \mid \mathbf{z}, c)$ and an approximating distribution $q_\phi(\mathbf{x} \mid \mathbf{z}, c)$. In terms of expectations with respect to the joint distribution $p_\theta(\mathbf{x}, \mathbf{z} \mid c)$, we can write:

$$-H(x \mid z, c) = E(\log p_\theta(x \mid z, c)) \geq E(\log q_\phi(x \mid z, c)). \tag{13}$$

Combing Equation 12 and 13 and assuming an approximating distribution $r_\alpha(z_i \mid c)$, parameterized by variational parameters α and ϕ, for $p_\theta(z_i \mid c)$, we obtain two inequalities:

$$\begin{aligned}I(x_i : z \mid c) &= H(x_i \mid c) - H(x_i \mid z, c) \\ &\geq H(x_i \mid c) + E(\log q_\phi(x \mid z, c)), \qquad (14) \\ I(z_i : x \mid c) &= D_{KL}\left(p_\theta(z_i \mid x, c)\|r_\alpha(z_i \mid c)\right). \qquad (15)\end{aligned}$$

Combining these bounds with β, we finally derive a lower bound for the objective function for dCVAE:

$$\begin{aligned}\mathcal{L}(\theta; x, c) \geq &\sum_{i=1}^{n} H(x_i \mid c) + E(\log q_\phi(x \mid z, c)) \\ &- \sum_{i=1}^{m} \beta D_{KL}\left(p_{(z_i|x,c)}\|r\left(z_i \mid c\right)\right).\end{aligned} \tag{16}$$

Equation 16 illustrates the lower bound objective function of dCVAE where $q_\phi(\mathbf{x} \mid \mathbf{z}, c)$ is the generative model or decoder and $p_\theta\left(\mathbf{z}_i \mid \mathbf{x}, c\right)$ is the recognition model or encoder.

IV. Experiments

In the experiments below, we compare our dCVAE method to five baseline methods: VAE, CVAE, β-VAE, Factor-VAE, and RFVAE. The first two methods were selected as well-known baselines that do not explicitly enforce disentanglement; on the other hand, the latter three methods seek to achieve a disentangled representation of the data.

A. Datasets

We evaluate dCVAE and other baseline models on the following four datasets. MNIST [17], Fashion-MNIST (FMNIST) [18] are considered the benchmark datasets, whereas the KMNIST [19] and EMNIST [20] datasets are used for testing accuracy on a real-world dataset to assess overall performance. A more detailed description of these datasets follows:

- **MNIST and FMNIST:** Firstly, we apply all models to two benchmark datasets, MNIST and Fashion-MNIST for training purpose. We used all 10 classes with 60000 and 10000 training and testing samples for both datasets with $28 \times 28 \times 1$ pixels channel.
- **KMNIST:** Secondly, we use another complex dataset, Kuzushiji-MNIST or KMNIST to test the models accuracy. KMNIST is a drop-in replacement for the MNIST dataset, a Japanese cursive writing style. KMNIST contains similar 10 classes with 60000 and 10000 training and testing samples with $28 \times 28 \times 1$ pixels channel.
- **EMNIST:** Finally, to further assess the performance, all models are employed on Extending MNIST or EMNIST dataset. EMNIST has extended 62 classes (digit 0-9, letters uppercase A-Z and lowercase a-z) with 700000 and 80000 training and testing samples with $28 \times 28 \times 1$ pixels channels. As a result, this dataset posses more challenges for the methods while conducting the downstream tasks. The EMNIST dataset was processed from NIST Special Database 19 [21] and contains handwritten digits and characters collected from over 500 writers.

Along with the benchmarked MNIST and FMNIST datasets, we performed extensive evaluations on KMNIST and EMNIST datasets emphasizing learning disentangled factors for UAD. In general, learning anomalous samples from the first two datasets is straightforward; however, EMNIST and KMNIST datasets contain small variations among normal and abnormal classes, sharp strokes, and lower distinguishable factors that result in a much more challenging UAD task.

B. Reconstruction error and Anomaly Score

Leveraging methods for the discriminator as the anomaly score and drawing separation between normal and anomalous data is challenging for the divergent architectures of autoencoders. Depending on the task the architecture is trained for, the discriminator varies greatly. In general, the UAD methods utilize reconstruction error [22], distribution-based error [23], and density-based error [24] scores to distinguish normal and anomalous data. Formally, for each input x, a test input $\widehat{x}_l$ is considered to be anomalous if reconstruction error or Anomaly score ($\mathcal{A}$) is greater than the minimum threshold value and denoted as follows:

$$\mathcal{A}(\hat{x}) = \|x - \mathrm{D}(\mathrm{G}(\hat{x}))\|_2. \tag{17}$$

C. Performance Metrics

One of the challenges of measuring the performance of disentanglement is to apply appropriate metrics based on the nature of the dataset, not of latent factors or dimensions in the latent space. Therefore, considering the different model architectures and datasets, we first measure the performance using Numerical AUC Score, reconstruction error ($\mathcal{A}$), and negative ELBO score ($\mathcal{E}$). These metrics provide a quantifiable method of accuracy, while also measuring the disentanglement among the latent factors.

We also measure performance qualitatively by visualizing the latent space and the 2D-manifold. Both allow us to visualize the orthogonality between latent features and demonstrate the accuracy of the models to handle reduced latent variables and the ability to reconstruct samples.

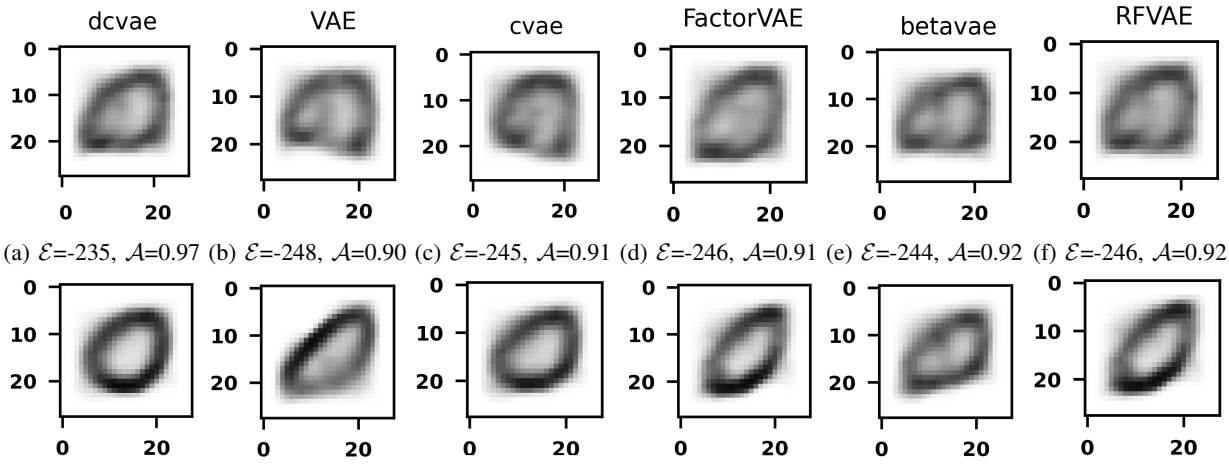
D. Model configuration

A fixed set of hyper-parameters are chosen to formulate a similar platform for all models and identify the computational cost and reproducibility of the models. Although baseline models that we chose, β-VAE, FactorVAE, RFVAE are highly sensitive to hyper-parameters tuning, the hyper-parameters throughout the experiment are kept consistent to observe how the models perform under similar values. A minimal 50 epochs are used to train the datasets. For MNIST, FMNIST, and KMNIST the batch size is kept to 64, with primary and secondary learning rates as $\alpha = 10^{-5}$ and $\alpha = 10^{-3}$ respectively. However, for the EMNIST dataset, the batch size increased to 128, and learning rates as $\alpha = 10^{-6}$ and $\alpha = 10^{-5}$.

V. Results and Discussion

In this section, we evaluate the results of dCVAE and other baseline methods on the downstream task of anomaly detection. A considerable volume of results was produced from our exhaustive evaluation. However, accounting for limitations of space here, we elected to focus on the results from EMNIST and KMNIST datasets in the main text. The remaining results (MNIST and FMNIST) are presented as Supplementary Material.

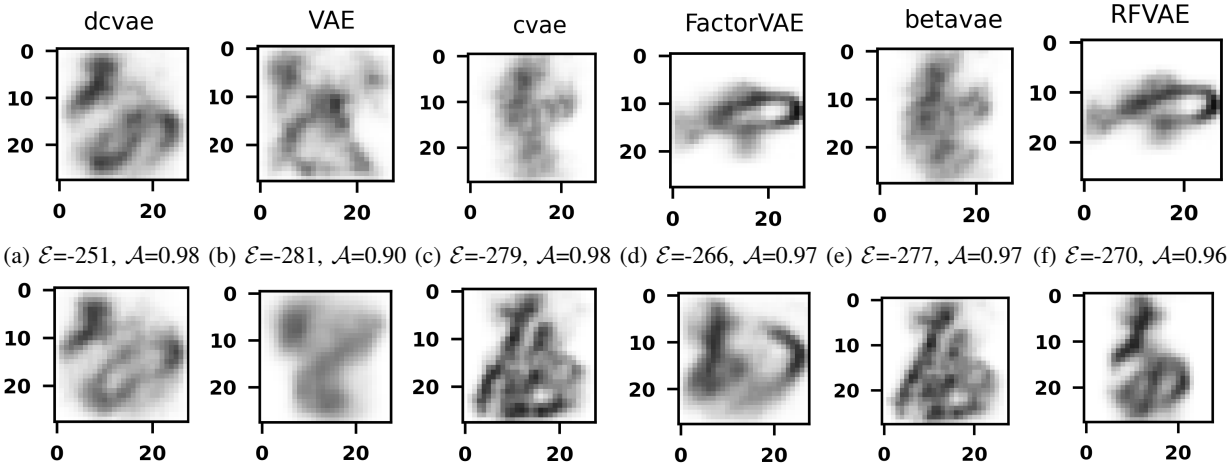
We show the results of our evaluation in three stages: firstly, using sample reconstruction and the negative ELBO score ($\mathcal{E}$) with reconstruction error $\mathcal{A}$, we evaluate and compare the disentanglement ability of dCVAE with baseline architectures. Secondly, we use the UMAP algorithm [25] to reduce dimensions and visualize both latent representation, as well as interpolation of the 2D-manifold to distinguish the TC by comparing information loss and effects of modeling known sources of variation. Finally, we present AUC scores and training time to summarize the overall accuracy of the



(a) $\mathcal{E}$=-235, $\mathcal{A}$=0.97 (b) $\mathcal{E}$=-248, $\mathcal{A}$=0.90 (c) $\mathcal{E}$=-245, $\mathcal{A}$=0.91 (d) $\mathcal{E}$=-246, $\mathcal{A}$=0.91 (e) $\mathcal{E}$=-244, $\mathcal{A}$=0.92 (f) $\mathcal{E}$=-246, $\mathcal{A}$=0.92

(g) $\mathcal{E}$=-171, $\mathcal{A}$=0.33 (h) $\mathcal{E}$=-199, $\mathcal{A}$=0.48 (i) $\mathcal{E}$=-180, $\mathcal{A}$=0.53 (j) $\mathcal{E}$=-185, $\mathcal{A}$=0.52 (k) $\mathcal{E}$=-195, $\mathcal{A}$=0.57 (l) $\mathcal{E}$=-190, $\mathcal{A}$=0.55

Fig. 1: Reconstruction for digit zero (0) and the capital letter O. Here, $\mathcal{E}$ refers to Negative ELBO score and $\mathcal{A}$ is the reconstruction error or anomaly score. Only dCVAE and FactorVAE show steady improvement for both types of reconstruction. All the other methods misclassify the samples. Moreover, we can observe higher reconstruction error and ELBO scores compared to MNIST and FMNIST.



(a) $\mathcal{E}$=-251, $\mathcal{A}$=0.98 (b) $\mathcal{E}$=-281, $\mathcal{A}$=0.90 (c) $\mathcal{E}$=-279, $\mathcal{A}$=0.98 (d) $\mathcal{E}$=-266, $\mathcal{A}$=0.97 (e) $\mathcal{E}$=-277, $\mathcal{A}$=0.97 (f) $\mathcal{E}$=-270, $\mathcal{A}$=0.96

(g) $\mathcal{E}$=-188, $\mathcal{A}$=0.28 (h) $\mathcal{E}$=-211, $\mathcal{A}$=0.47 (i) $\mathcal{E}$=-185, $\mathcal{A}$=0.41 (j) $\mathcal{E}$=-201, $\mathcal{A}$=0.57 (k) $\mathcal{E}$=-183, $\mathcal{A}$=0.41 (l) $\mathcal{E}$=-170, $\mathcal{A}$=0.38

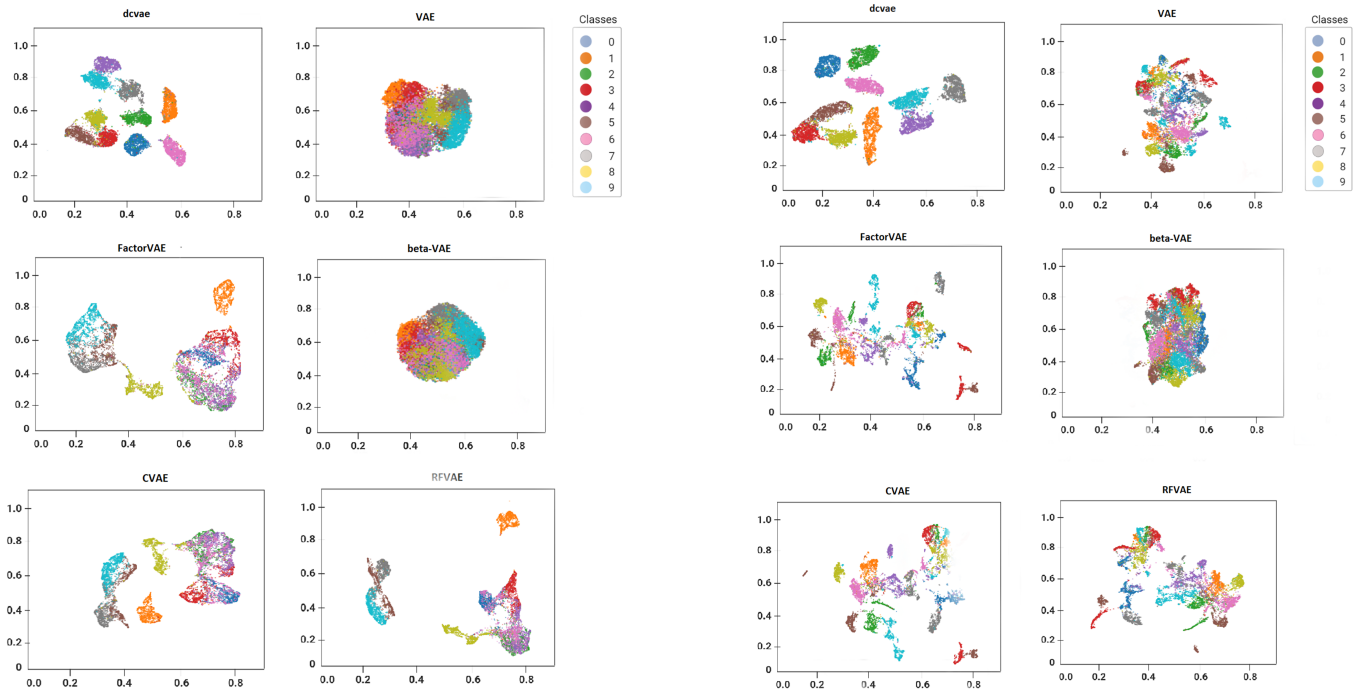
Fig. 2: In KMNIST dataset, without dCVAE, all other methods fail to classify both anomalous and normal samples. Reconstruction scores suggest FactorVAE, VAE almost fail to distinguish normal and anomalous observations. Since the strokes of the samples are similar in this dataset, methods that only emphasize disentanglement or empirical approximation lose more information in latent variables, resulting in false anomaly detection.

experimented methods. We evaluate the quality of disentanglement by considering explicit separation of $\mathcal{A}$ between normal and anomalous data and minimization of $\mathcal{E}$. A better disentanglement is achieved when:

(a) A higher reconstruction error $\mathcal{A}$ for anomalous sample and lower reconstruction error $\mathcal{A}$ for normal sample is obtained and
(b) $\mathcal{E}$ is minimized by enforcing regularization that either minimizes the negative ELBO decomposition $D_{KL}\left(p_{(z_i|x,c)\|r(z_i|c)}\right)$ or regularizes the approximate posterior $q_\phi(\mathbf{z} \mid \mathbf{x})$.

A clear boundary in terms of learning efficient disentanglement between the dCVAE and baseline methods can be observed from the reconstruction of both the reconstruction of EMNIST (Figure 1) and KMNIST (Figure 2) reconstruction. The first row corresponds to an anomalous reconstruction and the second row shows a normal sample reconstruction. Both $\mathcal{E}$ and $\mathcal{A}$ score suggest that dCVAE captures more independent factors and identifies anomalous and normal samples efficiently.

This observation strongly justifies one of our primary claims, namely that dCVAE incorporates the disentanglement



(a) EMNIST

(b) KMNIST

Fig. 3: Latent Representation of EMNIST and KMNIST

learning through enforcing TC and restricting independent latent variables to prioritize the minimization of the divergence. The other disentanglement methods presented here either only emphasize TC (indicated by the dependence between random variables) or introduce β (weighing the prior enforcement term), which limits the ability to learn randomness in a case where the hyperparameters are not tuned for certain dimensions. The second observation is drawn using latent representation (Figure 3) and 2D-manifold embeddings (Figure 4 and 5). Through this experiment, we observe the effect of modeling using a known source of variation (i.e. introducing conditional variable C into the objective function) and minimizing information loss through multivariate mutual information theory (i.e. decomposition of TC). We can observe clear similarities between KLD loss and modeling with known score of variance in a reduced latent space.

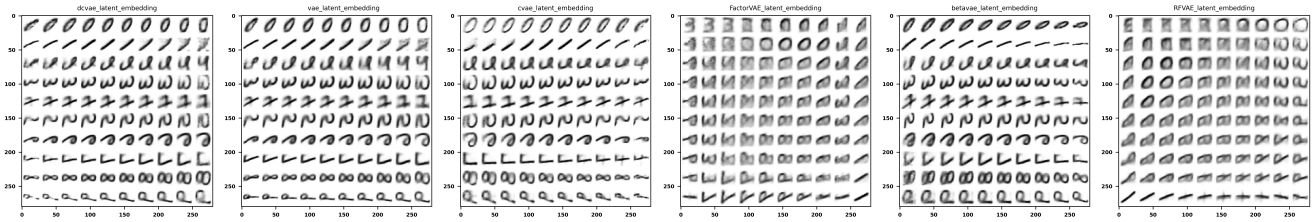
Due to enforced divergence loss, the plot of VAE and β-VAE are noticeably different from other architectures. The feature space is more compact for VAE, β-VAE, and we can see that the cluster of the different classes is not well separated. However, conditioning the generative function (encoder) of CVAE and dCVAE provides the leverage to construct a higher feature space and retain more accurate information in 2D-manifold (EMNIST, Figure 4; and KMNIST, Figure 5). Furthermore, TC reduces the correlation among disentanglement degrees when a specific feature is learned (shape, strokes, color, boundaries). Such classes can be observed to cluster together and the other is scattered with a larger feature space (Figure 3). Compared to other methods, it is evident that dCVAE maintains a consistent latent space and creates separate clusters more accurately. This indicates that more disentangled variables are captured and retain more information by conditioning the generative model by minimizing ELBO $D_{KL}\left(p_{(z_i|x,c)}\|r(z_i|c)\right)$.

Finally, Table I illustrates the results of the model evaluation through the AUC score and the training time. dCVAE outperforms other methods in terms of AUC score. However, for larger divergent datasets like KMNIST and EMNIST, VAE shows lower training time compared to dCVAE. Since VAE only optimizes the negative log-likelihood, reconstruction loss, and prior enforcement term, the training takes fewer latent variables to regularize, resulting in less training time. Nevertheless, compared to methods that incorporate TC (e.g. FactorVAE and RFVAE) or a constraint on the posterior (β-VAE), our proposed dCVAE scales to all larger datasets with higher classification accuracy.

The only trade-offs in our proposed method seem to occur when minimizing the negative ELBO loss. In certain conditions, dCVAE reaches a lower reconstruction loss (anomalous sample) yet minimizes the negative ELBO score (Figure 3, 4). In general, negative ELBO loss should illustrate symmetrical change with reconstruction error. Such inconsistency could lead to a significant drop in the classification accuracy, thus leading to a false anomaly detection result.

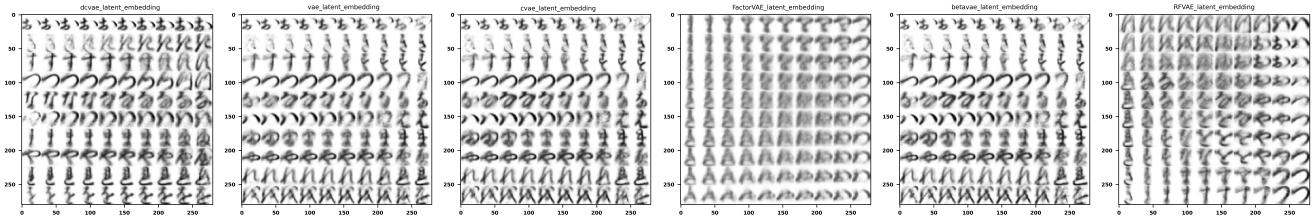
VI. Ablation Study

We performed the following ablation studies to evaluate the effect of tuning hyper-parameters on learning disentangled fac-



(a) dCVAE (b) VAE (c) CVAE (d) FactorVAE (e) β-VAE (f) RFVAE

Fig. 4: Manifold Embeddings (EMNIST)



(a) dCVAE (b) VAE (c) CVAE (d) FactorVAE (e) β-VAE (f) RFVAE

Fig. 5: Manifold Embeddings (KMNIST)

TABLE I: Evaluation metrics scores

Model	**MNIST**		**FMNIST**	
	AUC	**Training Time (min)**	**AUC**	**Training Time (min)**
dCVAE	**88.31**	37	**88.63**	44
VAE	88.21	37	84.12	39
CVAE	87.57	43	83.31	48
FactorVAE	87.11	53	82.78	50
β-VAE	85.31	51	82.31	53
RFVAE	85.31	**55**	81.11	**57**
Model	**EMNIST**		**KMNIST**	
	AUC	**Training Time (min)**	**AUC**	**Training Time (min)**
dCVAE	**78.98**	**102**	**61.02**	95
VAE	67.23	92	51.13	78
CVAE	66.01	117	**42.35**	104
FactorVAE	62.91	**138**	49.23	117
β-VAE	65.12	123	50.01	119
RFVAE	**55.03**	130	49.51	**132**

tors and minimizing information loss. Since both factors result in efficient UAD, we highlighted evaluation metrics (Table II) through the subsequent ablation studies:

1) **Unrestricted hyper-parameters:** We optimized the best hyperparameters (learning rate α, batch-size and β) for each model and tested them on KMNIST and EMNIST datasets. It is essential to find tuned parameters that result in efficient UAD. AUC score shows notable improvement for dCVAE and FactorVAE architecture.
2) **Random Generations:** We used the random generations to observe the reconstruction quality from the tuned hyper-parameters. dCVAE outperforms baseline methods on both benchmarked and test datasets.

VII. Conclusion

In this research, we present a novel generative variational model dCVAE, to improve the unsupervised anomaly detection task through disentanglement learning, TC loss, and minimizing trade-offs between reconstruction loss and reconstruction quality. Introducing a conditional variable to mitigate the loss of information effectively captures more disentangled features and produces more accurate reconstructions. Such architecture could be used in a wider range of applications, including generating controlled image synthesis, efficient molecular design and generation, source separation for bio-signals and images, and conditional text generation. Future research direction includes investigating the gap between the posterior and the prior distributions, resolving the trade-offs between loss function and reconstruction, and inspecting the dCVAE using different disentanglement metrics.

TABLE II: Results from Ablation Studies

Model	β	Learning Rate α	Batch Size	AUC	
				EMNIST	KMNIST
dCVAE	3	10^{-2}	128	81.23	70.25
VAE	1	10^{-5}	64	52.01	68.11
CVAE	1	10^{-5}	64	43.84	66.31
FactorVAE	2	10^{-6}	128	53.61	70.17
β-VAE	2.5	10^{-4}	64	52.71	66.51
RFVAE	2	10^{-5}	128	52.22	56.45

Acknowledgment

We would like to acknowledge support from the NSERC CREATE grant on the Visual and Automated Disease Analytics program. MT also acknowledges funding via a Discovery Grant from the Natural Sciences and Engineering Research

Council of Canada (NSERC), RGPIN-2021-04073. AAN acknowledges funding via a Research Dissemination Publish Grant from the Douglas College, S2024-RDG7.

REFERENCES

[1] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *2nd International Conference on Learning Representations, ICLR*, 2014.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[3] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*. PMLR, 2017, pp. 214–223.

[4] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–38, 2021.

[5] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *ICLR*, 2017.

[6] E. Mathieu, T. Rainforth, N. Siddharth, and Y. W. Teh, "Disentangling disentanglement in variational autoencoders," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4402–4412.

[7] R. T. Chen, X. Li, R. Grosse, and D. Duvenaud, "Isolating sources of disentanglement in VAEs," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 2615–2625.

[8] S. Zhao, J. Song, and S. Ermon, "Infovae: Information maximizing variational autoencoders," *arXiv preprint arXiv:1706.02262*, 2017.

[9] H. Kim and A. Mnih, "Disentangling by factorising," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2649–2658.

[10] S. Gao, R. Brekelmans, G. Ver Steeg, and A. Galstyan, "Auto-encoding total correlation explanation," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 1157–1166.

[11] M. Kim, Y. Wang, P. Sahu, and V. Pavlovic, "Relevance factor vae: Learning and identifying disentangled factors," *arXiv preprint arXiv:1902.01568*, 2019.

[12] A. A. Pol, V. Berger, C. Germain, G. Cerminara, and M. Pierini, "Anomaly detection with conditional variational autoencoders," in *2019 18th IEEE international conference on machine learning and applications (ICMLA)*. IEEE, 2019, pp. 1651–1657.

[13] M. D. Hoffman, C. Riquelme, and M. J. Johnson, "The β-vae's implicit prior," in *Workshop on Bayesian Deep Learning, NIPS*, 2017, pp. 1–5.

[14] M. D. Hoffman and M. J. Johnson, "Elbo surgery: yet another way to carve up the variational evidence lower bound," in *Workshop in Advances in Approximate Bayesian Inference, NIPS*, vol. 1, 2016.

[15] A. Makhzani and B. J. Frey, "Pixelgan autoencoders," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[16] M. Studenỳ and J. Vejnarová, "The multiinformation function as a tool for measuring stochastic dependence," in *Learning in graphical models*. Springer, 1998, pp. 261–297.

[17] L. Deng, "The MNIST database of handwritten digit images for machine learning research [best of the web]," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[18] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[19] T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, and D. Ha, "Deep learning for classical japanese literature," *arXiv preprint arXiv:1812.01718*, 2018.

[20] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik, "Emnist: an extension of mnist to handwritten letters (2017)," *arXiv preprint arXiv:1702.05373*, 2017.

[21] P. J. Grother, "Nist special database 19," *Handprinted forms and characters database, National Institute of Standards and Technology*, vol. 10, 1995.

[22] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab, "Deep autoencoding models for unsupervised anomaly segmentation in brain mr images," in *International MICCAI brainlesion workshop*. Springer, 2018, pp. 161–169.

[23] M. Goldstein and S. Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data," *PloS one*, vol. 11, no. 4, p. e0152173, 2016.

[24] B. R. Kiran, D. M. Thomas, and R. Parakkal, "An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos," *Journal of Imaging*, vol. 4, no. 2, p. 36, 2018.

[25] T. Sainburg, L. McInnes, and T. Q. Gentner, "Parametric UMAP Embeddings for Representation and Semisupervised Learning," *Neural Computation*, vol. 33, no. 11, pp. 2881–2907, 10 2021. [Online]. Available: https://doi.org/10.1162/neco_a_01434