



# MIT Open Access Articles

## *Generative Oversampling with a Contrastive Variational Autoencoder*

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

<b>Citation</b>	Dai, Wangzhi, Ng, Kenney, Sevenson, Kristen, Huang, Wei, Anderson, Fred et al. 2019. "Generative Oversampling with a Contrastive Variational Autoencoder." Proceedings - IEEE International Conference on Data Mining, ICDM, 2019-November.
<b>As Published</b>	10.1109/ICDM.2019.00020
<b>Publisher</b>	Institute of Electrical and Electronics Engineers (IEEE)
<b>Version</b>	Author's final manuscript
<b>Citable link</b>	<a href="https://hdl.handle.net/1721.1/137374">https://hdl.handle.net/1721.1/137374</a>
<b>Terms of Use</b>	Creative Commons Attribution-Noncommercial-Share Alike
<b>Detailed Terms</b>	<a href="http://creativecommons.org/licenses/by-nc-sa/4.0/">http://creativecommons.org/licenses/by-nc-sa/4.0/</a>

# Generative Oversampling with a Contrastive Variational Autoencoder

Wangzhi Dai\*, Kenney Ng<sup>†</sup>, Kristen A. Severson<sup>†</sup>, Wei Huang<sup>‡</sup>, Fred Anderson<sup>‡</sup> & Collin M. Stultz<sup>§</sup>

\* Department of EECS, MIT  
Cambridge, Massachusetts, USA  
Email: wzhdai@mit.edu

<sup>†</sup> Center for Computational Health and MIT-IBM Watson AI Lab  
Cambridge, Massachusetts, USA  
Email: Kenney.Ng@us.ibm.com, Kristen.Severson@ibm.com

<sup>‡</sup> University of Massachusetts Medical School  
Worcester, Massachusetts, USA

Email: {Wei.Huang, Fred.Anderson}@umassmed.edu

<sup>§</sup> Department of EECS & Institute of Medical Engineering and Science, MIT  
Division of Cardiology, Massachusetts General Hospital  
Cambridge, Massachusetts, USA  
Email: cmstultz@mit.edu

**Abstract**—Although oversampling methods are widely used to deal with class imbalance problems, most only utilize observed samples in the minority class and ignore the rich information available in the majority class. In this work, we use an oversampling method that leverages information in both the majority and minority classes to mitigate the class imbalance problem. Experimental results on two clinical datasets with highly imbalanced outcomes demonstrate that prediction models can be significantly improved using data obtained from this oversampling method when the number of minority class samples is very small.

**Keywords**—class imbalance, oversampling, generative model, contrastive learning

## I. INTRODUCTION

For a given classification problem, the term class imbalance refers to the scenario when the different class distributions are highly imbalanced. It is encountered in many real life situations including fraud detection [1] and disease prediction [2]. Applications of standard classification algorithms, which assume a balanced distribution, to imbalanced classification problems can lead to a reduction in performance [3]. As an example, clinicians may use classification algorithms to identify patients who are at high risk of death using clinical data available on admission [4]. Due to the low prior probability of death in the overall patient population, models trained retro-respectively can be highly biased towards the negative patients, making the model behave poorly on the positive patients. Given the class imbalance, the overall accuracy of such approaches may still be high, while the model’s ability to distinguish between positive and negative patients may be poor.

Approaches designed to cope with class imbalance can be roughly grouped into 2 categories, reweighting and resampling

[5] - [6]. Reweighting involves modifying the cost function to more heavily weight misclassifying samples in the minority class. As this involves customizing a cost function for each learning task, it could be hard to use them for other downstream applications. On the other hand, re-sampling methods either downsample the majority class or oversample the minority class, yielding a balanced dataset for training and testing. The Synthetic Minority Oversampling Technique (SMOTE) is one of the most used resampling methods [7]. Instead of simple oversampling with replacement, SMOTE creates synthetic new samples for the minority class by randomly interpolating between existing minority samples and their neighbors.

Despite the wide use of SMOTE, several drawbacks still exists. New samples created by interpolation always lie in the convex hull formed by the existing minority samples. This makes it hard to match the underlying distribution of the minority class, especially when the distribution contains a long tail.

Another common deficiency of SMOTE and other oversampling methods is that only the existing minority samples are used to fit and create new samples and the more abundant majority class, which contains vast information, is totally ignored. Though the minority and majority classes are distinct in many applications, they may also share a lot of common information. Suppose, for example, we are interested in classifying patients, who all share a common diagnosis, into those who will have an adverse outcome (high-risk) and those who do not (low-risk). High risk patients may be different than low risk patients in some aspects, but since all patients share the same diagnosis, both the positive and negative classes share some information. Moreover, in some class imbalance problems, the fraction of the total number of patients in the minority class can be

extremely low, which makes learning from the minority class challenging and at times misleading, due to the possible bias associated with a small number of samples.

Sharma et al. recently proposed a synthetic oversampling method with the majority class (SWIM) [8]. SWIM generates minority class samples that have the same Mahalanobis distance from the majority class. The generated samples are around the neighbourhood of the observed minority data and they are generated in regions that have similar densities with respect to the majority class as the observed minority data. Though the distributional information of the majority class is leveraged, SWIM does not explicitly discriminate between the shared and private information in each class, thus making the generated samples less intuitive and less reliable in cases of extreme class imbalance.

Contrastive learning algorithms provide a way to learn relationships between two data sets that share some common information [9] - [10]. Contrastive variational autoencoders (C-VAE) leverages deep neural network structures to learn nonlinear latent variables for both the shared variation and the unique variation enriched in a target dataset [9] - [11]. We build our generative oversampling method on top of the C-VAE in this work, and exploit the shared information in the positive class to build an improved oversampling procedure for the minority class.

## II. METHOD

Fig. 1 (a) illustrates the generative model for both the minority class  $x^+$  and majority class  $x^-$ . For  $x^+$ , there exist two distinct latent variables, the latent variable that consist the shared variation  $s$  and the private latent variable that consist the unique variation of the minority class  $z$ . For the generation of  $x^-$ , the private latent variable does not exist. It is only generated from the shared latent variable  $s$ . The variables  $z$  and  $s$  are drawn from a univariate normal distribution,  $N(0, I)$  in our implementation, where  $I$  is the identity matrix. The observed minority and majority samples are drawn from the conditional distribution given the latent variables. In the variational autoencoder frame work, this conditional distribution is a function  $f_\theta$  parameterized by  $\theta$ , and is shared by the minority class and majority class, as can be seen in Fig. 1 (b) as a shared decoder. Because the private latent variable  $z$  is absent for majority samples, they are set to 0 in the conditional distribution:

$$x_i^+ \sim f_\theta(x|z_i, s_i) \quad (1)$$

$$x_j^- \sim f_\theta(x|0, s_j) \quad (2)$$

In order to approximate the posterior distribution of the latent variables, two encoders  $q_{\phi_s}$  and  $q_{\phi_z}$  are introduced for the shared latent variables and private latent variables respectively. Similar to the standard variational learning algorithm, a lower bound can be derived for the likelihood for both the observed minority samples and majority samples:

$$\begin{aligned} L(x_i^+) &\geq \mathbb{E}_{q_{\phi_s} q_{\phi_z}} [f_\theta(x_i^+|s, z)] - KL(q_{\phi_s}(s|x_i^+)||p(s)) \\ &\quad - KL(q_{\phi_z}(z|x_i^+)||p(z)) \end{aligned} \quad (3)$$

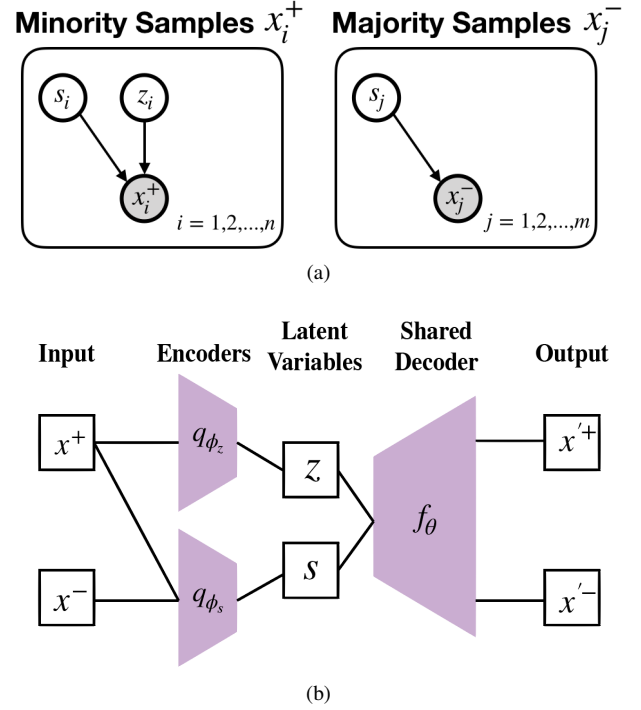


Fig. 1: (a) Generative model of minority samples and majority samples. The latent variable  $s$  is shared between the two classes while the private latent variable  $z$  only exists for the minority class. (b) Structure of C-VAE. Separate encoders  $q_{\phi_z}$  and  $q_{\phi_s}$  are used to approximate the posterior distribution of  $p(s|x^+, x^-)$  and  $p(z|x^+)$ . A shared decoder  $f_\theta$  represents the conditional distribution of  $p(x^+|z, s)$  and  $p(x^-|0, s)$ .

$$L(x_i^-) \geq \mathbb{E}_{q_{\phi_z}} [f_\theta(x_i^-|s, z)] - KL(q_{\phi_s}(s|x_i^-)||p(s)) \quad (4)$$

The encoders and decoders can be trained by maximizing the sum of the two lower bounds, using stochastic gradient descent from the observed minority and majority samples.

After training, new samples of the minority class can be generated by first drawing random samples for  $z$  and  $s$  from the prior distribution  $N(0, I)$ . Then the trained decoder is applied to map the latent variables to the observed space. The generated minority samples can be added to the original training set to make the two classes balanced, i.e., that the number of samples in the minority class + generated minority samples equals the number of samples in the majority class.

Fig. 2 shows the general workflow of our experiments. In addition to a C-VAE, we evaluated a number of other oversampling methods including a normal variational autoencoder (VAE), random oversampling (ROS), SMOTE and SWIM. Instead of using only the minority samples as is done in the traditional oversampling techniques, samples from both the majority class and the minority class are fed into the C-VAE and SWIM. The balanced data set with the generated minority samples can then be used to train new prediction models. For comparison, we call the prediction model without using the generated minority samples as base prediction model.

We implemented our oversampling method based on the

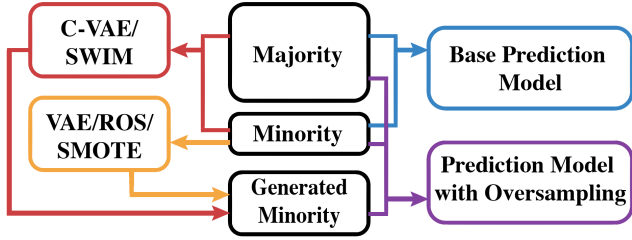


Fig. 2: Oversampling pipeline for C-VAE and other traditional techniques. Both majority and minority samples are fed into the C-VAE and SWIM, while only the minority samples are fed into the traditional methods. The number of samples in the minority class + generated minority samples equals the number of samples in the majority class. These three sets are used to train the prediction model with oversampling. The base prediction model is trained without the generated minority samples.

framework in [11] with Tensorflow. For ROS and SMOTE, we used the implementation in package *imblearn*. For SWIM, we used the code published by the authors of the paper [8].

### III. EXPERIMENTAL DESIGN

#### A. Dataset

We tested the generative oversampling method using the Global Registry of Acute Coronary Events (GRACE) [12] and another public Breast Cancer Dataset, which was obtained from the UCI Machine Learning Repository [13]. The GRACE registry was designed to track in-hospital and long-term outcomes of patients who presented with an acute coronary syndrome (ACS). GRACE enrolled over 70,000 patients from 1999-2009 from 250 hospitals in 30 countries. Patients enrolled in the GRACE registry experience a number of clinically relevant outcomes. To determine whether using generating synthetic data with a C-VAE improves classification performance, we focused on those outcomes that had the greatest class imbalance. Among them, we chose heparin-induced thrombocytopenia (HIT, a condition associated with decreased platelets), venous thromboembolic (VTE, a condition associated with blood clots in large veins) and stroke. We also included three more common in hospital events, myocardial infarction (MI), cardiogenic shock (CardShock) and recurrent ischemic symptoms (Ischemic) to investigate the effectiveness of the oversampling method in different situations.

On the other hand, the Breast Cancer dataset contains information on the rate of recurrent disease in breast cancer patients. It includes 286 patients with 201 of them had recurrence of breast cancer within five years of the initial tumor resection (positive outcome) and 85 of them did not (negative outcome). The patients are described by 9 prognostic features including age, menopausal status and other descriptive features for the tumor.

A summary of the two data sets and the chosen outcomes is shown in Table I.

TABLE I: Number and fraction of minority samples for the outcomes of interest in GRACE and Breast Cancer datasets.

Dataset	Outcome	# Minority	Fraction
GRACE	HIT	35	0.21%
	VTE	51	0.31%
	Stroke	85	0.51%
	MI	361	2.17%
	CardShock	528	3.17%
	Ischemic	3330	19.99%
Breast	Recurrence	85	29.7%

#### B. Prediction Tasks

For the ACS patients in the GRACE dataset, our task is to predict in hospital outcomes using all data available to the clinicians within the first 24 hours of the patients' admission. The extracted features include demographic information (e.g. age, gender), medical history, vital signs on admission (e.g. blood pressure), electrocardiographic (ECG) findings, lab tests (e.g. creatinine) and medications used (e.g. Aspirin). In sum, 198 features were used.

For the patients in the Breast Cancer dataset, we predict breast cancer recurrence by using all of the 9 features available.

We applied both a logistic regression model with L2 regularization and a simple neural network for the prediction tasks. The neural network is feed forward and contains 1 hidden layer with a Relu activation function and 1 output layer with a Sigmoid activation function. The implementation of the logistic regression model is based on scikit-learn and the neural network is based on Keras.

#### C. Oversampling Experiments

For each outcome, the data set is split into a training set and a test set, stratified according to the outcome of interest. Data for different outcomes were treated independently so that each prediction task had its own training and test set. Two base models for both logistic regression and neural network without any oversampling were trained using the training set. During the training of the base models, hyper-parameters such as learning rate, number of hidden units for the predictions models were selected by a 3-fold cross validation using the training sets. Then, both the majority and minority samples from the training set are used to train the C-VAE model. For comparison, we included a normal VAE and two baseline oversampling methods, random oversampling (ROS) and SMOTE, where only the minority samples are used for training, as shown in Fig. 2. We also included SWIM as an alternative oversampling method that uses some information from the majority class. Hyper-parameters including model architecture, learning rates, etc. for the C-VAE and the VAE, as well as the parameter that controls the spread of the synthetic samples in SWIM algorithm, were tuned by further splitting a validation set from the training data. A new logistic regression model and a neural network for each of the oversampling methods were then trained with the generated minority samples together with

TABLE II: AUC of logistic regression on different skewed clinical outcomes and comparison between different oversampling methods. Bold number indicates the corresponding model is significantly better than all other models ( $p < 0.05$ ). Standard errors of the AUCs are shown in parenthesis.

Dataset	Outcome	Base	ROS	SMOTE	SWIM	VAE	C-VAE
GRACE	HIT	0.66(0.06)	0.56(0.07)	0.57(0.07)	0.57(0.07)	0.80(0.06)	<b>0.87(0.04)</b>
	VTE	0.62(0.08)	0.56(0.10)	0.56(0.10)	0.60(0.08)	0.72(0.07)	<b>0.78(0.05)</b>
	Stroke	0.67(0.05)	0.64(0.06)	0.62(0.06)	0.66(0.07)	0.72(0.05)	<b>0.78(0.04)</b>
	MI	0.58(0.04)	0.57(0.03)	0.57(0.04)	0.58(0.04)	0.59(0.03)	0.63(0.04)
	CardShock	0.88(0.02)	0.87(0.02)	0.87(0.02)	0.87(0.02)	0.86(0.02)	0.87(0.02)
	Ischemic	0.62(0.01)	0.62(0.01)	0.62(0.01)	0.62(0.01)	0.59(0.01)	0.61(0.01)
Breast	Recurrence	0.69(0.05)	0.71(0.05)	0.70(0.05)	0.69(0.05)	0.71(0.04)	<b>0.72(0.04)</b>

TABLE III: PR-AUC of logistic regression on different skewed clinical outcomes and comparison between different oversampling methods. Bold number indicates the corresponding model is significantly better than all other models ( $p < 0.05$ ). Standard errors of the PR-AUCs are shown in parenthesis.

Dataset	Outcome	Base	ROS	SMOTE	SWIM	VAE	C-VAE
GRACE	HIT	0.004(0.003)	0.004(0.002)	0.004(0.002)	0.004(0.002)	0.011(0.006)	<b>0.030(0.020)</b>
	VTE	0.013(0.007)	0.014(0.013)	0.017(0.013)	0.011(0.007)	0.053(0.046)	<b>0.069(0.042)</b>
	Stroke	0.017(0.018)	0.010(0.005)	0.010(0.005)	0.014(0.019)	0.013(0.004)	0.032(0.019)
	MI	0.034(0.007)	0.036(0.012)	0.035(0.009)	0.032(0.006)	0.037(0.008)	0.049(0.013)
	CardShock	0.337(0.040)	0.329(0.044)	0.325(0.049)	0.332(0.045)	0.311(0.042)	0.344(0.043)
	Ischemic	0.295(0.013)	0.289(0.016)	0.290(0.011)	0.289(0.013)	0.265(0.012)	0.278(0.015)
Breast	Recurrence	0.477(0.074)	0.528(0.072)	0.525(0.073)	0.467(0.089)	0.533(0.058)	0.550(0.072)

the original training set.

#### D. Evaluation

All trained prediction models, including the base model, were then tested on the same held-out test set. We utilized the Area Under the receiver-operator Curve (AUC) and the Area Under the Precision-Recall Curve (PR-AUC) to evaluate the performance of the prediction models. AUC summarizes the trade-off between the true positive rate and false positive rate using different thresholds, and is widely used for evaluation of different prediction models [14]. However, the AUC may present an overly optimistic view of the results when there is a large class imbalance because of a very low false positive rate [15]. The PR-AUC on the other hand, is based on the fraction of true positives among positive predictions (precision), thus providing a better assessment of future classification performance for the positive class [16].

Here, we evaluated the prediction models with both of these two metrics. We did 10 bootstraps for every model and conducted a pair T test between different models to check statistical significance of their difference.

## IV. RESULTS

Table II summarizes the average AUC and standard error of the 10 bootstraps for the logistic regression model trained on different oversampling method, as well as the base model. Numbers in bold font indicate the corresponding model is significantly better than all other models ( $p < 0.05$ ). We find that the C-VAE method outperforms other methods when the number of the minority samples is small (e.g., for the outcomes HIT, VTE, Stroke, MI and breast cancer Recurrence). When the number of the minority samples is relatively large, as in

CardShock and Ischemic datasets, none of the applied oversampling methods enhanced the prediction model significantly compared to the base model.

The same experiment results are shown in Table III where the PR-AUC is used for evaluation. As the test data are highly skewed in that the prevalence of the outcome is very small, all PR-AUC values are small. Similar to the AUC results, the proposed C-VAE oversampling helped to improve the prediction models significantly when the number of the minority samples is low and on the other hand, no oversampling method is effective for the outcomes with a relatively large amount of minority class data.

Table IV and V show results of the same experiments using a neural network as the prediction model, evaluated by AUC and PR-AUC respectively. Similar to logistic regression, an obvious enhancement to the prediction performance can be seen when the minority samples are extremely well. In such cases, the neural network model without oversampling could not be appropriately trained due to lack of training data for the minority class. However, when there were enough data to train a neural network with discriminative power such as for the outcome CardShock, the oversampling methods did not improve the prediction performance and even made the results worse. This is potentially because the neural network is a much more complex model compared to logistic regression. When training with the enhanced dataset, it overfitted the over-sampled data, especially in light of the fact that the neural network models (in contrast to the logistic regression models) were constructed without regularization.

## V. DISCUSSIONS

Our results on the GRACE dataset suggest that the C-VAE oversampling method performs best when the number of

TABLE IV: AUC of feed-forward neural network on different skewed clinical outcomes and comparison between different oversampling methods. Bold number indicates the corresponding model is significantly better than all other models ( $p < 0.05$ ). Standard errors of the AUCs are shown in parenthesis.

Dataset	Outcome	Base	ROS	SMOTE	SWIM	VAE	C-VAE
GRACE	HIT	0.48(0.11)	0.56(0.06)	0.53(0.07)	0.59(0.14)	0.62(0.12)	<b>0.73(0.11)</b>
	VTE	0.45(0.08)	0.58(0.09)	0.61(0.07)	0.58(0.07)	0.61(0.06)	<b>0.68(0.07)</b>
	Stroke	0.55(0.07)	0.61(0.05)	0.64(0.04)	0.63(0.06)	0.63(0.05)	0.63(0.05)
	MI	0.60(0.03)	0.59(0.02)	0.58(0.02)	0.59(0.03)	0.54(0.03)	0.55(0.03)
	CardShock	<b>0.88(0.01)</b>	0.84(0.02)	0.84(0.02)	0.85(0.02)	0.81(0.04)	0.81(0.03)
	Ischemic	0.62(0.01)	0.62(0.01)	0.61(0.01)	0.61(0.01)	0.57(0.01)	0.57(0.01)
Breast	Recurrence	0.66(0.05)	0.64(0.06)	0.65(0.06)	0.66(0.05)	0.67(0.04)	<b>0.70(0.03)</b>

TABLE V: PR-AUC of feed-forward neural network on different skewed clinical outcomes and comparison between different oversampling methods. Bold number indicates the corresponding model is significantly better than all other models ( $p < 0.05$ ). Standard errors of the PR-AUCs are shown in parenthesis.

Dataset	Outcome	Base	ROS	SMOTE	SWIM	VAE	C-VAE
GRACE	HIT	0.004(0.004)	0.003(0.001)	0.004(0.003)	0.006(0.006)	0.009(0.012)	0.015(0.020)
	VTE	0.017(0.031)	0.013(0.019)	0.017(0.028)	0.017(0.024)	0.026(0.032)	0.038(0.046)
	Stroke	0.020(0.025)	0.013(0.012)	0.027(0.039)	0.021(0.030)	0.030(0.027)	0.031(0.027)
	MI	0.030(0.003)	0.30(0.003)	0.030(0.005)	0.032(0.008)	0.028(0.006)	0.029(0.003)
	CardShock	0.317(0.052)	0.274(0.042)	0.275(0.050)	0.298(0.054)	0.197(0.053)	0.021(0.043)
	Ischemic	0.288(0.012)	0.296(0.022)	0.278(0.009)	0.281(0.017)	0.0246(0.009)	0.245(0.010)
Breast	Recurrence	0.461(0.055)	0.459(0.084)	0.451(0.069)	0.459(0.047)	0.471(0.070)	<b>0.697(0.032)</b>

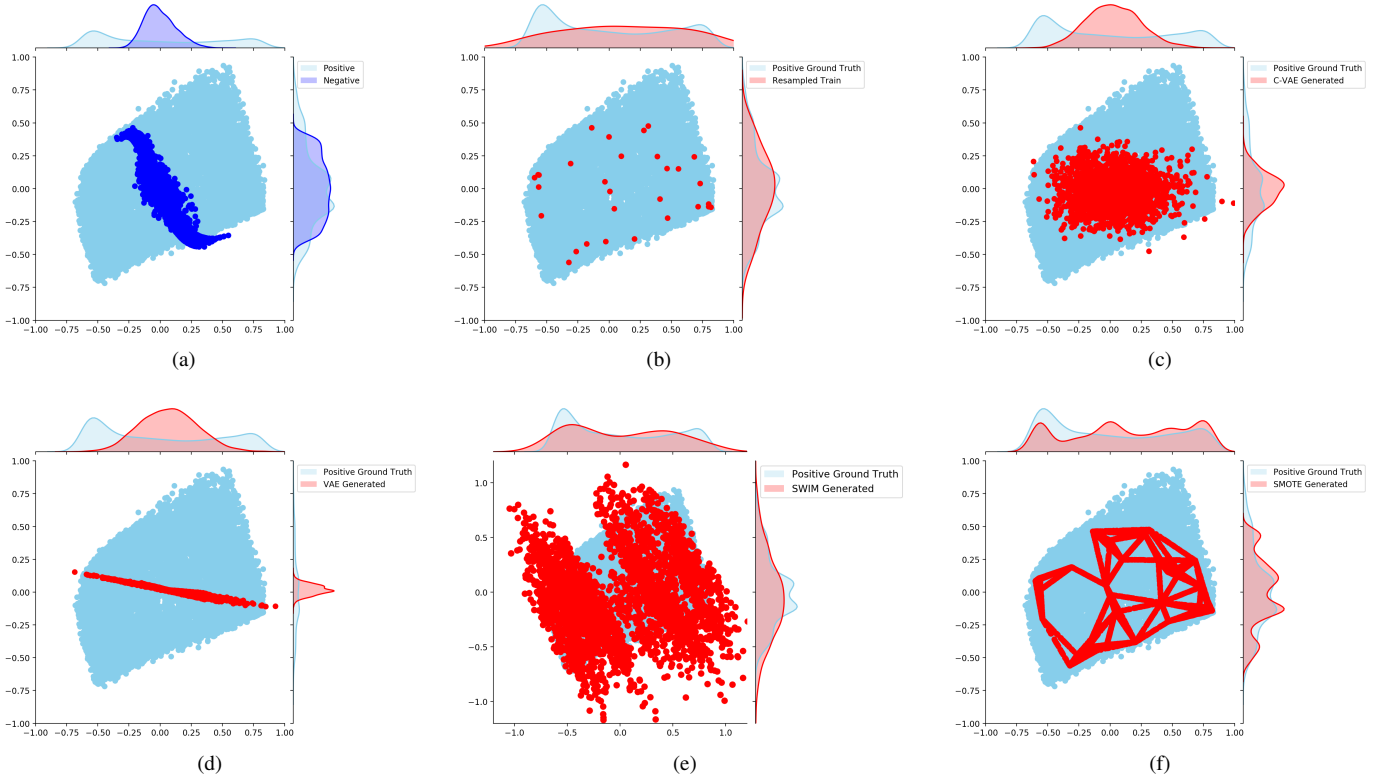


Fig. 3: Situation 1. Synthetic data experiments when there is shared information between positive and negative classes and private variation for the positive class exists. Shared latent variables exist for both classes but private latent variables only exist for the minority class. (a) PCA plots of synthetic data of both classes. (b) Re-sampled positive minority training data. (c)-(f) Generated positive minority samples from different oversampling methods compared with the ground truth positive data.



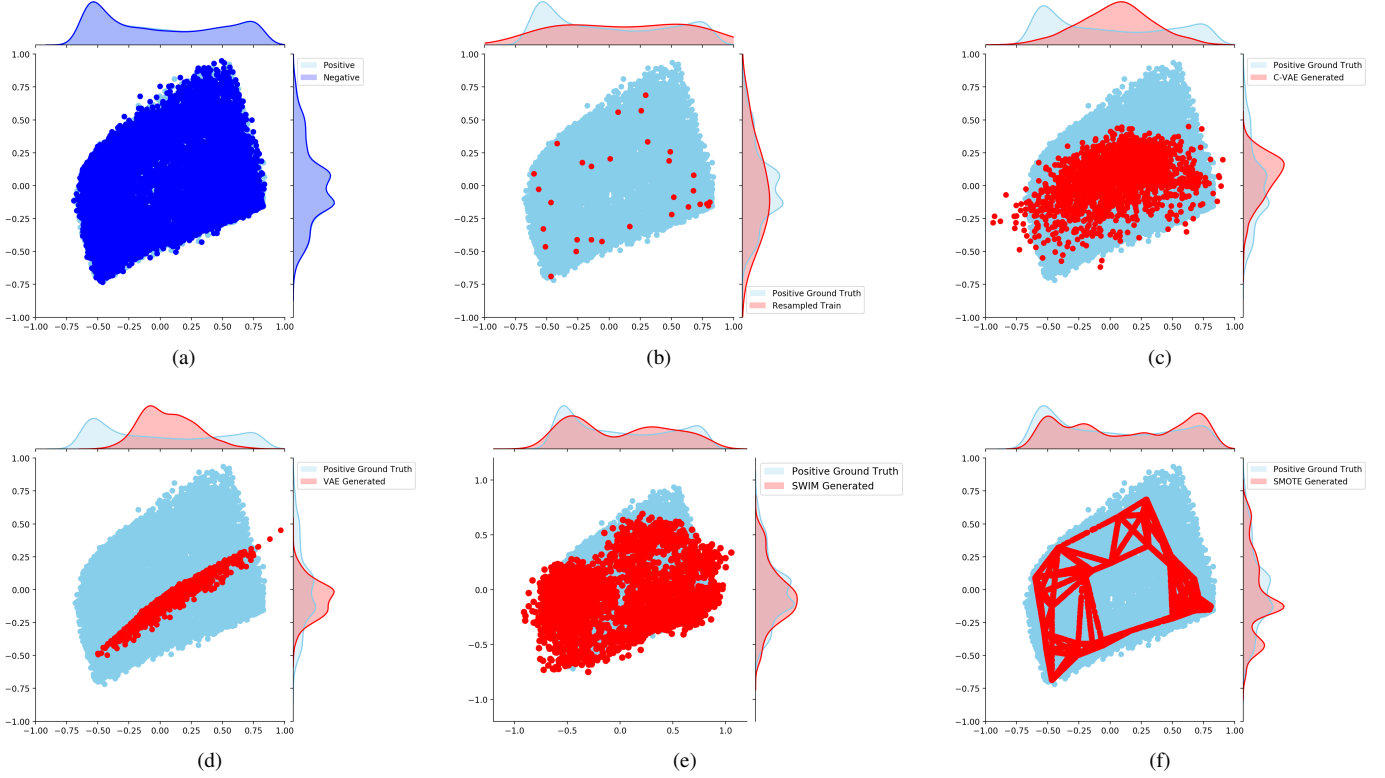


Fig. 4: Situation 2. Synthetic data experiments when positive and negative samples are drawn from the same distribution. Private latent variables are set to 0 and shared latent variables exist for both classes. (a) PCA plots of synthetic data of both classes. (b) Re-sampled positive minority training data. (c)-(f) Generated positive minority samples from different oversampling methods compared with the ground truth positive data.

samples in the minority class is very low. In order to further understand under what conditions the C-VAE would perform best, we conducted a series of experiments using synthetic data. We consider 3 different situations: 1) The positive and negative classes are drawn from two different distributions that share some common information (e.g., the multivariate distributions have similar variance in some dimensions); 2) The two classes samples are drawn from the same distribution; 3) The two classes are drawn from two different distributions that do not have any information in common. We construct the synthetic data using a latent variable model, similar to the generation process as described in Fig. 1. The shared information is represented by the variation of shared latent variables and the specific information of the positive class is captured by the variation of the private latent variables of the positive class only for situation 1. In situation 2, the private latent variables are set to 0 and only shared latent variables exist for both classes. In situation 3 the shared latent variables are set to 0 and both classes have its own private latent variables. Both the latent and private variables are drawn from a 2 dimensional Gaussian prior and we applied a 2 layer dense network with Sigmoid activation function to map them into a 16 dimensional data in the observed space. The weights of the network are pre-assigned with random number. Fig. 3

- Fig. 5 (a) show 2-dimensional PCA plots of the generated two class data in these 3 situations.

We consider the case where there is considerable class imbalance - a situation similar to many real world applications. We first generated 10,000 samples for each class, which represent the ground truth. We then re-sampled 30 data points from the positive class (0.3% positive), which corresponds to an observed imbalanced dataset that is available for model building, as shown by subfigure (b) in Fig. 3 - Fig. 5. We then trained different oversampling methods with those 30 training samples. Negative class data are also used for training C-VAE and SWIM. We generated positive samples using generative models trained on the synthetic data. The resulting data samples arising from the generative models are compared to the ground truth. Results are shown in (c) - (f) in Fig. 3 - Fig. 5.

The C-VAE clearly learned a better distribution when the positive and negative classes share some common information as shown in Fig. 3. The variation along the y-axis in this 2-dimensional representation was shared between positive and negative classes, making the negative samples helpful for the C-VAE to learn this variation. The VAE on the other hand, had only access to the small amount of positive training data as shown in Fig. 3 (b), failed to learn this shared variation

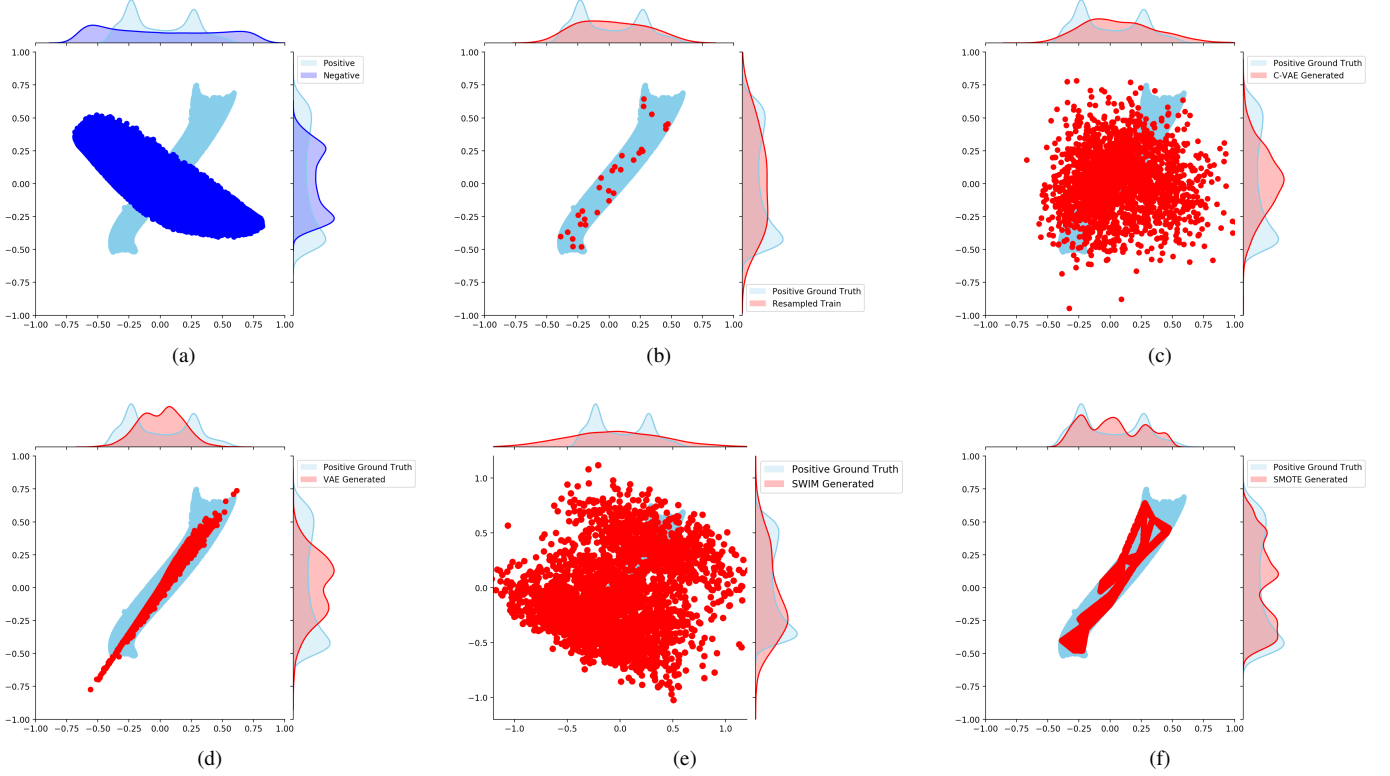


Fig. 5: Situation 3. Synthetic data experiments when there is no information shared between minority and majority classes. Shared latent variables are set to 0 and each class has its own private latent variables. (a) PCA plots of synthetic data of both classes. (b) Re-sampled minority training data. (c)-(f) Generated minority samples from different oversampling methods compared with the ground truth data.

but exaggerated the variation along x-axis in the training data. SWIM also learned the variation along both the x and y axis, but it generated a fair amount of outliers to the ground truth distribution.

When the positive and negative samples were drawn from the same distribution, both the C-VAE and SWIM learned a better distribution for the minority class, likely because it had more relevant data available to learn from. The VAE on the other hand, again, learned a distribution that exaggerates the variation along the x-axis, as shown in Fig. 4 (c) and (d).

By contrast, when the positive and negative samples were drawn from distinct distributions that do not share any common variance, the negative class does not provide useful information that the C-VAE can use, as can be seen in Fig. 5 (d). Outlier samples were generated due to the large variation along the diagonal in the negative training samples for C-VAE. SWIM, as can be seen in (e), was also misled by the negative class in generating samples with large variance along the wrong diagonal. In contrast, the normal VAE was not affected because it only had access to the positive training samples.

We also evaluated oversampling using SMOTE, as seen in part (f) of Fig. 3 - Fig. 5. In all 3 situations, the generated data from SMOTE matched the training samples well, but it did not

learn the distributional information of the positive samples, which makes it not generalizable when the training data are scarce.

From the above discussions, we conclude that C-VAE is useful when the following conditions are met: 1) shared information exist between the majority and minority classes and additional private variation exists for the minority class. 2) The minority class samples are scarce so that a normal generative model or oversampling model cannot be learned very well. 3) The distribution of the data are complex and high-dimensional, which makes simple oversampling methods fail.

Domain knowledge is necessary to judge whether shared information exists between classes. For example, patients diagnosed with the same disease can be a strong support for the common information and it is also reasonable to assume specific variation exists for patients with adverse outcomes in the cohort.

## VI. CONCLUSION

In this work, we proposed a generative oversampling method based on a contrastive variational autoencoder. Instead of using only the observed minority samples as done in most traditional oversampling methods, C-VAE oversampling



leverages shared information in both the majority and minority classes to build a better generative model for the minority class. We tested our method on a real life clinical data set where several outcomes (corresponding to the minority class) are highly skewed and extremely scarce. Results show that a logistic regression prediction model can be improved significantly when the original minority samples are rare. The C-VAE method out-performed the recent proposed SWIM method which also utilized the majority class.

Using a C-VAE is appropriate when there is information common to both the majority and minority class. Our synthetic data experiments argue that when the majority class has some variance in common with the minority class, the C-VAE can exploit this shared variance to better model the underlying distribution of the minority class. When the two classes arise from different distributions, which have no variance in common, then the C-VAE can yield misleading results. Prior domain specific information about the underlying distributions of the positive and negative classes can therefore help to when this oversampling method is most applicable.

## REFERENCES

- [1] Phua, Clifton, Daminda Alahakoon, and Vincent Lee. "Minority report in fraud detection: classification of skewed data." *Acm sigkdd explorations newsletter* 6.1 (2004): 50-59.
- [2] Guo, Xinjian, et al. "On the class imbalance problem." 2008 Fourth international conference on natural computation. Vol. 4. IEEE, 2008.
- [3] He, Haibo, and Edwardo A. Garcia. "Learning from imbalanced data." *IEEE Transactions on Knowledge & Data Engineering* 9 (2008): 1263-1284.
- [4] Tang, Eng Wei, Cheuk-Kit Wong, and Peter Herbison. "Global Registry of Acute Coronary Events (GRACE) hospital discharge risk score accurately predicts long-term mortality post acute coronary syndrome." *American heart journal* 153.1 (2007): 29-35.
- [5] Sun, Yanmin, Andrew KC Wong, and Mohamed S. Kamel. "Classification of imbalanced data: A review." *International Journal of Pattern Recognition and Artificial Intelligence* 23.04 (2009): 687-719.
- [6] Seiffert, Chris, et al. "Resampling or reweighting: A comparison of boosting implementations." 2008 20th IEEE International Conference on Tools with Artificial Intelligence. Vol. 1. IEEE, 2008.
- [7] Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research* 16 (2002): 321-357.
- [8] Sharma, Shiven, et al. "Synthetic oversampling with the majority class: A new perspective on handling extreme imbalance." 2018 IEEE International Conference on Data Mining (ICDM). IEEE, 2018.
- [9] Severson, Kristen A., Soumya Ghosh, and Kenney Ng. "Unsupervised learning with contrastive latent variable models." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019. APA
- [10] Abid, Abubakar, et al. "Exploring patterns enriched in a dataset with contrastive principal component analysis." *Nature communications* 9.1 (2018): 2134.
- [11] Abid, Abubakar, and James Zou. "Contrastive Variational Autoencoder Enhances Salient Features." *arXiv preprint arXiv:1902.04601* (2019).
- [12] Grace Investigators. "Rationale and design of the GRACE (Global Registry of Acute Coronary Events) Project: a multinational registry of patients hospitalized with acute coronary syndromes." *American heart journal* 141.2 (2001): 190-199.
- [13] Dua, D. and Graff, C. (2019). *UCI Machine Learning Repository* [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.
- [14] Hanley, James A., and Barbara J. McNeil. "The meaning and use of the area under a receiver operating characteristic (ROC) curve." *Radiology* 143.1 (1982): 29-36.
- [15] Davis, Jesse, and Mark Goadrich. "The relationship between Precision-Recall and ROC curves." *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006. APA
- [16] Saito, Takaya, and Marc Rehmsmeier. "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets." *PloS one* 10.3 (2015): e0118432.