Tackling Key Challenges to Guide Clinical Decisions in Cardiovascular Diseases

by

Wangzhi Dai

B.S., Peking University (2017)
S.M., Massachusetts Institute of Technology (2021)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2022

Author . . . .
Department of Electrical Engineering and Computer Science
June 30, 2022

Certified by . . . .
Collin M. Stultz
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by . . . .
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Tackling Key Challenges to Guide Clinical Decisions in Cardiovascular Diseases

by

Wangzhi Dai

Submitted to the Department of Electrical Engineering and Computer Science
on June 30, 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Machine learning models in healthcare have been widely studied in a number of contexts ranging from clinical risk stratification to image-guided diagnosis and prognostication. Nevertheless, key challenges remain from both clinical and technical perspectives. In the case of prediction models, for example, predicting the occurrence of rare clinical events is often challenging, mainly because of extreme class imbalance in the training data. Estimating treatment effect, on the other hand, is hindered by the fact that the common support assumption is not *a priori* guaranteed to be valid in non-randomized data. This thesis develops and applies approaches that address these challenges in order to obtain clinically useful insights.

In the first part of the thesis, we tackle these obstacles in the context of Acute Coronary Syndrome (ACS) - a condition where blood flow to the heart suddenly becomes compromised. We use a contrastive Variational Autoencoder (contrastive-VAE), an approach that models both the majority and minority classes as having shared latent properties, to address the following challenges: 1) Predicting rare adverse clinical outcomes after ACS; 2) Quantifying common support for estimating the effect of therapies for ACS; and 3) Causal feature selection for estimating individual treatment effects (ITE). For the first challenge, we demonstrate that generative oversampling with a contrastive-VAE significantly improves the discriminatory ability of predictive models relative to other traditional methods like SMOTE (Synthetic Minority Oversampling Technique). Similarly, for the problem of common support estimation, we show that a contrastive-VAE can effectively model the overlap between multiple treatment groups, yielding a quantitative estimate of the common support for the individual treatment effect and concomitant confidence intervals for the ITE estimate. Lastly, by modeling the joint distribution of patient features, treatments, and outcomes, we demonstrate that one can effectively identify a subset of patient features that are most important for ITE estimation, and that this smaller subset yields more precise ITEs with smaller confidence intervals.

In the second part of the thesis, we turn to a challenging clinical problem that uses ultrasound imaging for diagnosis and prognostication. Cardiac ultrasound (or echocardiography) plays a central role in the diagnosis and management of patients with suspected aortic stenosis (AS) - a disorder where one of the valves in the heart does not fully open. A complete echocardiographic study is typically performed by a trained sonographer who acquires videos of multiple views of the heart, and echocardiographers (cardiologists who specialize in the analysis of echocardiograms) interpret these videos, yielding clinically useful information. To facilitate the acquisition and interpretation of echocardiographic data, we developed a deep learning model that uses a single echocardiographic view (as opposed to use all of the acquired views) to diagnoses severe AS. We trained and evaluated the model based on spatial-temporal convolution that can accurately identify two key indicators of severe AS: large mean gradient over valve (0.88 AUC) and narrowed aortic valve area (0.78 AUC). Our approach might enable early detection of severe AS by non-specialists.

Thesis Supervisor: Collin M. Stultz
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

There are so many people who gave me enormous help during my journey in the past five years. But among them, I am most grateful to my advisor Professor Collin Stultz. Without his immerse knowledge and patient guidance, I cannot imagine myself surviving and finishing my study.

I also deeply appreciate the help from my committee, Professor John Guttag and Professor Marzyheh Ghassemi. Their insightful comments and encouragement, as well as the hard questions they raised during my defense, really helped me finish this thesis and widen my view of my research from various perspectives.

My sincere thanks also goes to my graduate mentor Professor Roger Mark who gave me precious advice every semester when we met to discuss my research and study plan.

I also want to mention collaborators at the Echo lab in Mass General Hospital, especially Dr. Judy Hung, cardiac fellow Hamed Nazzari, and Carl Anderson who helped me on the echocardiogram project. I'd also like to thank Dr. Kenny Ng and collaborators from IBM research.

I would also like to extend my deepest gratitude to all the lab mates at MIT CCRG research group, as well as our lab assistant Megumi for all her work to help us in the past few years.

Lastly, I'd like to thank all my friends, and most importantly my parents and my beloved Zhenzhuo Lan. I know they have always got my back no matter what happens in life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The past few years have witnessed the rapid growth of machine learning approaches in the healthcare domain. The curation of large datasets have fostered the development of data driven models for risk stratification [4] [5] and image-guided diagnosis [6] [7]. Nevertheless, deployment of those models in a real world setting is still relatively rare [8] and key challenges remain from both clinical and technical perspective.

Class imbalance is a common challenge in disease prediction models, since some adverse outcomes can be rare in the population. Traditional approaches like SMOTE [9] oversamples the minority group by generating synthetic data points. One common deficiency of SMOTE and other oversampling methods is that only the existing minority samples are used to fit and create new samples, the more abundant majority class is totally ignored. Though the minority and majority classes are distinct in many applications, they may also share a lot of common information. In this thesis, we use a Contrastive Variational Autoencoder (Contrastive-VAE) that leverages deep neural network structures to learn nonlinear latent variables for both the shared variation and the private variation enriched in a target dataset. We build our generative oversampling method on top of the Contrastive-VAE, and exploit the shared information in the majority class to build an improved oversampling procedure for the minority class.

Estimating the effect of a given medical treatment on individual patients from an observational dataset is possible only when the assumptions of ignorability and common support are valid [10]. Essentially, features having the greatest common support correspond to regions of significant overlap between the distributions of the different treatment groups [11]. In observational datasets, however, all possible treatment options are usually not uniformly represented, and therefore robust estimation of their effect may only be possible for the patients in the overlapping region. We use the Contrastive-VAE to estimate where there is significant overlap between patient distributions corresponding to different treatment options. By exploiting shared information between different treatment groups, a Contrastive-VAE can provide an improved estimation of the distribution of the groups with a small number of data points. By estimating the likelihood for each group with annealed importance sampling, we are able to quantitatively identify the area of overlap between multiple treatment groups and obtain an effective confidence interval for the estimated individual treatment effect.

For the purpose of satisfying the ignorability assumption, traditional wisdom in treatment effect estimation suggests including as many features as possible, as more features indicate a lower probability that a confounding factor is missing [10]. However, too many features in a dataset may not lead to better results in certain situations [12]. Large feature sets necessarily increase the dimensionality and complexity of models and, consequently training such models will require large datasets, and inference for downstream tasks may also be more compute intensive. In addition, a large number of features may make treatment effect estimation more difficult because the resulting feature may contain features that are irrelevant for the outcome of interest. To give an extreme example, a patient's date of admission typically appears in the electronic health record. Using this as a feature, however, would violate the common support assumption if no other patients in the dataset at hand were admitted that day. Therefore, removing irrelevant features in an observational dataset helps to ensure that the common support assumption is met and that robust estimates of the individual treatment effect are obtained. We use a Contrastive-VAE to model the distribution of patients data in different treatment groups. We then use these distributions to find a reduced subset that only contains features that are most relevant. In sum, the probability of the outcome conditioned on the full set of patient features is equal to the probability of the outcome conditioned only on the smaller subset containing only relevant features. We use annealed importance sampling to calculate the needed conditional probabilities using the learned distributions from the Contrastive-VAE.

We discuss the use case of the proposed approaches in two types of common, yet sometimes fatal cardiovascular diseases, Acute Coronary Syndrome (ACS) and Aortic Stenosis (AS). ACS is a condition related to decreased blood flow in the coronary arteries so that part of the heart muscle is unable to function properly or dies. Clinical risk scores like TIMI [13] and GRACE [14] have been successfully used to stratify the risks of adverse outcomes (death, recurrent myocardial infarction, etc) post ACS, but other rare events like venous thromboembolic (VTE, a condition associated with blood clots in large veins) are much harder to predict [15]. We also discuss the use of Contrastive-VAE in treatment estimation of two major treatments for ACS, Percutaneous Coronary Intervention (PCI) and Coronary Artery Bypass Grafting (CABG).

Unlike ACS, AS is a progressive disease. In AS, the aortic valve of patients narrows with time [16]. The diagnosis of Aortic Stenosis is typically made with echocardiography, an ultrasound study that requires a trained sonographer to obtain many (~ 100) videos of the heart, each one corresponding to a different view. A quick, accurate and automatic detection using just one echo view would provide easier and more frequent access to diagnoses for many patients. In the second part of the thesis, we build computational tools to pre-process and extract the Parasternal Long Axis (PLAX) view, one of the most common and easy-to-get part in an echo study. We then identify key indicators of severe aortic stenosis by training models that use only a single PLAX video.

1.1 Organization of the Thesis

The thesis is composed of 4 parts. The first 3 sections describe the use of contrastive-VAE, a generative model that leverages shared information between the majority and minority groups to build models that can predict rare events and that can estimate the individual treatment effect, all in the context of acute coronary syndromes. The 4th part of the thesis discusses processing and modeling Echocardiogram videos for the diagnosis of severe Aortic Stenosis.

1.1.1 Classifying Rare Clinical Outcomes with Generative Oversampling

Motivation

Class imbalance is a common yet serious problem for all data driven models, especially in the healthcare domain where rare clinical outcomes may lead to serious conditions. In this section, we aim to explicitly model the shared information between the majority and minority group with a Contrastive-VAE with a partitioned latent space [17]. By leveraging the common information, we build more accurate generative models and generate samples that can improve prediction models.

Main Contributions

- We develop a contrastive-VAE that leverages shared information in both the majority and minority classes to build a generative model for the minority class.

- We test our method on a clinical data set where several events (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 minority class is augmented with samples arising from a contrastive-VAE

1.1.2 Quantifying common support for individual treatment effect estimation

Motivation

Robust estimates of the treatment effect for a given patient with a pre-specified set of clinical characteristics, are possible to obtain when there is sufficient common support for these features [11]. In this section, we aim to use a Contrastive-VAE to model patients in different treatment groups and estimate where there is significant overlap between patient distributions corresponding to different treatment options and obtain an effective confidence interval for the estimated individual treatment effect.

Main Contributions

- Contrastive-VAE explicitly models the distributions of each treatment groups in a parametric way. The method allows us to model multiple treatment groups simultaneously, and effectively deals with data scarcity - a common problem in datasets where patients can receive multiple different treatments.

- By estimating the likelihood for each group with annealed importance sampling, we are able to quantitatively identify the area of overlap between multiple treatment groups and obtain an effective confidence interval for the estimated individual treatment effect.

1.1.3 Selecting features for individual treatment effect estimation

Motivation

To fulfill the ignorability assumption [10], the usual practice in estimating treatment effect with observational data is to include as many features as possible. A large number of irrelevant features lead to complex models and make the common support assumption difficult to meet [12]. The challenge, however, is reliably identifying, *a* *priori*, which features are irrelevant to the outcome of interest using observational data. In this section, we propose and develop a method for selecting an unbiased subset of clinical features for estimating the individual treatment effect.

Main Contributions

- We develop an algorithm to infer from a trained Contrastive-VAE for distributions of patients from various treatment groups. The algorithm finds a reduced subset that only contains features that are most relevant, where the probability of the outcome conditioned on the full set of patient features is approximately equal to the probability of the outcome conditioned only on the smaller subset containing the relevant features. We use annealed importance sampling to calculate the needed conditional probabilities using the learned distributions from the Contrastive-VAE.

- Results on both synthetic and actual clinical data demonstrate that the algorithm can successfully exclude irrelevant features and give an unbiased estimation of the treatment effect in the synthetic case.

1.1.4 Identifying Aortic Stenosis with a Single Echocardiogram Video

Motivation

Cardiac ultrasound plays a central role in the diagnosis and management of patients with suspected aortic stenosis (AS) [18]. However, a comprehensive assessment of the aortic valve requires a full echocardiographic study, which is performed by a skilled sonographer and interpreted by a clinician who has expertise in evaluating echocardiographic studies [19]. A quick and accurate detection method, which minimizes the need for specialized clinical interpretation, would make AS screening more accessible in settings where access to clinical specialists is limited. In this section, we aim to develop a model to identify severe AS patients with a single PLAX video for quick screening and detection of the disease.

Main Contributions

- We build a series of computational tools for echocardiogram pre-processing including de-identification, ECG extraction, PLAX view identification and data augmentation

- We build a classification model for two key indicators of severe AS: high mean gradient ($>$ 40 mmHG) and narrowed aortic valve area (< 1 cm^2). We achieve an AUC of 0.88 and 0.78 for the two tasks respectively. At the prevalence of severe AS in our cohort, the model can obtain 0.90 NPV for mean gradient and 0.87 NPV for valve area, while maintaining 0.9 sensitivity. Such method may enable quick screening and detection of AS and dramatically reduce the expenses and efforts of a full echo study.

Chapter 2

Generative Oversampling with a Contrastive VAE

2.1 Introduction

2.1.1 Class Imbalance

For a given classification problem, the term class imbalance refers to the scenario when the different class distributions are highly imbalanced, as shown in the diagram of Figure 2-1. It is encountered in many real life situations including fraud detection [20] and disease prediction [21]. Applications of standard classification algorithms, which assume a balanced distribution, to imbalanced classification problems can lead to a reduction in performance [22]. As an example, clinicians may use classification algorithms to identify patients who are at high risk of death using clinical data available on admission [14]. 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 [23] - [24]. Reweighting involves modifying the

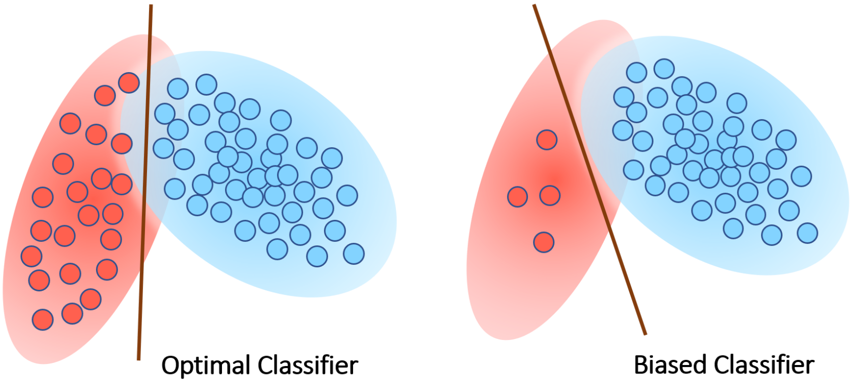


Figure 2-1: Biased classifier in case of class imbalance

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 [9]. 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. An illustration of downsampling, oversampling and SMOTE is shown in Figure 2-2.

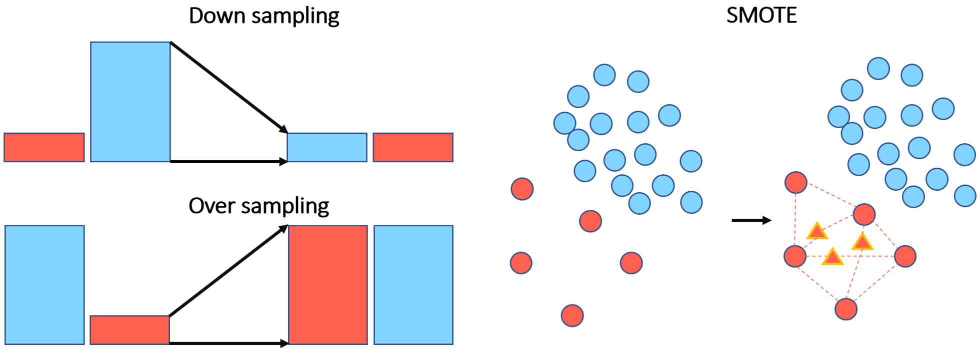


Figure 2-2: Downsampling, oversampling and SMOTE algorithm

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.

2.1.2 Oversampling with Shared Information between Minority and Majority Samples

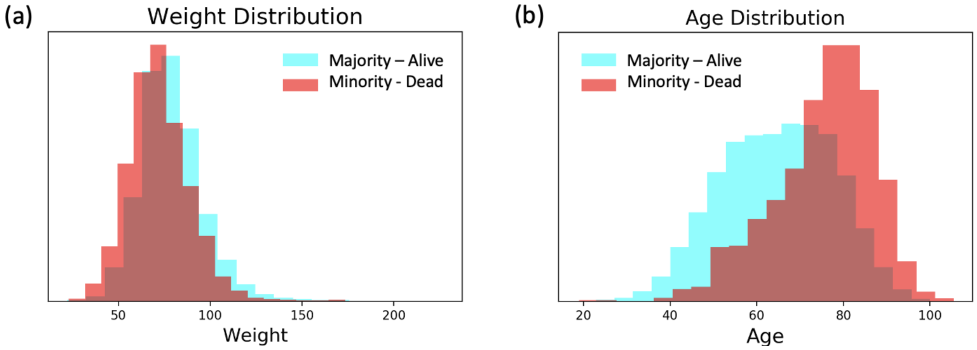


Figure 2-3: Shared and special characteristics of minority and majority samples

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. Figure 2-3 gives a concrete example of a real word clinical data where the majority group is patients who are alive and the minority group is patients who are dead. Both groups of patients are diagnosed with acute coronary syndrome. Figure 2-3 (a) shows the distribution of weight is roughly the same between the two groups, while the age distribution differs dramatically as shown in Figure 2-3 (b). Such example illustrates that shared informing exists between the two groups, despite their difference. 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

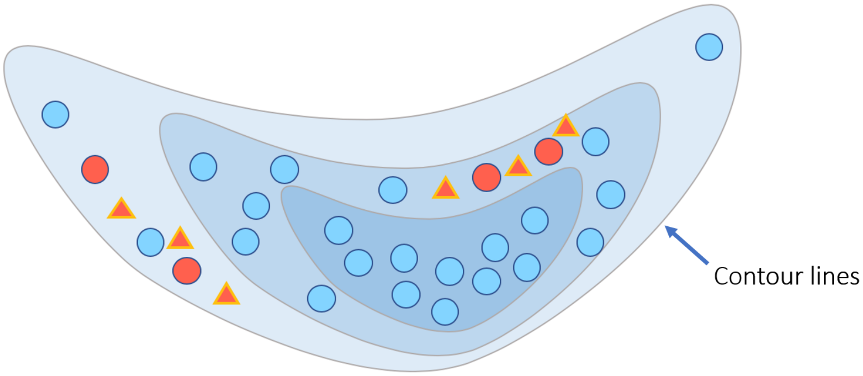


Figure 2-4: SWIM oversampling

and at times misleading, due to the possible bias associated with a small number of samples. Modeling the shared information and learning from the majority group can be helpful in such cases.

Sharma et al. proposed a synthetic oversampling method with the majority class (SWIM) [25]. 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, as shown in Figure 2-4. 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.

2.2 Methods

2.2.1 Contrastive VAE

Unsupervised learning with contrastive latent variables provide a way to learn relationships between two data sets that share some common information [17] - [26]. 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 [27]. 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.

Figure 2-5 (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 identify 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_θ parameterized by θ, and is shared by the minority class and majority class, as can be seen in Fig. 2-5 (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:

$$x_i^+ \sim f_\theta(x|z_i, s_i) \tag{2.1}$$

$$x_j^- \sim f_\theta(x|0, s_j) \tag{2.2}$$

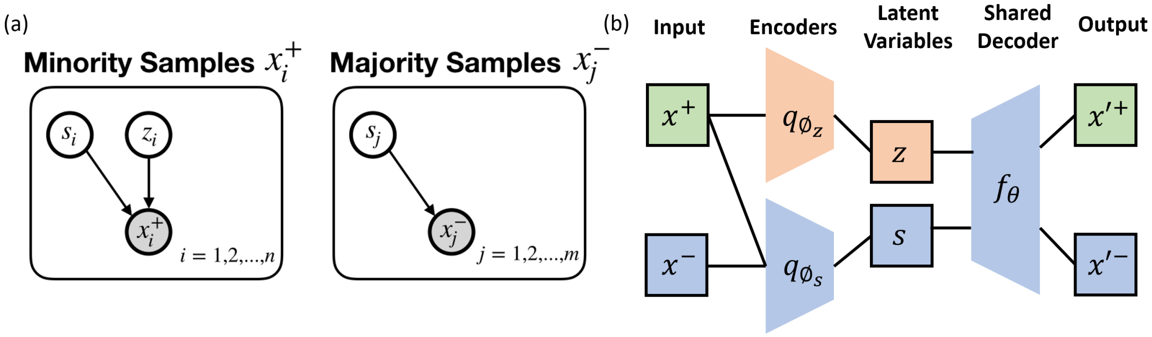


Figure 2-5: (a) Generative model of minority samples and majority samples. The latent variable s is shared between the two classes while the private latent variable z only exists for the minority class. (b) Structure of C-VAE. Separate encoders q_{ϕ_z} and q_{ϕ_s} are used to approximate the posterior distribution of $p(s|x^+, x^-)$ and $p(z|x^+)$. A shared decoder f_θ represents the conditional distribution of $p(x^+|z, s)$ and $p(x^-|0, s)$).

In order to approximate the posterior distribution of the latent variables, two encoders q_{ϕ_s} and q_{ϕ_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:

$$\begin{aligned} L(x_i^+) \geq E_{q_{\phi_s} q_{\phi_z}}[f_\theta(x_i^+|s,z)] - KL(q_{\phi_s(s|x_i^+)}||p(s)) \\ - KL(q_{\phi_s(z|x_i^+)}||p(z)) \end{aligned} \tag{2.3}$$

$$L(x_i^-) \geq E_{q_{\phi_z}}[f_\theta(x_i^-|s,z)] - KL(q_{\phi_s(s|x_i^-)}||p(s)) \tag{2.4}$$

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. 2-6 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 the base prediction model.

We implemented our oversampling method based on the framework in [27] with Tensorflow. For ROS and SMOTE, we used the implementation in package *imblearn*.

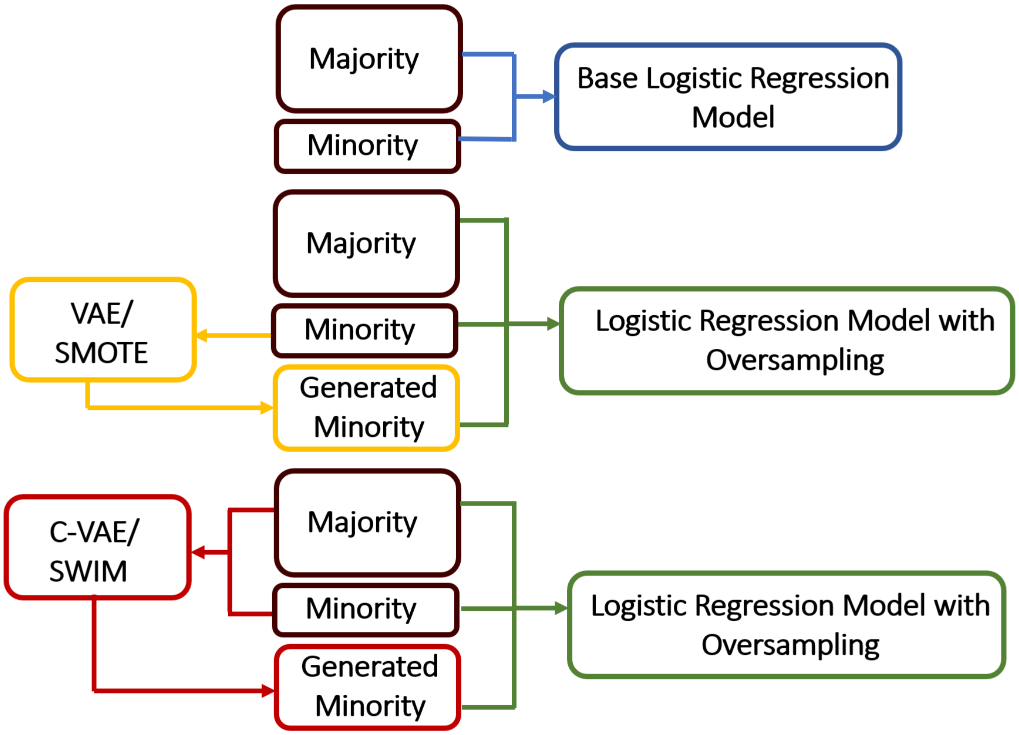

Majority
Minority
Base Logistic Regression Model
VAE/ SMOTE
Majority
Minority
Generated Minority
Logistic Regression Model with Oversampling
C-VAE/ SWIM
Majority
Minority
Generated Minority
Logistic Regression Model with Oversampling

Figure 2-6: 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.

For SWIM, we used the code published by the authors of the paper [25].

2.2.2 Datasets

We tested the generative oversampling method using the Global Registry of Acute Coronary Events (GRACE) [14] and another public Breast Cancer Dataset , which was obtained from the UCI Machine Learning Repository [28] . 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 effectivenss 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 2.1.

2.2.3 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 pa-

Table 2.1: 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%

tients' admission. The extracted features include demographic information (e.g. age, gender), medical history, vital signs on admission (eg. 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.

2.2.4 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. 2-6. We also included SWIM as an alternative oversampling method that uses some information from the majority class. Hyperparameters 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 the original training set.

2.2.5 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 [29]. 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 [30]. 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 [31].

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.

2.3 Results

Table 2.2 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, Sroke, 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 2.3 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 2.2: 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)	**0.87(0.04)**
	VTE	0.62(0.08)	0.56(0.10)	0.56(0.10)	0.60(0.08)	0.72(0.07)	**0.78(0.05)**
	Stroke	0.67(0.05)	0.64(0.06)	0.62(0.06)	0.66(0.07)	0.72(0.05)	**0.78(0.04)**
	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)	**0.72(0.04)**

Table 2.4 and 2.5 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

Table 2.3: 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)	**0.030(0.020)**
	VTE	0.013(0.007)	0.014(0.013)	0.017(0.013)	0.011(0.007)	0.053(0.046)	**0.069(0.042)**
	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)

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.

Table 2.4: 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)	**0.73(0.11)**
	VTE	0.45(0.08)	0.58(0.09)	0.61(0.07)	0.58(0.07)	0.61(0.06)	**0.68(0.07)**
	Stroke	0.55(0.07)	0.61(0.05)	0.64(0.04)	0.63(0.06)	0.63(0.05)	0.63(0.0.05)
	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	**0.88(0.01)**	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)	**0.70(0.03)**

2.4 Discussions

Our results on the GRACE dataset suggest that the C-VAE oversampling method performs best when the number of samples in the minority class is very low. In

Table 2.5: 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)	**0.697(0.032)**

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. 2-1. 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. 2-7 - Fig. 2-9 (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. 2-7 - Fig. 2-9. 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. 2-7 - Fig. 2-9.

The C-VAE clearly learned a better distribution when the positive and negative classes share some common information as shown in Fig. 2-7. 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. 2-7 (b), failed to learn this shared variation 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.

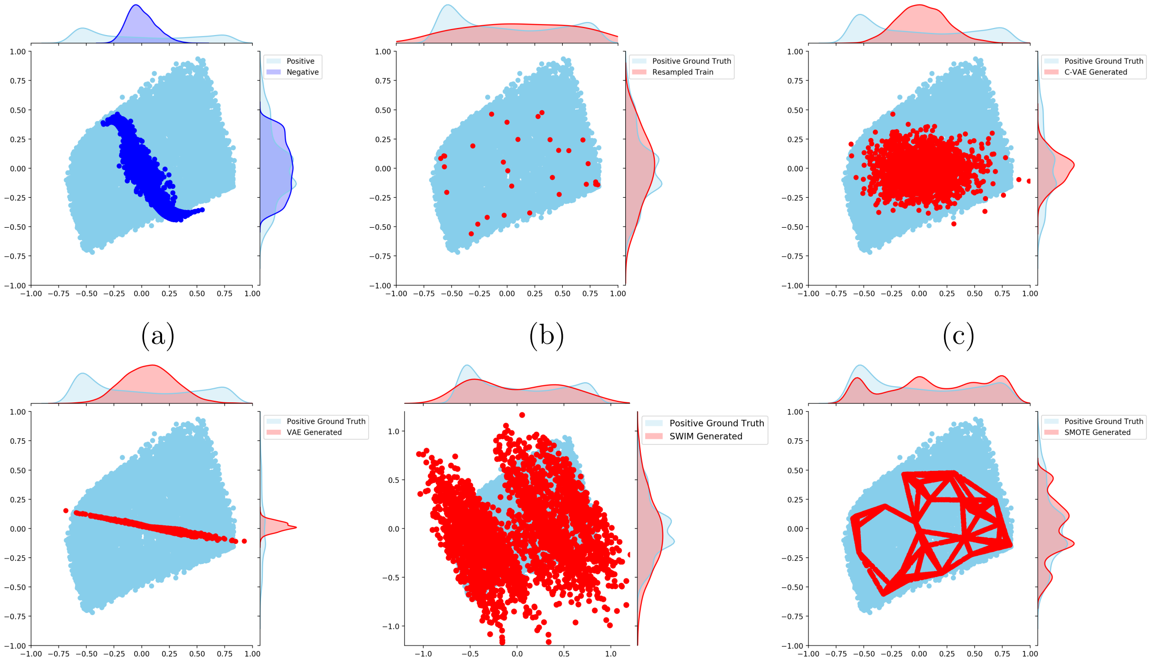
When the positive and negative samples were drawn from the same distribution, both the C-VAE and SWIM learned a better distribution for the minority class, likely because it had more relevant data available to learn from. The VAE on the other hand, again, learned a distribution that exaggerates the variation along the x-axis, as shown in Fig. 2-8 (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. 2-9 (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. 2-7 - Fig. 2-9. 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.

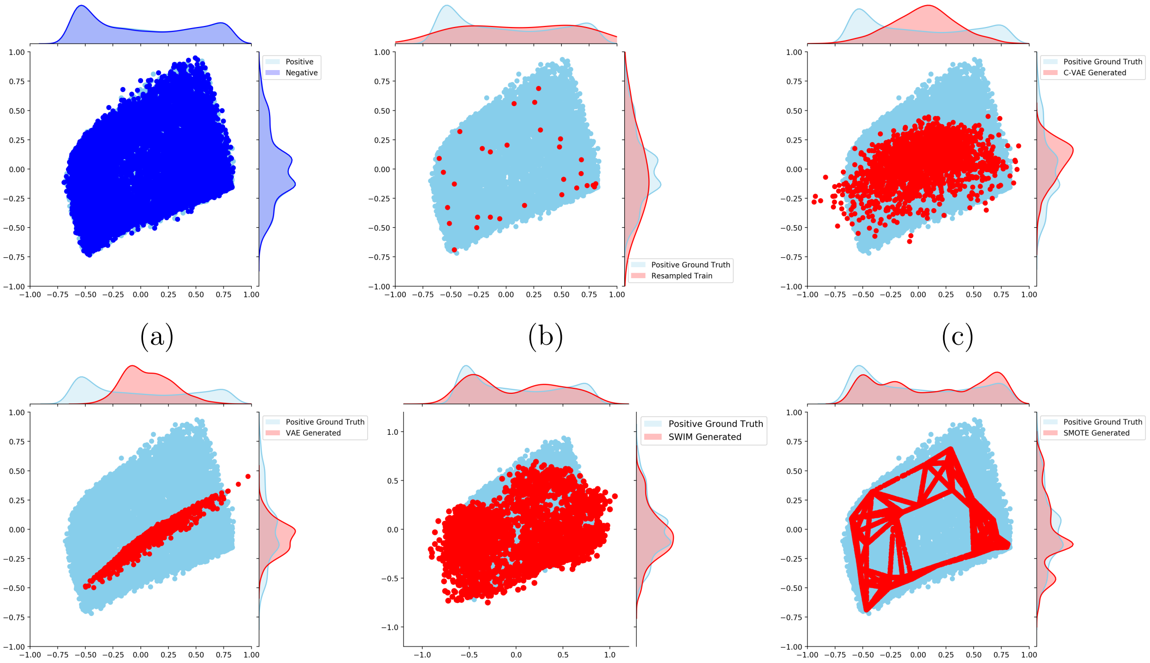
Domain knowledge is necessary to judge whether shared information exists between classes. For example, patients diagnosed with the same disease can be a strong support for the common information and it is also reasonable to assume specific variation exists for patients with adverse outcomes in the cohort.



(a) (b) (c)

(d) (e) (f)

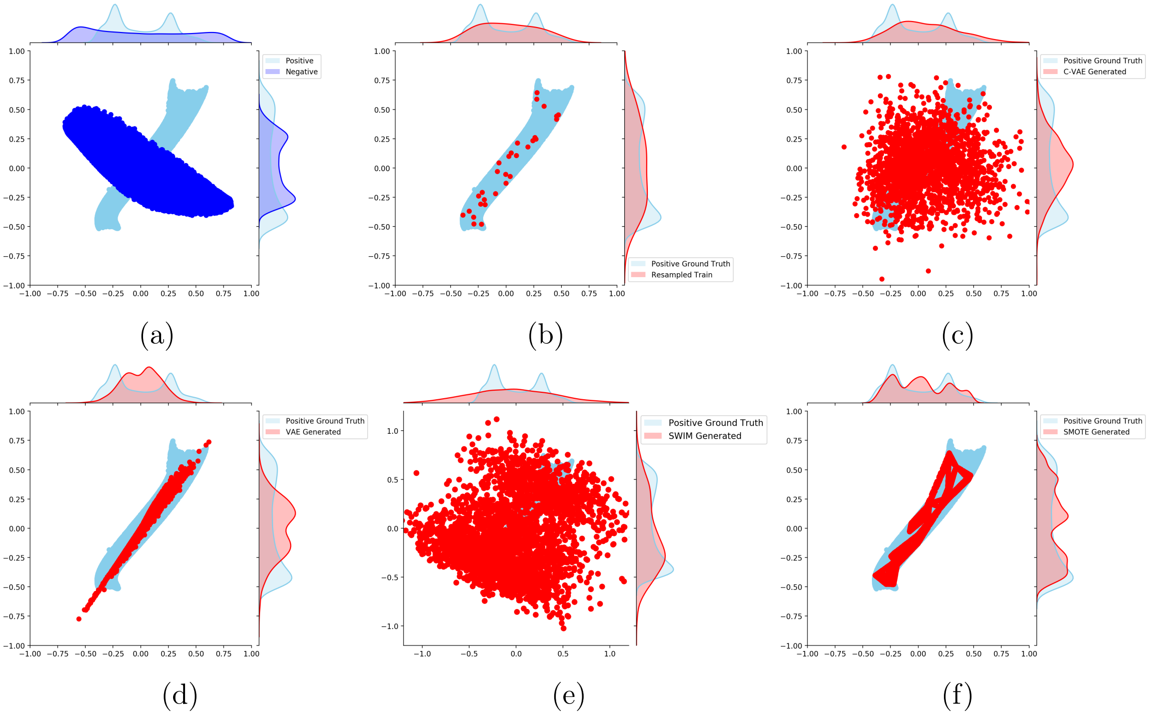
Figure 2-7: Situation 1. Synthetic data experiments when there is shared information between positive and negative classes and private variation for the positive class exists. Shared latent variables exist for both classes but private latent variables only exist for the minority class. (a) PCA plots 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.



(a) (b) (c)

(d) (e) (f)

Figure 2-8: 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.



(a) (b) (c)

(d) (e) (f)

Figure 2-9: 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.

2.5 Conclusions

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.

2.5.1 Limitations

Using a C-VAE is appropriate when there is information common to both the majority and minority classes. Our synthetic data experiments demonstrate 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 decide when this oversampling method is most applicable.

While we demonstrated the use case of this approach with both synthetic and real word data, the type of data we used is limited to be tabular clinical registry. Other types of data like time series, images and videos are not discussed in this work. Nevertheless, oversampling, and more broadly, data augmentation is a common task in learning with all different kinds of data [32]. Therefore, one future direction of work is to expand and transfer the idea of shared and private information into different data types and different generative models.

Another limitation of the method is that only two classes of data can be modeled.

Class imbalance in more sophisticated situations like multi-class classification need further modification of the model. In the next chapter, we extend the C-VAE to enable multiple groups by adding additional private latent variable to each of the groups, while maintaining a shared latent across variable across different groups.

Chapter 3

Quantifying Common Support between Multiple Treatment Groups Using a Contrastive VAE

3.1 Introduction

3.1.1 Estimating Treatment Effect with Observational Data

Data driven machine learning models are being applied with ever increasing frequency in the clinical domain [33]. One fundamental problem that limits their application is that most machine learning models are trained on retro-respective, observational data [34, 35]. This makes it difficult to identify causal relationships, estimate treatment effects, and make unbiased predictions when the model is deployed in practice [36]. Take clinical risk stratification as an example. A risk score typically estimates patient risk using a set of predefined patient characteristics; e.g., predicting death after a heart attack from patient demographics and labs available at admission[4]. Although such models may have significant discriminatory ability, it is not guaranteed that the chosen patient features are causally related to the outcome of interest. Indeed, the existence of unappreciated confounding factors limits the one's ability to make causal statements from such models. As a case in point, patients with high risk features who receive aggressive therapies may have a lower adverse event rate than many patients with low risk features because the administered treatments are effective at lowering the risk of inimical events [37]. However, classifying such patients, with high risk features, as low risk is clearly misleading because their outcome is affected by the treatment decisions of their health care providers. The risk provided by such a model is therefore not an unbiased prediction and may not be appropriate for many patients.

Traditional causal inference methods on observational data estimate such treatment effects by reducing the selection bias via simple statistical methods like matching and re-weighting [38, 39]. These methods usually depend on strong assumptions such as un-confoundedness [40] and common support[41]. Moreover, they typically can only be applied in the setting of a binary treatment decision [42]. Real word clinical data, by contrast, are much more sophisticated; e.g., these assumptions are usually hard to meet, and patients are usually given more than one treatment at a time. Modeling such data requires more complex modeling choices that must deal with class imbalance and data scarcity, as some complex treatment decisions may not be well represented in the dataset.

In this paper, we develop a method that estimates both the treatment effect and the common support of this estimate in a multiple treatment group scenario. Furthermore, the approach effectively addresses the class imbalance and data scarcity - common problems that arise when analyzing more than one treatment at a time. By leveraging this knowledge, we obtain insights into the observed data and develop more accurate clinical risk scores that can help guide clinical decision making.

3.1.2 The Common Support Assumption

Common support is a key assumption in treatment effect estimation models; e.g., convariate adjustment and propensity score matching [43, 44]. Although a number of strategies have been developed to identify and assess the common support assumption in treated vs. control scenarios, simple methods such as comparing bounds of covariates between groups [45] might fail when the corresponding covariate distributions and their overlap are complex and non-linear. Other methods usually can be viewed as a by-product of causal inference models, for example, by bounding the treatment propensity score [46], thresholding data points in matching algorithms [47], or comparing individual-specific posterior distributions for each potential outcome using Bayesian Additive Regression Trees [11]. Recently, Johansson et. al. proposed an interpretable assessment by rephrase the problem into finding minimum volume sets subject to coverage constraints with Boolean rule classifiers [48].

However, all of these methods require accurate modeling for each of the treatment group, and this makes it challenging to extend them to more than two treatment groups. Moreover, class imbalance and data scarcity makes it difficult to build separate models for individual treatment groups.

Contrastive learning algorithms provide a way to learn relationships between two or more data sets that share some common information [17]. A Contrastive-VAE, for example, leverages deep neural network structures to learn nonlinear latent variables for both the shared variation and the unique variation in distinct treatment groups within a given dataset [27]. Contrastive-VAE models can therefore model multiple groups of data simultaneously and yield improved performance relative to individual models, especially in situations of severe class imbalance and data scarcity[49].

3.2 Methods

3.2.1 Contrastive VAE for Multiple Treatment Groups

We extend the Contrastive VAE as described in Chapter 2 to model the distribution of multiple groups of patients features and their outcomes. Instead of having private latent variables only for the minority group, we assume individual private latent variables exist for each of the groups, in addition to common latent variables that model the shared variation between the different groups. Without loss of generality, Figure 3-1 shows the generative model for three treatment groups (T=1, 2, 3).

Let s denote the shared latent variables between treatment groups, $z^{(i)}$ as the private latent variable for the ith treatment group. The generative distribution for the features $x^{(i)}$ and outcome y^i will be

$$p(x^{(i)}, y^{(i)}) = \int_{s,z^{(i)}} p(x^{(i)}, y^{(i)}|s, z^{(i)})p(s)p(z^{(i)})dsdz^{(i)} \tag{3.1}$$

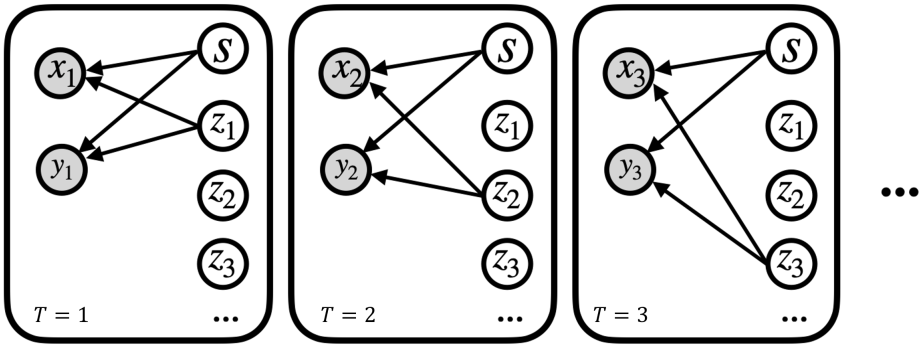


Figure 3-1: Generative model for multiple treatment groups

Here, $p(s)$ and $p(z^{(i)})$ are standard Gaussian prior distribution of the latent variables. The conditional distributions for the observed variables are modeled using a shared neural network decoder f_θ, which takes the shared latent variable and the group specific private latent variable as input.

$$p(x^{(i)}, y^{(i)}|s, z^{(i)}) = \mathcal{N}(f_\theta(s, z^{(i)}, 0), 1) \tag{3.2}$$

Figure 3-2 shows the structure of the Contrastive-VAE, where the optimization object is the sum of the evidence lower bounds (ELBO) of each group, i.e., $\sum_i L(x^{(i)}, y^{(i)})$, where

$$L(x^{(i)}, y^{(i)}) \geq E_{q_{\phi_s}, q_{\phi_{z^{(i)}}}}[f_\theta(x^{(i)}|s, z^{(i)})] \tag{3.3}$$

$$- KL(q_{\phi_s}(s|x^{(i)})||p(s)) - KL(q_{\phi_{z^{(i)}}}(z|x^{(i)})||p(z^{(i)})) \tag{3.4}$$

3.2.2 Estimating Distribution Overlap

The common support is defined by the distributional overlap between different groups of patients. Without loss of generality, we describe the estimation for two treatment

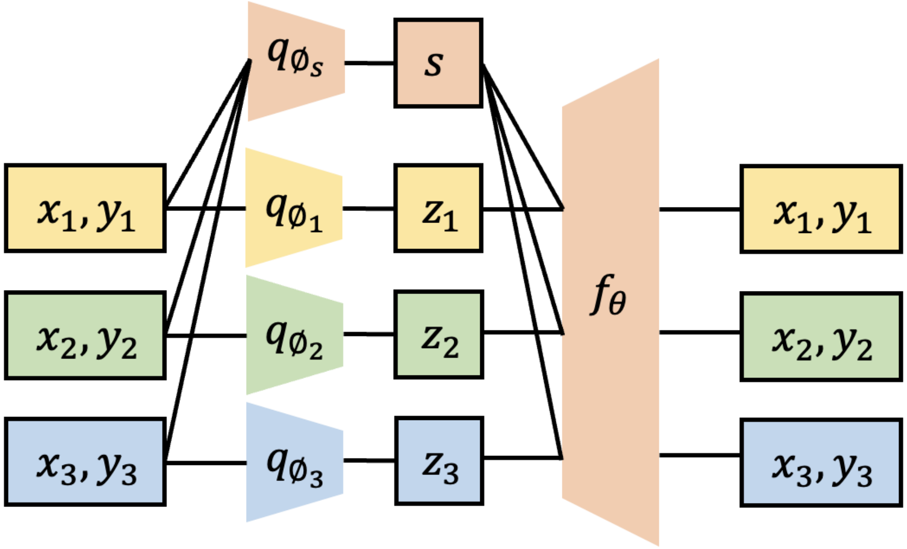


Figure 3-2: Structure of contrastive VAE for multiple treatment groups

groups, where we note the private latent variables to be z^+ and z^-. For a given patient's set of clinical features x, we need to compute

$$\begin{aligned} &support(x) = \min\{P(x|T=1), P(x|T=0)\} \\ &= \min\{\int_{z^+}\int_s\int_y p(x,y|s,z^+)p(s)p(z^+)dydsdz^+, \\ &\int_{z^-}\int_s\int_y p(x,y|s,z^-)p(s)p(z^-)dydsdz^-\} \end{aligned} \tag{3.5}$$

However, direct computation of the data likelihood given an arbitrary generative model is challenging because the associated integral over the entire latent space generally does not have a closed form solution. We therefore used Annealed Importance Sampling (AIS) to estimate the likelihood [50]. The idea behind AIS is to first find a distribution $p_0(x)$ that we can rigorously compute, and then define K intermediate distributions between $p_0(x)$ and $p_K(x) = p(x)$. By estimating the ratio between each of the intermediate distributions, the desired probability can be obtained by multiplying the estimated ratio and the initial probability:

$$p(x) = \hat{r}p_0(x) \tag{3.6}$$

where $\hat{r}$ is the ratio estimated by a Markov Chain Monte Carlo procedure:

$$\hat{r} = \frac{1}{M} \sum_{i=1}^{M} w_{AIS}^{(i)} \tag{3.7}$$

$$w_{AIS} = \frac{p_1(x, z_0)}{p_0(x, z_0)} \frac{p_2(x, z_1)}{p_1(x, z_1)} \dots \frac{p_K(x, z_K)}{p_{K-1}(x, z_K)} \tag{3.8}$$

Here M is the number of independent Markov Chains, z_0 is sampled from the initial prior distribution $p_0(z)$, and z_k for $1 \le k \le K$ are sampled from the transition kernel $\mathcal{T}_{k-1}(z_k|z_{k-1})$.

We chose the intermediate distributions to be

$$p_k(x, z) = p_0(x, z)^{1-\beta_k} p_K(x, z)^{\beta_k} \tag{3.9}$$

where β_0,...,β_K are monotonically increasing numbers from 0 to 1.

3.2.3 Estimating Individual Treatment Effects

We constructed one Contrastive-VAE to model both the features and the outcome because this allows us to estimate both the treatment effect as well as the corresponding common support for this estimate. Under the assumption of ignorability, the individual treatment effect (ITE) can be computed as,

$$\begin{aligned} ITE(x) &= E[y^+|x, T = 1] - E[y^-|x, T = 0] \\ &= \int_{y^+, s, z^+} y^+ p(y^+|s, z^+) p(s, z^+|x^+) dy ds dz^+ \\ &\quad - \int_{y^-, s, z^-} y^- p(y^-|s, z^-) p(s, z^-|x^-) dy ds dz^- \end{aligned} \tag{3.10}$$

To estimate this conditional probability of the outcome y given the features x, we used Gibbs Sampling to sample y while keeping x fixed.

3.2.4 Confidence Interval with Regard to Distribution Overlap

In order to intuitively explain the effect of overlap on the estimated ITE, we introduce a confidence interval with regard to distribution overlap. The vanilla definition of confidence interval is

$$CI = I\bar{T}E \pm Z_\alpha \frac{\sigma}{\sqrt{n}} \tag{3.11}$$

where σ is the standard deviation of the estimated treatment effect, which can be estimated using samples from the Contrastive-VAE. Z_α is the confidence value at level α, so that

$$\begin{aligned} & P\left[I\bar{T}E - Z_\alpha \frac{\sigma}{\sqrt{n}} < IT\bar{E} < I\bar{T}E + Z_\alpha \frac{\sigma}{\sqrt{n}} \right] \\ & = \alpha \end{aligned} \tag{3.12}$$

where α is the associated probability of the confidence level and n is the sample size.

By replacing the sample size to an effective local sample size in the overlap distribution, we get

$$\begin{aligned} & P[I\bar{T}E - Z_\alpha \frac{\sigma}{\sqrt{N \min(P^{+}(x), P^{-}(x))}} \\ & < I\bar{T}E < I\bar{T}E + Z_\alpha \frac{\sigma}{\sqrt{N \min(P^{+}(x), P^{-}(x))}}] \\ & = 0.95 \end{aligned} \tag{3.13}$$

where $P^{+}(x)$ and $P^{-}(x)$ are probabilities of the given feature x in the two groups. For continuous variables, we convert the density to probabilities by discretizing the feature space so that the minimum will be a number between 0 and 1.

3.3 Experiments on Synthetic Data

3.3.1 Experimental Design

We designed a series of synthetic data experiments to evaluate a Contrastive-VAE's ability to estimate the distributional overlap as well as the treatment effect. For these experiments we construct two groups of patients. The first group receives treatment (T=1) and the second group does not receive the treatment (T=0). We assume that patients data represent samples from 3D Dirichlet distributions. Samples of 3D Dirichlet distributions lie in a 2-simplex, a 2D triangle in 3D space, which mimics the realistic scenario where patient data corresponds a relatively low dimensional manifold in a high dimensional space [51]. We also assume that both treatment groups share some information - which is typically true in practice. For example, it is often of interest to assess the effect of a given therapy on a specific patient population. Although patients in different treatment groups receive different therapies, they nonetheless have the same diagnosis and/or disease. To simulate this situation, we assume one of the marginal distributions to be the same for different treatment groups, i.e.

$$\mathbf{x}^{+} \sim Dir(\alpha_1^{+}, \alpha_2^{+}, \alpha_3^{+}) \tag{3.14}$$

$$\mathbf{x}^{-} \sim Dir(\alpha_1^{-}, \alpha_2^{-}, \alpha_3^{-}) \tag{3.15}$$

where, $\alpha_1^{+} = \alpha_1^{-}$ and $\alpha_1^{+} + \alpha_2^{+} + \alpha_3^{+} = \alpha_1^{-} + \alpha_2^{-} + \alpha_3^{-}$. The outcomes for each group is defined by 2 non-linear functions $f^{+} : \mathcal{X} \rightarrow \mathcal{Y}$ and $f^{-} : \mathcal{X} \rightarrow \mathcal{Y}$, that maps the patients feature $\mathbf{x} \in \mathcal{X}$ to an outcome $y \in \mathcal{Y}$ for either received the treatment $(T = 1)$, or dose not received $(T = 0)$. We set $\{\mathbf{x}^{+}, y^{+} = f^{+}(\mathbf{x}^{+})\}$ and $\{\mathbf{x}^{-}, y^{-} = f^{-}(\mathbf{x}^{-})\}$ as the observed data, while the counter-factual outcomes $y'^{+} = f^{-}(\mathbf{x}^{+})$ and $y'^{-} = f^{+}(\mathbf{x}^{-})$ were concealed from the model and only used for evaluation of the treatment effect.

In the first experiment, we demonstrate the distributional overlap of the two groups of patients can be reproduced by the Contrastive-VAE. Ground truth probability density was used to evaluate the density estimated by AIS of the trained

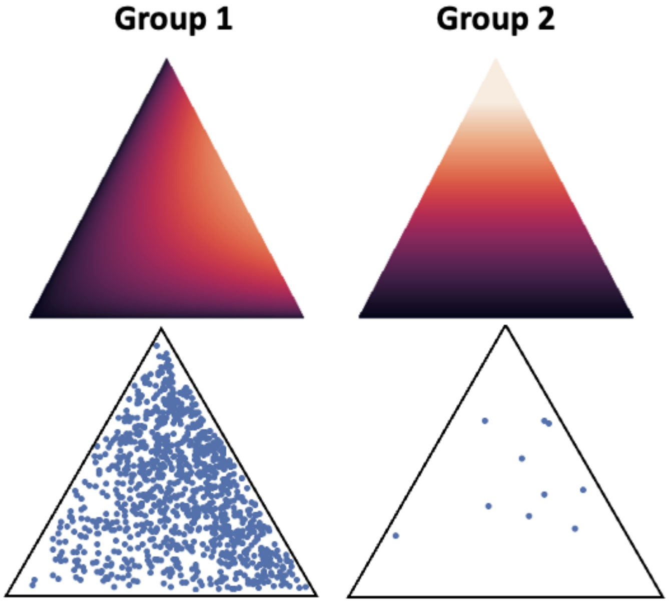


Figure 3-3: Probability densities of two simulated Dirichlet distribution and imbalanced samples of 1000 vs 10.

Contrastive-VAE. We also compared the Contrastive-VAE to Kernel Density Estimation (KDE) and standard VAEs that model the distribution of each group independently. The KDEs used a Gaussian kernel and their bandwidths were decided by a 5-fold cross validation. Both the KDEs and standard VAEs were trained separately for different groups of data, while the contrastive-VAE was trained with the two groups together. We conducted the experiments with different levels of class imbalance. An example of probability densities of two simulated Dirichlet distribution and imbalanced samples of 1000 vs 10 is shown in Figure 3-3. The 3 dimensional densities and samples are plotted in a 2-simplex.

We did an additional experiment to show that a Contrastive-VAE can model patients with more than two treatment groups. This allows us to extend the model when more than one treatment is given to patients. For example, if we want to compare the effect of treatment A and B, we can find out the overlap of three populations, those who received A, those who received B and those who received nothing. Only for patients with support in all of the three populations, the estimated treatment effect can be thought of as reliable. In this experiment, we sampled three groups of data

x_1, x_2 and x_3 from three different Dirichlet distributions. The number of samples from each of the groups are imbalanced to mimic the real observational dataset. A Contrastive-VAE with three private latent space, KDE and three independent VAEs were compared to reconstruct the probability distribution and the overlap of the three groups.

In the second experiment, we trained a Contrastive-VAE on imbalanced data to learn the joint distribution of features $\mathbf{x}$ and outcome y. We then use the trained model to predict the counter-factual outcome for patients of the 2 groups, i.e. y'^{+} for $\mathbf{x}^{+}$ and y'^{-} for $\mathbf{x}^{-}$, and the treatment effects for the 2 groups, i.e. $y^{+} - y'^{+}$ and $y'^{-} - y^{-}$. As the treatment effect estimation is only valid for patients that satisfied the common support assumption, we excluded patients with a overlap probability below a certain threshold. Then the estimated treatment effects were compared with the ground truth simulated by functions f^{+} and f^{-}.

Additionally we demonstrated the confidence interval with regard to common support using samples from 1 dimensional normal distributions. We considered three different situations where the distribution of the two treatment groups are partially overlapped, identical or apart from each other. In each situation, we estimated the treatment effects with a linear function using the samples and computed their confidence interval using equation 3.13.

3.3.2 Results of Distributional Overlap for Two Treatment Groups

Figure 3-4 shows the result of the first experiment with a highly imbalanced training set, 1000 training points for group 1 vs. 10 points for group 2. Probability density for the two groups and their overlap are plotted on the 2D simplex, the plane where all the data exist. It can be seen that the KDE failed to estimate the probability density for both of the 2 groups. An alternate approach is to model each of the two groups separately, using two independent VAEs - one for each treatment group. However, the VAE trained on the group that contains only 10 samples is a poor presentation

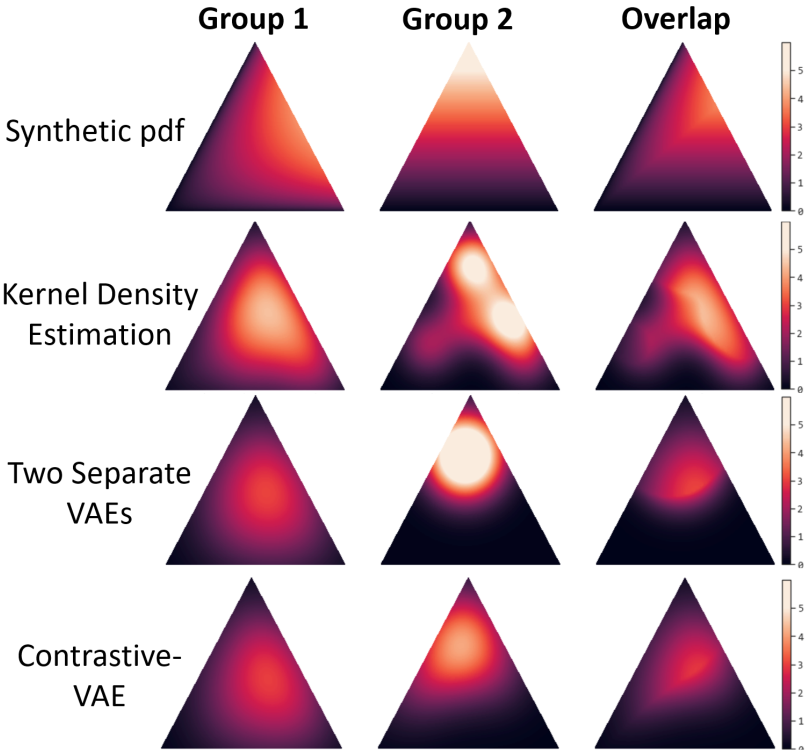


Figure 3-4: Values of probability densities plotted on the 2D simplex with color. Group 1 and 2 refers to the 2 Dirichlet distributions in the synthetic experiments.

of the underlying distribution for this group.

Figure 3-6 (a) shows the mean squared error of the overlap probability estimated by the above three methods, in different level of class imbalance. Contrastive-VAE gives the significantly lower error, compared to KDE and the standard VAEs, in all situations.

3.3.3 Distributional Overlap of Three Treatment Groups

Figure 3-5 shows the results of three groups of data and their overlap. The training set size of the three groups are 1000, 50 and 10 to mimic the class imbalance in real world datasets. Similar to the overlap experiments in 3.3.2, the KDE failed to restore the smooth distribution in all three groups. Standard VAEs works well when efficient training data were provided for group 1 and 2, but did poorly for group 3 when only 10 points were used to training. The Contrastive-VAE was best able to reproduce the ground truth probability densities and their overlap.

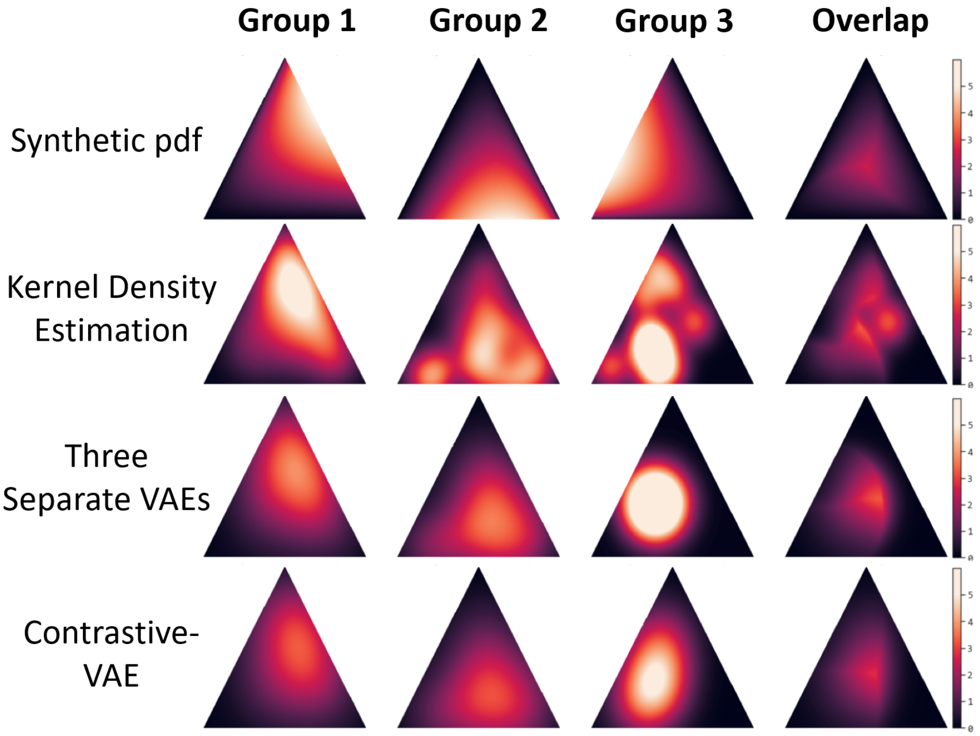


Figure 3-5: Values of probability densities plotted on the 2D simplex with color. Group 1-3 refers to the 3 Dirichlet distributions. Training set size for each of the groups are 1000, 50 and 10. Reconstructions of the probability densities and their overlap are compared using KDE, 3 independent VAEs and the Contrastive-VAE. Mean squared errors are 0.66 for KDE, 0.12 for 3 VAEs and 0.08 for Contrastive-VAE.

3.3.4 Results of Treatment effect estimation

Figure 3-6 (b) - (c) shows the mean squared error of the estimated treatment effect compared to simulated ground truth. To see the effect of the common support assumption, we altered the overlap threshold that decides which predictions are used to compute the treatment effect. With 0 overlap probability threshold, the common support assumption is completely ignored and the error in the estimated treatment effect is the largest. When non-zero overlap probability thresholds are used, the treatment effect estimation is only computed for patients who have an overlap probability that is larger than this threshold. The mean squared error decreases as the threshold increases, thereby demonstrating that accurate estimation of the treatment effect requires significant common support.

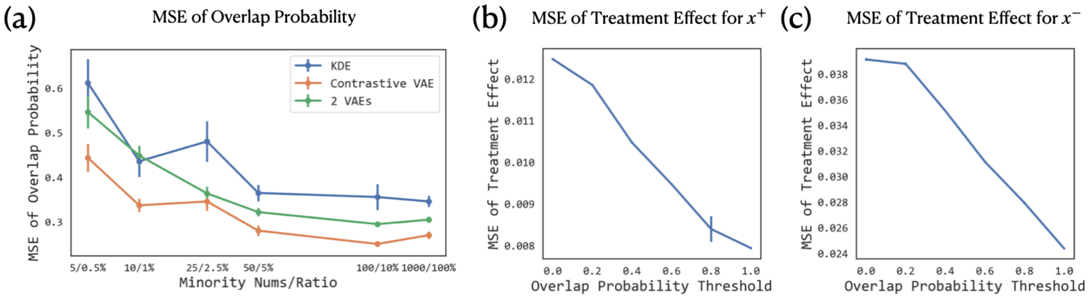


Figure 3-6: (a) Mean squared error of the overlap probability estimated by KDE, Contrastive-VAE and 2 standard VAE for 2 Dirichlet distributions with different levels of class imbalance. (b)-(c) Mean squared error of predicted treatment effect vs. overlap probability threshold for patients who received the treatment and who did not. All error bars represent the standard error over 10 bootstraps.

3.3.5 Simulation Results of Confidence Interval

Figure 3-7 shows the simulated results of three different feature distributions and the estimated ITE as well as its confidence interval. Figure 3-7 (a) - (c) show the situation where the two groups distribution are identical, where (a) gives the ground truth distribution of the feature in two treatment groups. The simulated outcome and training samples are shown in (b). (c) shows the estimated ITE and 95% confidence interval with regard to the overlap distribution. In this case, we see an extremely small interval in the region that contain training samples, as the common support assumption is fully satisfied. (d)-(f) show a partial overlapping situation. Here, the interval is dramatically smaller in the overlap region, compared to outside, which intuitively demonstrate the importance of common support for the ITE estimation. (g) - (i) show the other extreme case where there's hardly any overlap between the two distributions, where the extremely large interval indicates the ITE estimation is barely reliable when considering the common support assumption.

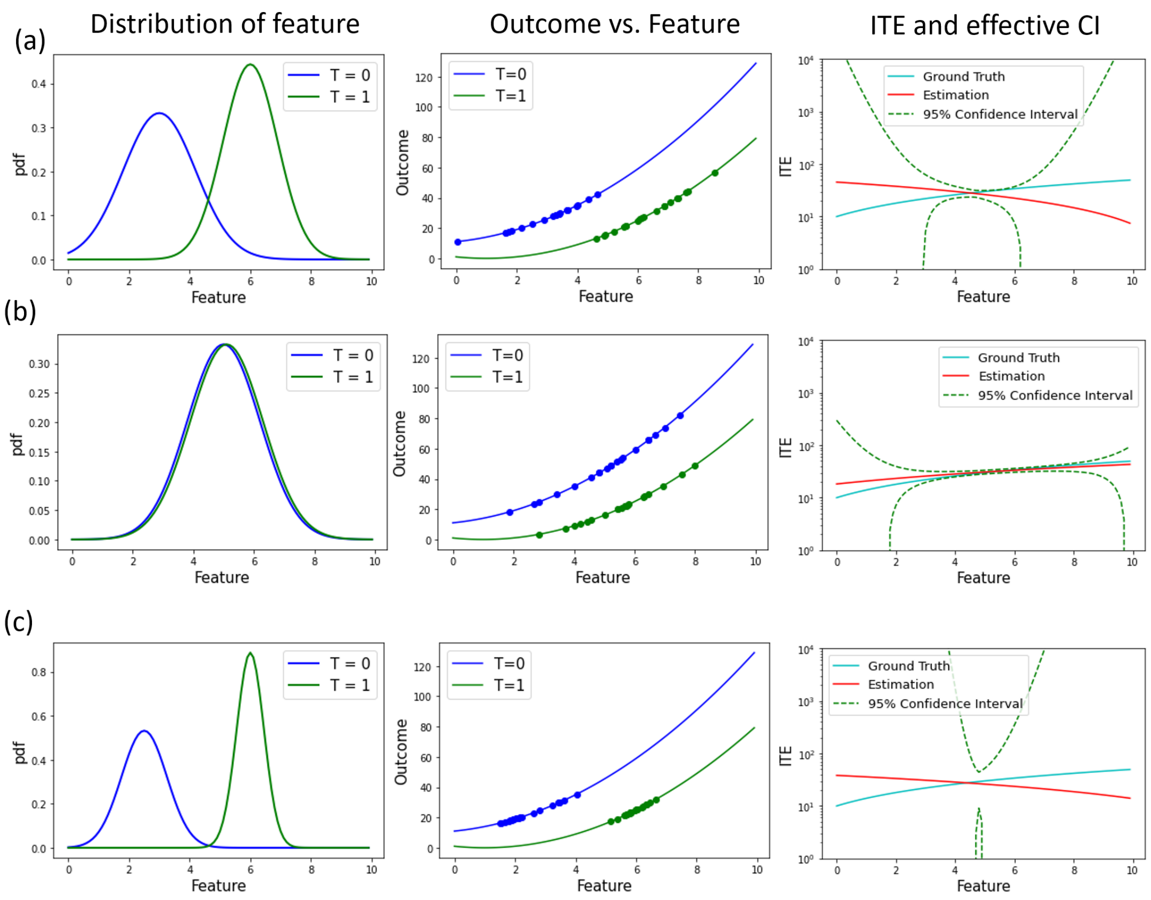


Figure 3-7: Confidence interval with regard to common support for different overlap levels. (a) some overlap; (b) full overlap; (c) no overlap

3.4 Experiments on Real Clinical Data

3.4.1 Experimental Design

We applied our method to a real world clinical data set, the Global Registry of Acute Coronary Events (GRACE). GRACE enrolled over 70,000 patients from 250 hospitals in 30 countries [52]. Patients enrolled in the GRACE registry were diagnosed with an acute coronary syndrome (a constellation of signs and symptoms consistent with reduced blood flow to the heart). Patients were followed and their outcomes and therapeutic interventions were recorded.

We chose 2 major treatments, Percutaneous Coronary Intervention (PCI) and Coronary Artery Bypass Grafting (CABG), and trained a Contrastive-VAE with three separate private latent variables, representing each of the three groups - those who only receive a PCI, those who only had a CABG, and patients who received neither treatment as the control group. For patients characteristics, we used 8 features that is used in GRACE score, a risk stratification for ACS patients [14].

In order to estimate the effect of the treatments, we consider patients that have common support in the treated and the control groups. For example, for the treatment PCI, we compare the distribution of the patients who received PCI to those who did not receive either PCI or CABG, and selected patients with confidence interval of ITE that below a threshold. Within the selected patients, we used the trained Contrastive-VAE to estimate the treatment effect of PCI, where the outcome of interest is death within 6 months of presentation. Those patients with a treatment effect greater than 6% were considered to be the effective group. A cutoff of 6% was used because this corresponds to the prevalence of death in the overall dataset. The non-effective group, on the other hand, corresponds to the patients whose estimated treatment effect is smaller than 6%. We compared patient characteristics between the effective group and the non-effective group for different level of effective confidence to analyze the importance of the common support. Figure 3-8 shows a diagram for different confidence levels. At 0% CI, the common support assumption is not considered at all and the ITE is a simple point for each patient. In this case, patients are divided into the effective and non-effective groups simply by comparing their ITE to the threshold. At 90% confidence level, each point of ITE has an error bar showing the effective confidence interval of the ITE. A wider interval is shown in the 95% case. With the effective CI for the ITE, we now only selected patients with lower bound greater than the threshold into the effective group and those with their upper bound

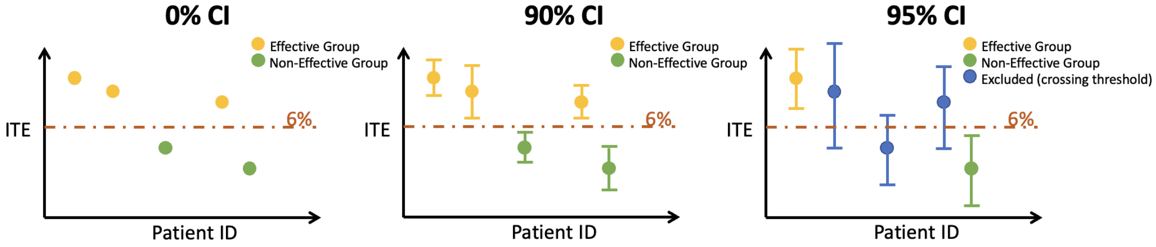


Figure 3-8: Diagram to show effective and non-effective groups with different level of proxy confidence interval.

lower than the threshold into the non-effective group. We compare patients features between the effective group and the non-effective group in order to see for which patients the treatment is more effective. By changing the confidence level, we show the impact of the common support assumption on the conclusions we can make about the observational data.

3.4.2 Treatment Effects and Common Support

Figure 3-9 (a) - (b) shows the average age and systolic blood pressure for patients in the CABG effective and non-effective group, and the corresponding p value that quantifies the statistical significance of this difference. Similarly, Figure 3-9 (c) shows the expected KILLIP class (a metric that quantifies the extent of heart failure on clinical exam at presentation) for patients in the PCI effective and non-effective group. As we can see from the figures, considering the confidence interval with regard to the distribution overlap can change the group characteristics significantly, and therefore leads to completely different clinical conclusions. For example, for the CABG treatment in (a) and (b), by considering only the patients with ITE within a small confidence interval, one can conclude that CABG is effective for patients with a younger age and lower systolic blood pressure ($p < 0.05$), however, a similar analysis that uses all the data (i.e., a large confidence interval) suggests that CABG does not derive a benefit irrespective of age ($p > 0.05$). Similarly, for the PCI treatment, conclusions that are made without considering the extent of the common support might be not valid when the overlap is considered, as shown in Figure 3-9 (c). When the threshold for the confidence interval is large, one can conclude the treatment is effective for patients with a larger KILLIP class, but the conclusion would be not valid if we restrict the patients to those with smaller confidence interval for their ITE.

3.5 Discussions

In this work, we demonstrate that both the treatment effect and the common support can be accurately estimated using a single Contrastive-VAE. The key point that

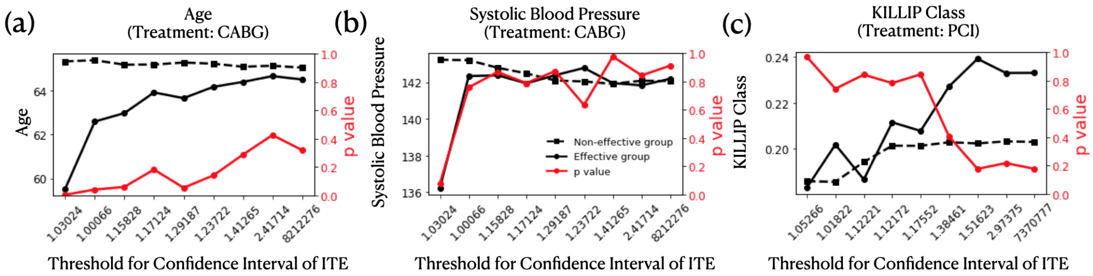


Figure 3-9: (a) - (b) Average age and systolic blood pressure compared between effective and non-effective groups for treatment CABG. (c) Similar result for KILLP class in threatment PCI.

makes our method different from traditional propensity matching approaches is that we approach the problem in a parametric way, where we model distributions explicitly for each of the treatment groups. The method allows us to model multiple treatment groups simultaneously and effectively deals with data scarcity - a common problem in real world datasets where patients can receive multiple different treatments. We demonstrate that a Contrastive-VAE can be used to discover meaningful clinical insights, even when data are highly imbalanced and sometimes scarce for certain treatment combinations.

A Contrastive-VAE is appropriate for this class of problems because it leverages the shared information between different treatment groups. Although patients in different groups may be treated differently, they often share the same diagnosis, and latent factors that lead to similar observed clinical or demo-graphical features. Having said this, it is important to stress that the shared information is an assumption and should be treated as a inductive-bias that arises from domain specific knowledge. If the two groups do not share any common information, a situation that is not typical of treatment groups in observational datasets, then the Contrastive-VAE may not yield suitable estimates of the common support.

In order to explain the estimated overlap probability and makes it easier for clinical applications, we proposed a effective confidence interval with regard to overlap. The number of sample size in the vanilla definition of confidence interval is replaced by a effective sample size $N \min(P^{+}(x), P^{-}(x))$. Here we use the probability, in stead of density, to make the weighting factor a number between 0 and 1. For continuous variables, this was achieved by discretizing the feature space and multiplying the density by a pre-chosen volume size. However, the volume size for different data sets and distributions might be chosen differently and therefore makes it difficult to compare the confidence interval between data sets. One possible solution is to first map the feature space to a latent space of fixed size and asses the common support assumption in the latent space. In this way, the volume size will be a fixed factor and comparison between different data sets will be unbiased.

3.5.1 Limitations

One major limitation of the method is the computational complexity to estimate the probability density from a Contrastive VAE. The annealed importance sampling we used in this section depends on a Markov Chain Monte Carlo process that is sequential in its nature. The process can be accelerated by substituting the Metropolis Hastings transition kernel with Hamiltonian Monte Carlo, which can dramatically improve the random walk with knowledge of the gradient. Another potentially interesting direction is to use models like normalizing flows [53] where the probability densities can be analytically expressed. Normalizing flows construct complex distribution by transforming a simple probability density through a series of invertible mappings. A normalizing flow with shared and private sectors in the latent density can potentially be a more efficient way to model and estimate probability densities across different treatment groups.

The other major limitation of this approach is we still need to assume the ignorability assumption, that all confounding factors must be observed, in order to evaluate the observational data for reliable treatment effect estimation. In another word, this approach alone can only evaluate the common support assumption, another method to evaluate the ignorability must be applied in order to make sure the observational data satisfy the two assumptions for unbiased treatment effect estimation.

Chapter 4

Feature selection for Treatment Effect Estimation

4.1 Introduction

Estimating the individual treatment effect using observational data is in general a challenging problem. Current methods rely on two critical assumptions: 1) Ignorability: that all confounding factors are observed, and 2) Common support, that it is possible to observe any given set of patient features in each treatment group [10]. For the purpose of satisfying the ignorability assumption, it is common to consider as many features as possible, as more features reduce the chance that a confounding factor is missing. However, too many features in a dataset do not always lead to better results. Including a large number of features in a model may make treatment effect estimation more difficult because the resulting feature set may contain features that are irrelevant for the outcome of interest. To give an extreme example, a patient's date of admission typically appears in the electronic health record. Using this as a feature, however, would violate the common support assumption if no other patients in the dataset at hand were admitted that day. Therefore, removing irrelevant features in an observational dataset helps to ensure that the common support assumption is met and that robust estimates of the individual treatment effect are obtained. The challenge is reliably identifying, *a priori*, what features are irrelevant to the outcome of interest using observational data.

In this work, we use a Contrastive Variational Autoencoder [17] to model the distribution of patients data in different treatment groups. We then use these distributions to find a reduced feature subset that only contains features that are most relevant to the outcome of interest, such that the probability of the outcome conditioned on the full set of patient features is equal to the probability of the outcome conditioned on the smaller subset containing only relevant features. We use annealed importance sampling to calculate the needed conditional probabilities using the learned distributions from the CVAE.

4.1.1 Related Work

Estimating treatment effect from observational data has been the focus of a number of studies in recent years [54] [42]. In order to minimize the number of non-confounding features (e.g., features that are least relevant to the outcome of interest), it is common to select a subset of variables using a formalism that models the outcome as a linear function of patient features. Shortreed et al. proposed outcome adaptive lasso on the logistic regression of the propensity score to do this [55]. A similar method is considered in Greenewald et al's work [12] where a regularized regression on the outcome given a subset of features is used.

Another growing body of recent research focuses on projecting the variables to a different space to fulfill the assumptions, instead of using the features from the observed space. Johansson et al. [56] mapped the observed features into a latent representation where there model is trained to predict the outcome using the representation and minimize the discrepancy of the representation between different treatment groups at the same time. Yao et al [57] proposed a deep representation learning model that maps covariate space to a latent space, which balances the data distribution and that preserves the local similarity between neighbors.

4.2 Methods

4.2.1 Problem Setting

Let X be the set of all observed features, Y be the outcome and T be the treatment of interest. For simplicity, we assume Y and T are both binary variables. The goal of feature selection for treatment effect estimation is to find a subset $S \subset X$ so that the estimate of the treatment effect with S is unbiased compared to the entire set X.

Our proposed method consists of two parts. In the modeling part, we utilized a Contrastive VAE, which leverages the shared information between different treatment groups to model the distribution of the features, treatments and the outcome, as described in Section 3.2.1. At inference time, we check the conditional independence of the outcome Y and a certain feature x_i by comparing $p(Y|X,T)$ and $p(Y|S,T)$, where $S = X \setminus x_i$. If there's no difference between $p(Y|X,T)$ and $p(Y|S,T)$ then x_i is not a confounding variable and it is removed from S for future comparison. An overview of the procedure can be seen in the flowchart shown in Figure 4-1.

Central to this success of this approach is an efficient method for comparing two conditional distributions, $p(Y|X,T)$ and $p(Y|S,T)$. This, however, is not straightforward as the dimension of X and S can be high. Therefore, we reformulate this task into an optimization problem where the goal is to find a point, x_{max}, which maximizes the difference between the conditional distributions $p(Y|X,T)$ and $p(Y|S,T)$. If there exists a point where the difference between $p(Y|X,T)$ and $p(Y|S,T)$ is greater than zero, then S is missing at least one confounding variable (Figure 1).

We used a simulated annealing (SA) protocol to search for the point that maximizes the Jensen-Shannon divergence (JSD) between the distributions. To compute the associated probabilities, we used Annealed Importance Sampling (AIS) with the learned joint distribution arising from the trained Contrastive VAE.

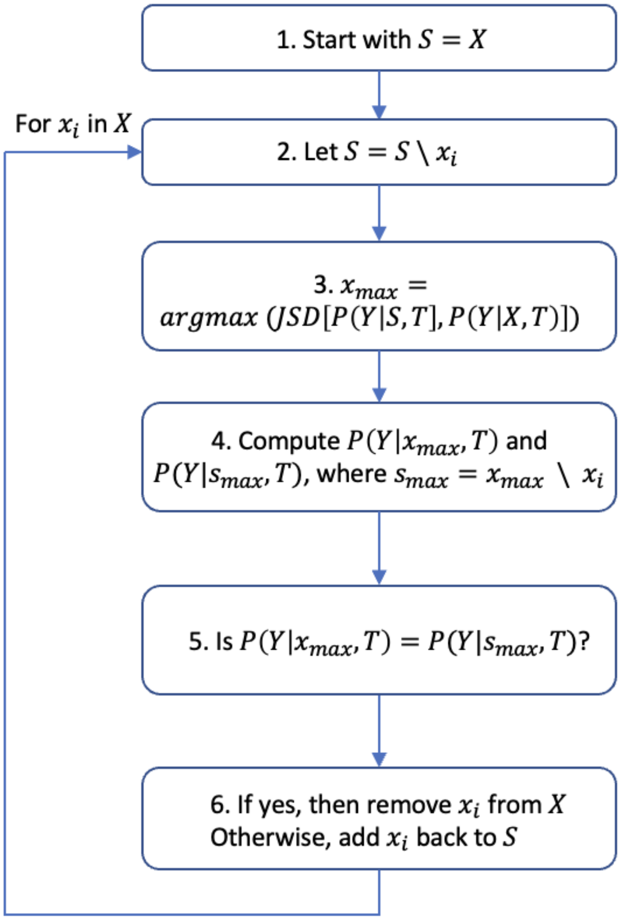

1. Start with $S = X$
For x_i in X
2. Let $S = S \setminus x_i$
3. $x_{max} = argmax\ (JSD[P(Y|S,T], P(Y|X,T)])$
4. Compute $P(Y|x_{max}, T)$ and $P(Y|s_{max}, T)$, where $s_{max} = x_{max} \setminus x_i$
5. Is $P(Y|x_{max}, T) = P(Y|s_{max}, T)$?
6. If yes, then remove x_i from X Otherwise, add x_i back to S

Figure 4-1: Procedure of proposed feature selection algorithm

4.2.2 Inferring Conditional Independence

In order to check conditional independence of certain features with regard to the outcome, we used a leave-one-out algorithm to compare $p(Y|X,T)$ and $p(Y|S,T)$, where S is the set of all features excluding the ith variable x_i, $S = X \setminus x_i$. If removing such a feature results in no difference between $p(Y|X,T)$ and $p(Y|S,T)$, then feature x_i would not be a confounding variable and excluding it would not make the estimation of treatment effect biased. This procedure in described in the flowchart of Figure 4-1 from step 2 to step 6.

In step 2, we choose to remove feature x_i from S. The comparison between $p(Y|X,T)$ and $p(Y|S,T)$ is then converted to an optimization problem, as shown in step 3,

$$\mathbf{x}_{max} = arg \max_{\mathbf{X}}(JSD(p(Y|X,T), p(Y|S,T))) \tag{4.1}$$

We used a simulated annealing algorithm to maximize the Jensen Shannon Divergence (JSD) between the two distributions, as no analytical expression for the conditional probabilities exist. After finding x_{max}, we computed the probability of $p(Y|x_{max},T)$ and $p(Y|s_{max},T)$ where $s_{max} = x_{max} \setminus x_i$. To determine whether these two probabilities are the same, as shown in step 5, we did a paired t test for the null hypothesis $H_0 : P(Y|x_{max},T) = P(Y|s_{max},T)$ with 10 bootstraps of computing the two values. If the p-value from the t test is smaller than 0.05, we would reject the hypothesis and conclude $P(Y|X,T)$ is different from $P(Y|S,T)$ and therefore x_i is a confounding factor that cannot be removed. We then added x_i back to S and move on to the next feature. If the p-value is greater than 0.05, then x_i can be safely removed as $P(Y|X,T)$ is the same as $P(Y|S,T)$ with x_i removed.

Exact computation of the conditional probabilities, $P(Y|X,T)$ and $P(Y|S,T)$ for a Variational Autoencoder is in general not possible, as it involves an intractable integration over latent variables. We therefore used Annealed Importance Sampling (AIS) to estimate these probabilities [50]. For simplicity, assume we want to estimate $p(x) = \int p(x,v)dv$ where v is a high dimensional latent variable. AIS defines a series intermediate distributions between the target distribution $p(x)$ and a reference distribution $p_0(x)$ where the latter can be computed exactly. By estimating the ratio between each of the intermediate distributions, the desired probability can be obtained by multiplying the estimated ratio and the initial probability:

$$p(x) = \hat{r} p_0(x) \tag{4.2}$$

where $\hat{r}$ is estimated using a Markov Chain Monte Carlo procedure:

$$\hat{r} = \frac{1}{M} \sum_{i=1}^{M} w_{AIS}^{(i)} \tag{4.3}$$

$$w_{AIS} = \frac{p_1(x, v_0)}{p_0(x, v_0)} \frac{p_2(x, v_1)}{p_1(x, v_1)} ... \frac{p_K(x, v_K)}{p_{K-1}(x, v_K)} \tag{4.4}$$

Here M is the number of independent Markov Chains, v_0 is sampled from the initial prior distribution $p_0(v)$, and v_k for $1 \leq k \leq K$ are sequentially sampled from the transition kernel $\mathcal{T}_{k-1}(v_k|v_{k-1})$.

We chose the intermediate distributions to be

$$p_k(x, v) = p_0(x, v)^{1-\beta_k} p_K(x, v)^{\beta_k} \tag{4.5}$$

where β_0,...,β_K are monotonically increasing numbers from 0 to 1. For the Markov Chain transition kernel $\mathcal{T}_{k-1}(v_k|v_{k-1})$, we used Metropolis-Hastings algorithm with a Gaussian proposal distribution

$$g(v_k|v_{k-1}) = \mathcal{N}(v_{k-1}, \sigma) \tag{4.6}$$

where σ determines the step size of the Gaussian random walk. The candidate sample from g would be accepted with a probability equals to

$$A(v_k, v_{k-1}) = \min\left(1, \frac{p_k(x, v_k)}{p_k(x, v_{k-1})}\right) \tag{4.7}$$

If not accepted, then $v_k = v_{k-1}$.

4.3 Experiments on Synthetic Data

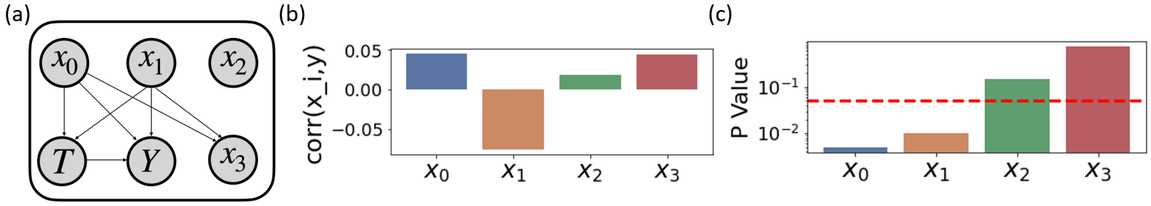


Figure 4-2: Binary synthetic experiments. (a) Generative process of 4 feature variables and an outcome variable. (b) Correlations with the outcome of each features. (c) p-value for the hypothesis $H_0 : P(Y|S,T) = P(Y|X,T)$ where $S = X \setminus x_i$

To test the method, we designed a synthetic data experiment using a pre-specified causal relationship between random variables as shown in Figure 4-2 (a). Here we have 4 features x_0, x_1, x_2, x_3 and an outcome y. Among the 4 features, x_0 and x_1 are causally related to y and x_3, so that x_3 will be correlated to y, but still independent to y when conditioned on x_0 and x_1. The other variable x_2 is irrelevant to the outcome and the treatment. Using generated synthetic data, we computed the correlation between each of the features and their outcome (Figure 4-2 (b)). We then trained variational auto-encoders on the generated data and applied the inference algorithm to see whether the proposed method can correctly identify the conditional independence relationship and exclude $x2$ and x_3 from the list of confounding features for further causal discoveries.

We used the procedure outlined in Figure 4-1 to identify features causally related to the outcome using only the synthetically generated data. As described in section 4.2.2, the procedure testing whether each feature, x_i, in X, is causally related to the outcome Y. Feature x_i is not causally related to the outcome Y if $p(Y|X,T) = P(Y|X \setminus x_i, T)$. We used AIS to compute estimates of these conditional probabilities over 10 bootstraps and perform a paired t-test to determine whether these two values are statistically different. A low p-value suggests that $p(Y|X,T)! = p(Y|X \setminus x_i, T)$ and therefore that x_i is causally related to the outcome. As shown in Figure 4-2 (c), the method identifies features x_0 and x_1 as having p-values below 0.05; i.e., the standard threshold for statistical significance. Moreover, from Figure 4-2 (c) and (d), we can see that though x_3 is highly correlated with the outcome Y, our algorithm was still able to identify the conditional independence between x_3 and Y.

4.4 Experiments on Clinical Data

We applied our algorithm to the Global Registry of Acute Coronary Events (GRACE) [14]. GRACE enrolled over 70,000 patients from 1999-2009 from 250 hospitals in 30 countries. Patients in the registry were diagnosed with an acute coronary syndrome, a constellation of signs and symptoms consistent with reduced blood flow to the heart. Patients were followed and their outcomes and therapeutic interventions were recorded in the dataset. The outcome of interest is mortality within 6 months of admission. We define the treatment effect to be the reduced probability of death from two interventions: Coronary Artery Bypass Grafting (CABG) and Percutaneous Coronary Intervention (PCI), two major interventions for ACS patients. If either CABG or PCI was applied to the patient, the patient would be assigned to the treated group ($T = 1$).

4.4.1 Feature Choices

For the purpose of evaluating the causal feature selection algorithm, we included three types of features. 1) Clear non-confounders, which includes the month of birth and month of hospitalization. 2) 8 clinical features that are believed to be closely related to the outcome (these features are often used in predictive algorithms to quantify patient risk [14] and are thought to be causally related to the outcome) 3) Additional features to explore, including sex and transfer status; i.e., whether the patient was transferred from another hospital. For computational efficiency, we binarized each feature using the median value as a threshold. Missing feature values were estimated using mean imputation.

4.4.2 Results on Clinical Data

The resulting p-values for each clinical feature are shown in Figure 4-3. Our algorithm successfully discovered the two irrelevant features, month of birth and month of hospitalization. Among the 8 GRACE Score features, we were able to identify all of them as confounding factors. The algorithm also generated an above threshold p-value for the feature sex and transfer, considering them neglectable for treatment effect estimation.

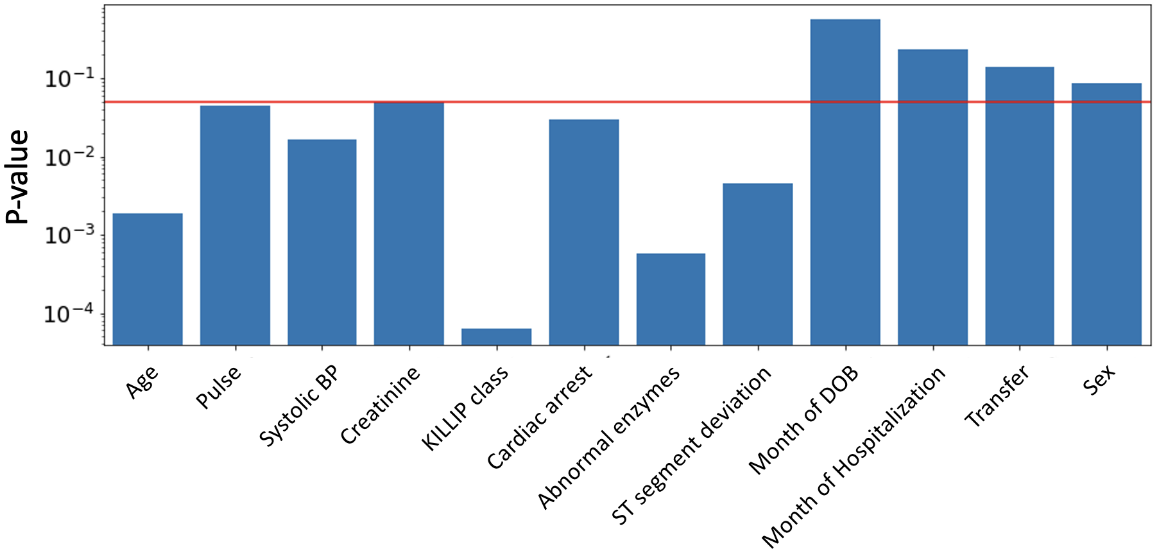


Figure 4-3: Result on GRACE dataset. Features with p-values smaller than 0.05 are selected for treatment effect estimation.

4.4.3 Robustness of ITE with selected feature set

In chapter 3, we show an effective confidence Interval (CI) with regard to common support provides a quantitative way to assess whether the observational data provides enough overlap for robust treatment estimation [58]. As we stated in section 4.1, including irrelevant features may artificially increase the effective CI and make the common support assumption hard to meet. Therefore, selecting a subset of the features can maintain an unbiased estimation of treatment effect, while keeping a tight effective CI. We selected 10 patients whose effective CI are the largest using the original set of features and the new effective CI using the selected features, as

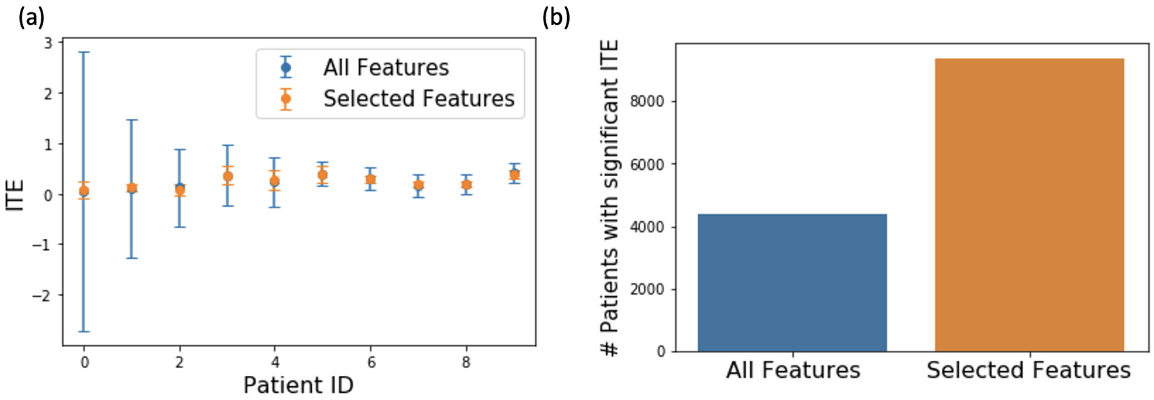


Figure 4-4: Robustness of ITE with selected features. (a) ITE and effective CI of 10 patients with largest effective CI using the original feature set (b) Number of patients with significant ITE using all features vs. selected features

shown in Figure 4-4 (a). Patients with the selected feature set are estimated to have a much tighter interval, making the ITE more robust and reliable. We define an ITE to be significant when its confidence interval does not cross 0, or in another word, we have a $P > 95\%$ that the ITE is either positive or negative. Estimating the treatment effect with the selected feature set dramatically increase the number of patients with significant ITE, as can be seen in Figure 4-4 (b).

4.5 Conclusion

In this Chapter we present a method for addressing the challenge of having irrelevant features in observational datasets, with respect to estimating individual treatment effects. datasets for treatment effect estimation. We used a Contrastive VAE to model the distribution of patients data in different treatment groups and then used these distributions to find a reduced subset that only contains features that are most relevant. As a result, the probability of the outcome conditioned on the full set of patient features is equal to the probability of the outcome conditioned only on the smaller subset containing only relevant features. We use annealed importance sampling to calculate the needed conditional probabilities using the learned distributions from the CVAE. We demonstrated the method on a synthetic dataset and showed the model can successfully exclude non-relevant features in an observational dataset.

By applying the algorithm to a real world dataset, we found the method can exclude features like month of birth and month of hospital admission. We further showed a reduced feature set can make the common support assumption easier to meet and enable more robust estimation of treatment effects.

4.5.1 Limitations

Computational complexity of both simulated annealing and annealed importance sampling are the major limitation of the proposed approach. In order to estimate the density of a given point with a contrastive-VAE, we used annealed importance sampling with intermediate distributions between the target distribution and a known distribution. The required number of intermediate distribution increases with the dimensionality of the data. The sampling cannot be paralleled as the distributions are in a sequential order. Moreover, the annealed importance sampling is needed in every step of simulated annealing, which strongly limits the dimension and complexity of the model in practice. One potentially solution for the computational complexity is to use models like normalizing flows [53] where the probability densities can be analytically expressed. Normalizing flows construct complex distributions by transforming a simple probability density through a series of invertible mappings. If the shared and private information principle can be transferred into the framework of normalizing flow, one can potentially model patients data from different treatment groups accurately while maintaining an efficient way to estimate the point-wise density given the model.

Chapter 5

Identifying Severe Aortic Stenosis with a Single Echocardiogram Video

5.1 Introduction

The accurate diagnosis of Aortic Stenosis (AS) involves both the acquisition of cardiac ultrasound images and the interpretation of these images by skilled personnel [59]. Access to such specialty care, however, may not be possible in many parts of the world, and regular echocardiogram studies can be expensive. Nonetheless, AS is a progressive disease and frequent echocardiographic studies are recommended for patients with suspected disease [60]. A quick and accurate detection method, which minimizes the need for specialized clinical interpretation, would make AS screening more accessible in settings where access to clinical specialists is limited. In this section, we therefore developed and validated a supervised deep learning model that utilizes limited data from cardiac ultrasound to identify patients with severe AS. The model only requires the acquisition of a single view of the heart, and does not require an echocardiologist interpretation for diagnosis.

We begin by introducing some background about the role of echocardiography in the diagnosis of AS. We then briefly review the literature on modeling and analyzing echocardiogram images and videos, as well as studies related to AS. In the subsequent section, we describe the dataset used for this study and the data pre-processing tools we used. We then discuss the deep learning model and training strategies we employed to build a classification model that only uses one echocardiographic view. Lastly, we present the evaluation and analysis of the model we constructed.

5.1.1 Aortic Stenosis

Aortic Stenosis (AS) occurs when the aortic valve narrows and blood flow from the left ventricle to the aorta is blocked or reduced [18]. Figure 5-1 shows a comparison between a healthy valve and a stenotic valve. If left untreated, aortic stenosis can lead to a variety of adverse sequelae including heart failure and death. The 1 year mortality rate for symptomatic patients with severe AS can be as high as 50% [61].

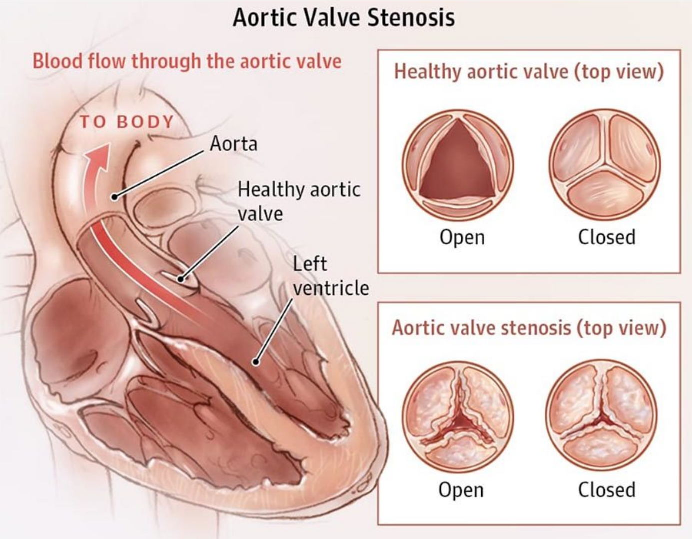

Aortic Valve Stenosis
Blood flow through the aortic valve
TO BODY
Aorta
Healthy aortic valve
Left ventricle
Healthy aortic valve (top view)
Open
Closed
Aortic valve stenosis (top view)
Open
Closed

Figure 5-1: Anatomy of the heart (left) and comparison of a normal aortic valve versus aortic stenosis (right) [1].

Clinically, the severity of AS ranges from mild to severe, as shown in Figure 5-2. It is usually assessed by measuring the maximum trans-aortic velocity, mean pressure gradient and valve area. Severe AS is defined by a peak trans-valvular velocity of > 4 m/s, a mean gradient of > 40 $mmHG$ or a valve area < 1 cm^2. Measurement of these values is typically made using echocardiography, which we briefly review in the next section.

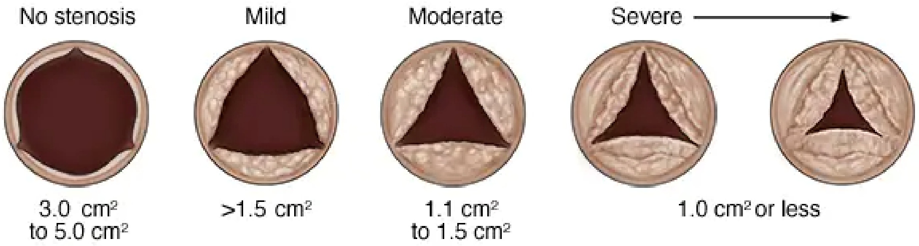


Figure 5-2: Severity of AS from normal to severe. [2]

5.1.2 Echocardiography

Echocardiography is the primary method used to diagnose Aortic Stenosis. During an echocardiographic study (also known as en echo study), ultrasound from a hand-held transducer is used to image the heart. Patients with AS typically undergo yearly echocardiographic studies to assess the rate of progression [62]. A full echo study contains several different views of the heart, where sonographers acquire images at different positions and/or angles on the chest. Common views include Parasternal Long Axis (PLAX), Parasternal Short Axis (PSAX), 4 chamber views and so on. Figure 5-3 shows a diagram of clinicians acquiring echo images with a transducer as well as example of 4 different views of echocariodgram.

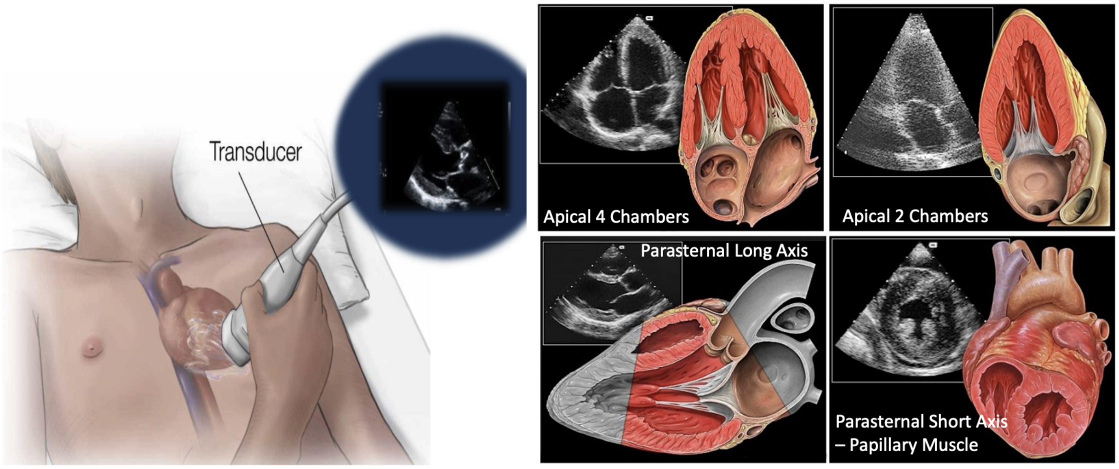


Figure 5-3: Echocardiogram study (left) and example of 4 different views of echocardiogram. [3]

In a full echo study, typically 100 videos and images are obtained. Figure 5-4 gives an example of all files contained in a study.

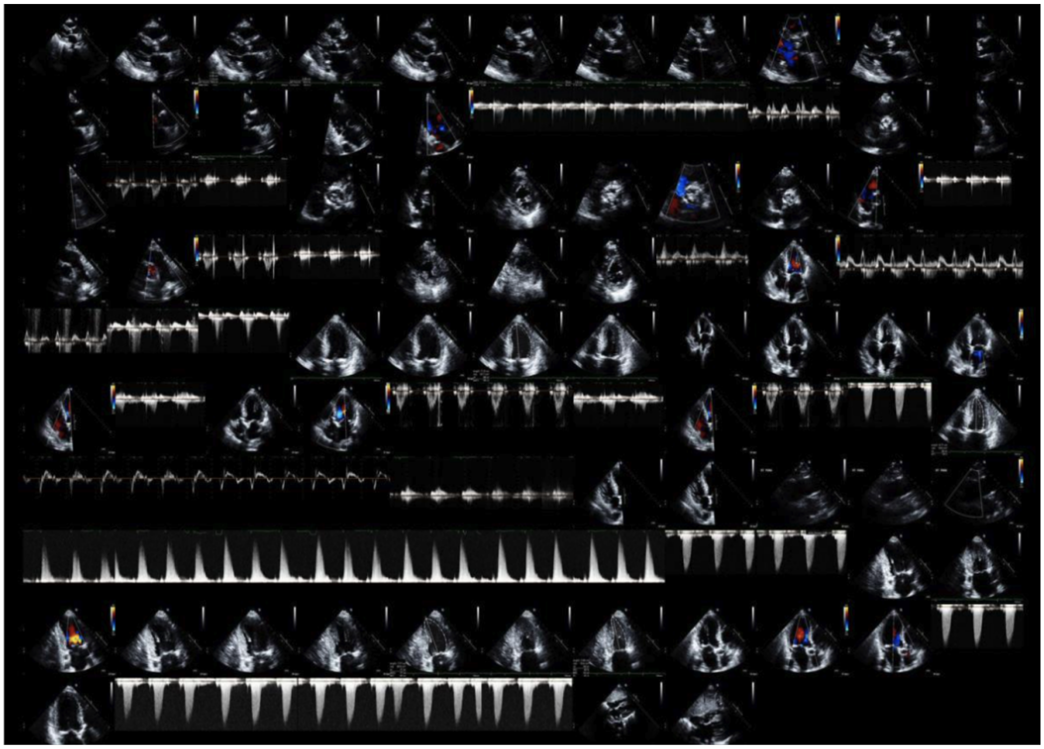


Figure 5-4: A full echo study contains more than 100 videos and images

5.1.3 Deep Learning with Echocardiographic Images

Deep learning models have been widely studied and used in echo image/video modeling and analysis. Models have been built for fundamental tasks like view classification and segmentation and various studies have also leveraged echocardiographic images for diagnosis. For view classification, convolutional neural networks have been reported to achieve good performance on a wide range of different views [63] [64]. Both CNN [65] and U-net [66] have been applied to echo image segmentation for 4 chamber views to identify cardiac chambers. These tools have been extended for various disease detection methods including predicting ejection fraction [6], post operative ventricle failure after heart transplant [67], and identify areas of the heart that may be damaged after myocardial injury [68]. Most of these studies use a supervised approach to build models on images extracted from echo videos. EchoNet [6] was the first reported model that also use the temporal information within an echo video, where a 2+1 D convolution was applied to continuous frames of echo videos to predict heart failure with reduced ejection fraction. Built upon EchoNet, models have been developed with transfer learning techniques and additional ways of feature extraction like optical flow has been reported to improve the performance [67].

In the case of Aortic Stenosis, Huang et el, reported a semi-supervised model with MixMatch, named TMED, to classify patients as having no AS, mild or severe AS [69]. In this model, they first classified each view of an echo image into PLAX, PSAX and other and use a weighted combination of all these 3 categories to predict the severity of the disease. They developed and evaluated their models on a relatively small datasets (260) patients but achieved a 90% accuracy in the classification task. Nevertheless, TMED still needs all views from a full study to classify the severity.

5.2 Methods

5.2.1 Dataset

We queried the MGH database for all echocardiogram studies performed in the past 20 years. Among those, we selected studies where the mean gradient or aortic valve area is reported by Level III trained echocardiographers. In the data cleaning process, we removed studies with unrealistic values for patient demographics, mean gradient (< 0 or $> 200mmHG$) and valve area (< 0 or $> 5cm^2$). After cleaning, we obtained a total number of 28,734 studies from 16,066 patients. We extracted all PLAX videos from these studies and divide the videos into training (60%), validation (20%) and held-out test set (20%), with a constraint that all videos from the same patient would be assigned to only one of the three sets. Among the collected PLAX videos, all of them were labeled with mean gradient and 53% were labeled with valve area. A summary statistics of the curated data set for patients labeled with mean gradient is shown in Table 5.1. Table 5.2 shows the same statistics for patients labeled with valve area. The valve area dataset is a subset of the mean gradient set, as only half of the patients are labeled with the actual value of valve area in their echo report. In general, these patients are sicker than the larger population, as can be seen in statistics about age and ejection fraction. The reason for this difference is that the valve area is more likely to be measured in sicker patients.

Statistics	**Total**	**Training**	**Validation**	**Test**
Number of Patients	16,066	9,639 (60%)	3,213 (20%)	3,214 (20%)
Number of Studies	28,734	17,139 (60%)	5,804 (20%)	5,791 (20%)
Number of PLAX Videos	109,971	98,376	5,804	5,791
Age (STD)	76 (12)	76 (12)	76 (12)	76 (12)
Female %	42 %	42 %	43 %	44 %
Height *cm* (STD)	168 (12)	168 (12)	167 (12)	167 (12)
Weight *kg* (STD)	79 (20)	79 (20)	80 (20)	79 (20)
Ejection Fraction (STD)	62 (14)	62 (14)	62 (14)	61 (14)
Mean Gradient *mmHG* (STD)	23.38 (15.41)	23.39 (15.41)	23.37 (15.49)	23.15 (15.43)

Table 5.1: Summary of patients characteristics of curated dataset

Statistics	**Total**	**Training**	**Validation**	**Test**
Number of Patients	8,749	5,228 (60%)	1,721 (20%)	1,800 (20%)
Number of Studies	15,014	8,939 (60%)	2,983 (20%)	3,075 (20%)
Number of PLAX Videos	57,999	51,941	2,983	3,075
Age (STD)	78 (10)	78 (10)	78 (10)	78 (10)
Female %	40 %	40 %	40 %	41 %
Height *cm* (STD)	168 (13)	168 (13)	168 (11)	167 (12)
Weight *kg* (STD)	79 (20)	79 (20)	80 (20)	78 (21)
Ejection Fraction (STD)	60 (15)	60 (15)	60 (15)	60 (15)
Valve Area cm^2 (STD)	0.98 (0.34)	0.98 (0.33)	0.98 (0.34)	0.99 (0.36)

Table 5.2: Valve Area Table

5.2.2 De-identification

A frame of an echocardiogram video contains several components, as can be seen in the example shown in Figure 5-5. The main ultrasound image is contained in a fan shape region, highlighted by the red dotted line in the example. The ECG signal, which is acquired simultaneously sits in the bottom part of the frame, highlighted by yellow dotted lines in the example. Patients name, age and other protected health information (PHI) are shown at the top region of the frame (we blurred the PHI in the example figure).

The goal of de-identification is to remove PHI and other metadata in the frame and isolate the ultrasound image and the ECG signal. We designed and implemented

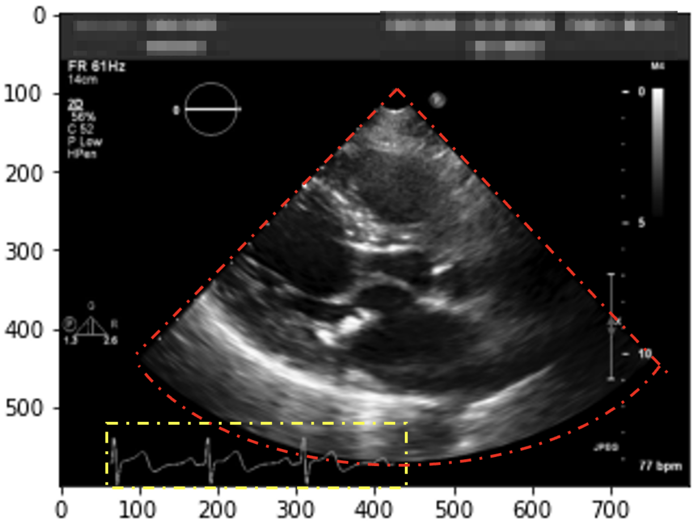


Figure 5-5: Example frame of echo cardiogram. Red and yellow dotted lines highlight the region contains the ultrasound image and the ECG signal.

a heuristic algorithm to extract the ultrasound region, which is schematically shown in Figure 5-6.

Each echo frame is a grey scale matrix with values ranging from 0 to 255. In the first step, we sum up all frames in a video so that any pixels greater than 0 in any frame will be captured. This gives us a large bright fan shape region and other small areas of pixels containing the meta data. In the second step, we apply an average kernel to the frames so that the pixels containing the metadata will be combined with the black background, resulting in a blurred image as shown in the figure. Then the metadata pixels can be removed by adding a threshold as shown in step 3. In step 4, we further remove the top white bar which corresponds to the area that contains patient protected health information. Finally, we apply the black and white image we obtained in step 4 as a mask to the original frame. The resulting de-identified echo frame is shown in step 5.

5.2.3 Extracting PLAX Videos

We applied a supervised classification model that was developed and published by Rahual et, al.[63] to identify PLAX videos within full echo studies. The model classifies an echo video into 1 of 23 different views with a Softmax probability output. We

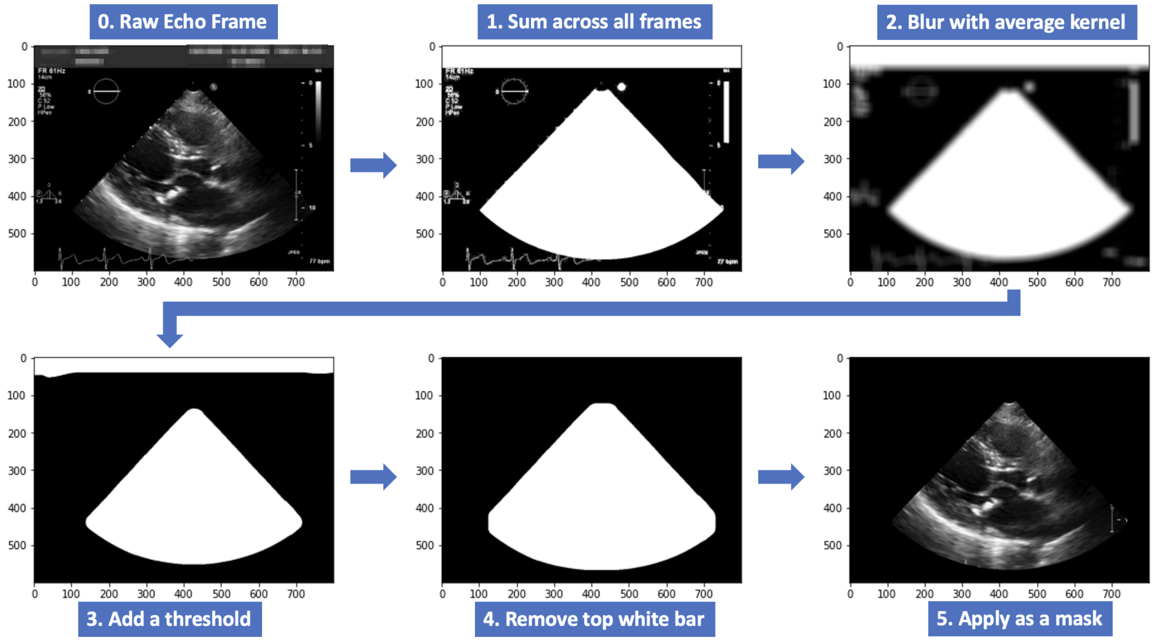


Figure 5-6: Step by step example of the de-identification process.

selected videos with $P(PLAX) > 0.5$ as the PLAX videos. To evaluate the model on our dataset, we manually labeled 5 studies with the help from an expert cardiologist. The precision of the PLAX classification is 0.9

5.2.4 Data Augmentation

We applied two types of data augmentation to improve generalization of supervised learning models for the downstream task. The first type is augmentation along frames. In each training epoch, we selected 20 random continuous frames from the video as input to the model. This can potentially prevent the model from overfitting to a particular frame in the video. The second type of augmentation is augmentation within a frame. Frames are randomly resized and cropped so that the zoom level and position of the heart in the videos would not be used as spurious features by the model. Concretely, we first resize each frame with a random ratio between 1.1 to 1.5. We then randomly cropped the enlarged images to the original size. An example of such augmentation is shown in Figure 5-7.

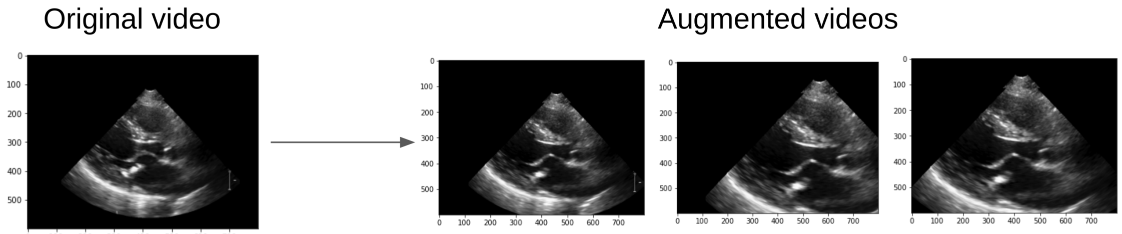


Figure 5-7: Augmentation of echo frames

5.2.5 Model Training and Evaluation

We implemented our model based on ResNet18 [70] with spatial temporal convolutions. The spatial temporal convolution consists of a 2 dimensional convolution across all pixels within each frame and a 1 dimensional convolution across the frames along the time axis. The network consists of 18 blocks and a linear layer with Sigmoid activation.

As more patients are labeled with mean gradient, we first trained the model to predict whether the gradient it greater than 40 mmHG, noted as $M_{pressure}$. Then we used the learned weights in $M_{pressure}$ as the initial weight for the training of valve area model, noted as M_{area}. Training of both $M_{pressure}$ and M_{area} models consists of two steps. In the first step, we used all videos with $P(PLAX) > 0.5$ for training. In the second step, we fine tune the model using only 1 PLAX video from each study by selecting the video with highest $P(PLAX)$. The same criteria is used to select one PLAX video from each study for the validation and test set, as shown in the schematic Figure 5-8 . During training, we computed model AUC on the validation set for every epoch and selected the best performing model according to the validation AUC. We

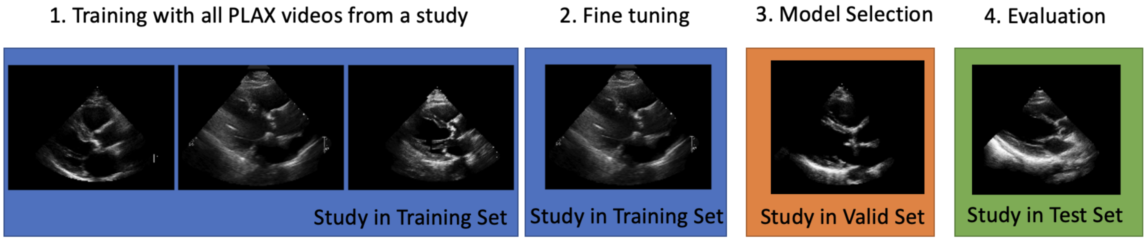


Figure 5-8: Training, validation and evaluation steps. We first train the model with all PLAX videos from a study, and then fine tune with videos of highest $P(PLAX)$

then evaluated the model on the held out test set using AUC and accuracy. To compute the accuracy, we used a threshold of 0.5. For both models, we performed 5 bootstraps with random splitting of training, validation and test sets. Both outcomes are stratified in the splitting.

Models are implemented in Python with Pytorch [71]. Training and evaluation are conducted on 2 Nvidia Tesla V100 GPUs, each with 32GB of memory.

5.2.6 Statistical Analysis

We further evaluated the models by computing the specificity and sensitivity at different thresholds. Using the same thresholds, we computed the positive predictive values (PPV) and the negative predictive values (NPV) for the models at different prevalence level.

We follow the computation method proposed by [72], where the formulas for PPV and NPV are

$$PPV = \frac{p \cdot se}{p \cdot se + (1 - p) \cdot (1 - sp)} \tag{5.1}$$

$$NPV = \frac{(1 - p) \cdot sp}{p \cdot (1 - se) + (1 - p) \cdot sp} \tag{5.2}$$

where p is the disease prevalence, se is the sensitivity and sp is the specificity.

5.3 Results

5.3.1 Classification Performance

Table 5.3 shows the result of the classification performance for the two tasks. We achieved decent AUC for both mean gradient and valve area classification. In general, the mean gradient model performs better than the valve area model. The main reason of that is the cohort to evaluate the valve area consists of more severe patients.

With a 0.5 cut-off, we computed the accuracy of these two models. For comparison, the accuracy reported in TMED [69] is 90% using all views from a study. The

Model	**Test Size**	**AUC**	**Accuracy**
$M_{pressure}$	5,792	0.88 ± 0.01	0.88 ± 0.004
M_{area}	2,976	0.78 ± 0.01	0.70 ± 0.007

Table 5.3: Test set size and classification performance of the two models. Numbers shown are mean and std from 5 random bootstraps.

sensitivity of $M_{pressure}$ is 0.42 ± 0.03 and the sensitivity of M_{area} is 0.71 ± 0.05.

5.3.2 Statistical Analysis

The specificity-sensitivity curve in Figure 5-9 and 5-10. PPV and NPV at different prevalence levels are alson shown in the same figures. Both the mean gradient and the valve area model show the ability to keep NPV relatively high regardless of the prevalence level. To gauge how this model would perform in the general population, we note that the prevalence of severe aortic stenosis is approximately 3% in patients over 75 years old in the United States [73]. At this prevalence level, the NPV of the mean gradient model is more than 99% with a 80% sensitivity. The high NPV value suggests the model could be used as an efficient screening tool for patients over 75 years old. On the other hand, the model could also be used to identify severe AS patients in an AS cohort. In the MGH cohort we collected with valve area labeled, 50% of the patients are considered in a severe condition. At this prevalence, the PPV of the mean gradient model is above 80% at 80% sensitivity.

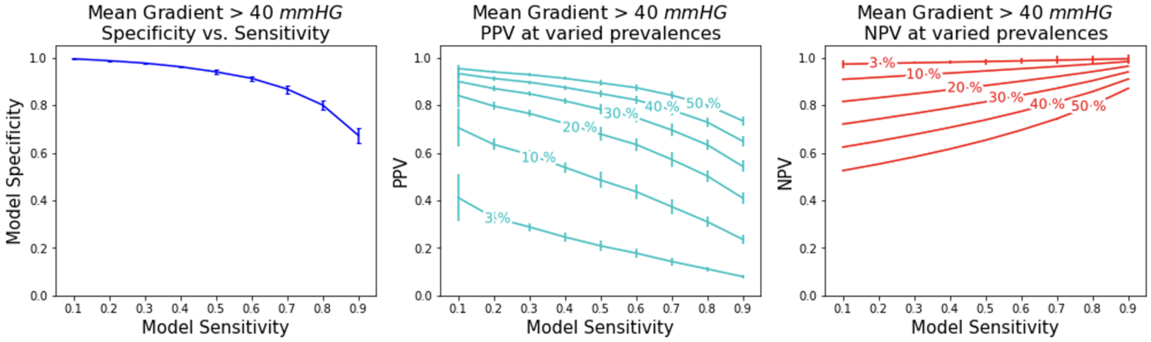


Figure 5-9: Specificity vs. sensitivity; PPV and NPV at different prevalence levels for the mean gradient model

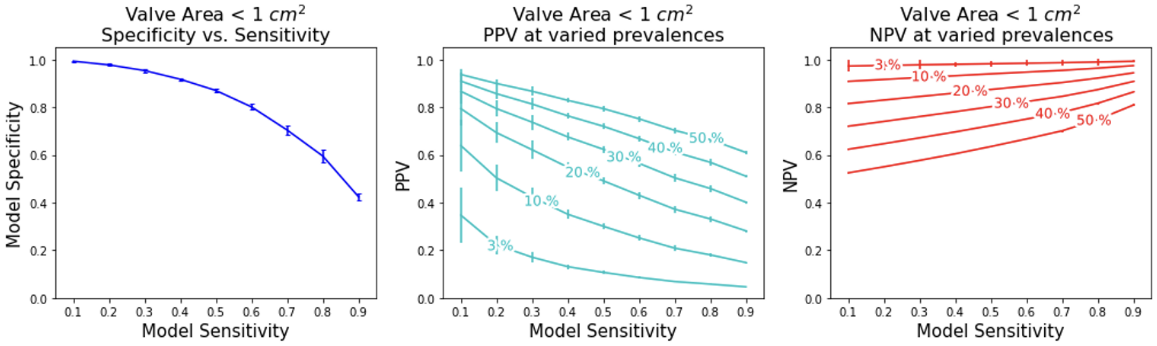


Figure 5-10: Specificity vs. sensitivity; PPV and NPV at different prevalence levels for the valve area model

5.4 Discussions

5.4.1 Model Calibration

The calibration curves of the two models are shown in Figure 5-11. As the model has a Sigmoid activation in the last layer and cross entropy loss, it's well calibrated after training. The R^2 scores for the two models are 0.95 and 0.93.

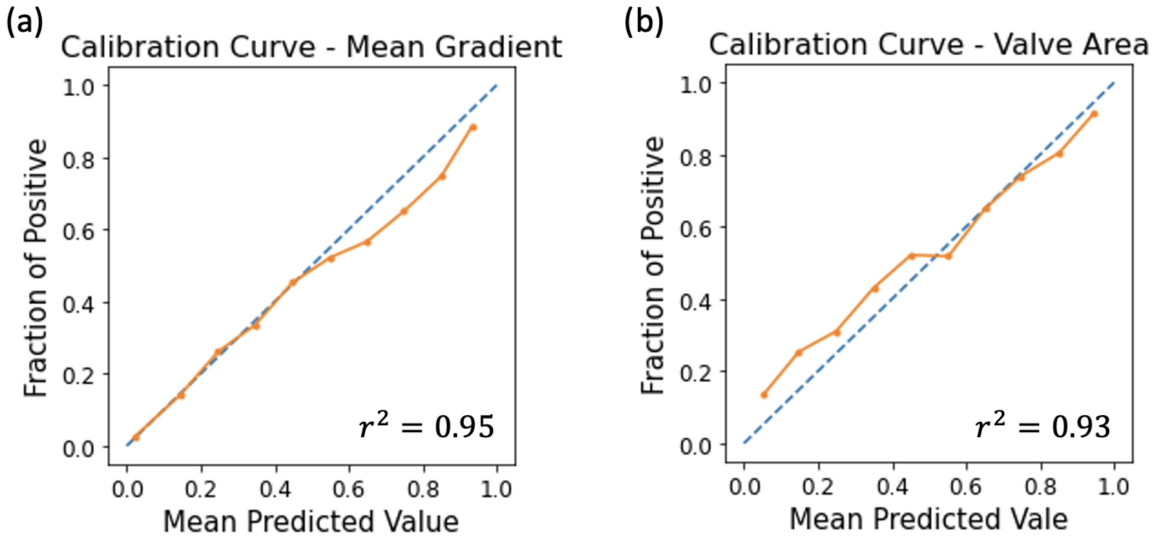


Figure 5-11

5.4.2 Model Interpretation

Deep learning models are often seen as black box and interpretability of such models are critical for the model to be trusted and applied in real clinical care. Saliency Map Analysis is an illustrative way to reveal which region the models focus on to make the prediction [74]. Examples of such analysis for frames in three different studies are shown in Figure 5-12. As can be seen from the figure, pixels around the aortic valve are clearly highlighted compared to the background in both models. This demonstrates deep learning models align with human expertise to focus on regions highly related to the valve disease itself and therefore the information extracted from the pixels can help to make more reasonable decisions about the two outcomes.

We also looked at the saliency value across different frames to see if any frame plays a more significant role in the inference process. By summing up the saliency value within a frame, we show the comparison in Figure 5-13. A corresponding ECG signal of these frames is shown along side. We find the frames with the largest value of saliency corresponds to the ejection period of the cardiac cycle, a period where the aortic valve is open and the left ventricle ejects blood to the aorta through the valve. This further suggests the trained model is able to leverage the useful information in the pixel data without any prior knowledge.

5.4.3 Regressing Mean Gradient and Valve Area

By changing the loss function from binary cross entropy to mean squared loss and feeding the true value of mean gradient and valve area, we trained two models to regress the value of these two targets. We adopted the same training strategies as in section 5.2.5, except for the bootstraps. We evaluated the regression results with metrics including root mean squared error (RMSE), R^2 score and Pearson Correlation, as can be seen in Table 5.4. Figure 5-14 further shows the scatter plot of prediction vs. ground truth for samples in the test set. Results suggest the regression model does not predict the values of mean gradient and valve area very well. Part of the reason can be the label of the two values can be noisy themselves, as accurately measuring

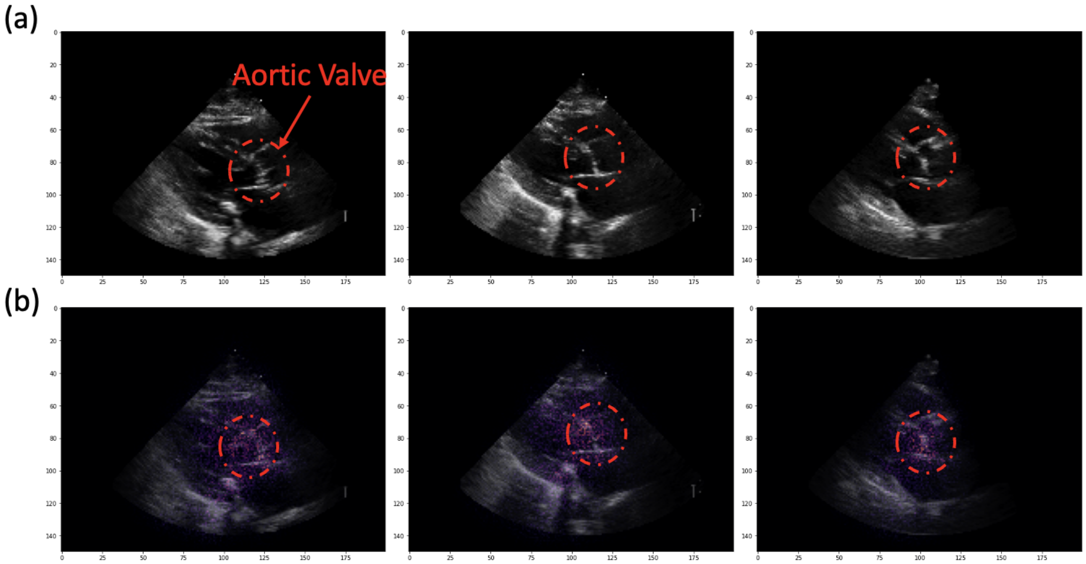


Figure 5-12: (a) Examples of Echo frames and (b) the corresponding saliency maps of the frames.

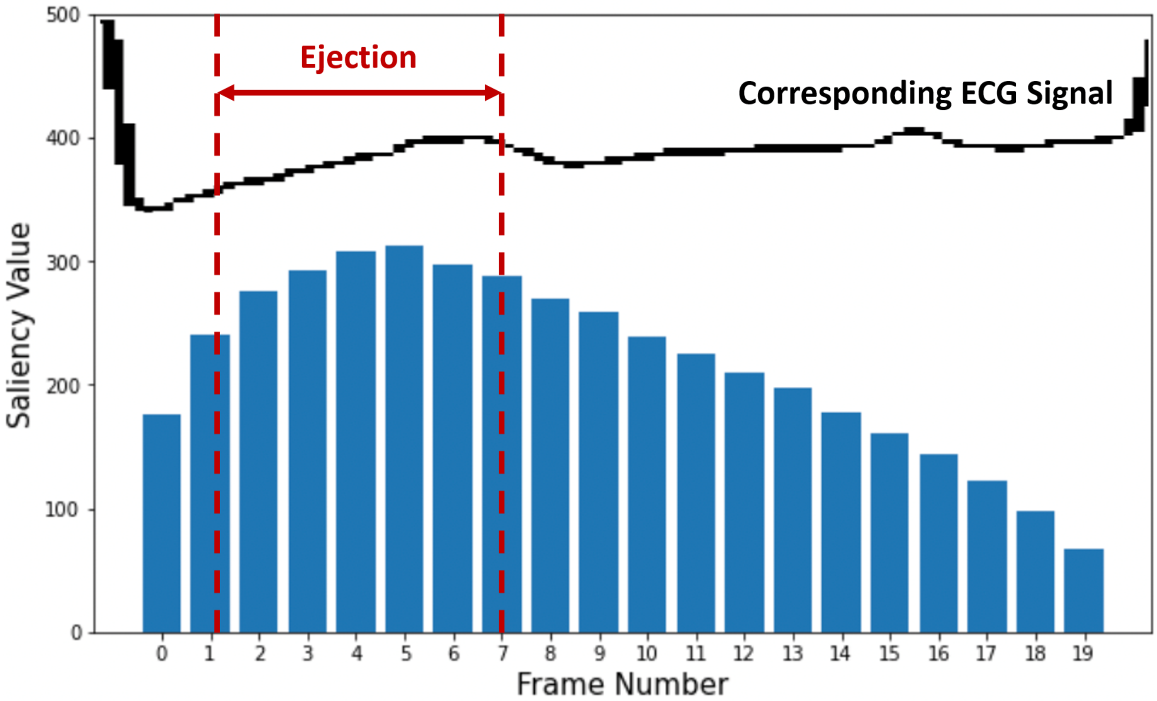


Figure 5-13: Comparison of saliency values across different frames and the Lead I ECG signal corresponding the these frames.

Metric	Mean Gradient Model	Valve Area Model
RMSE	12	0.08
R^2	0.43	0.38
Pearson Corr	0.65	0.63

Table 5.4: Regression Performance of predicting mean gradient and valve area

them is not an easy tasks for cardiologists.

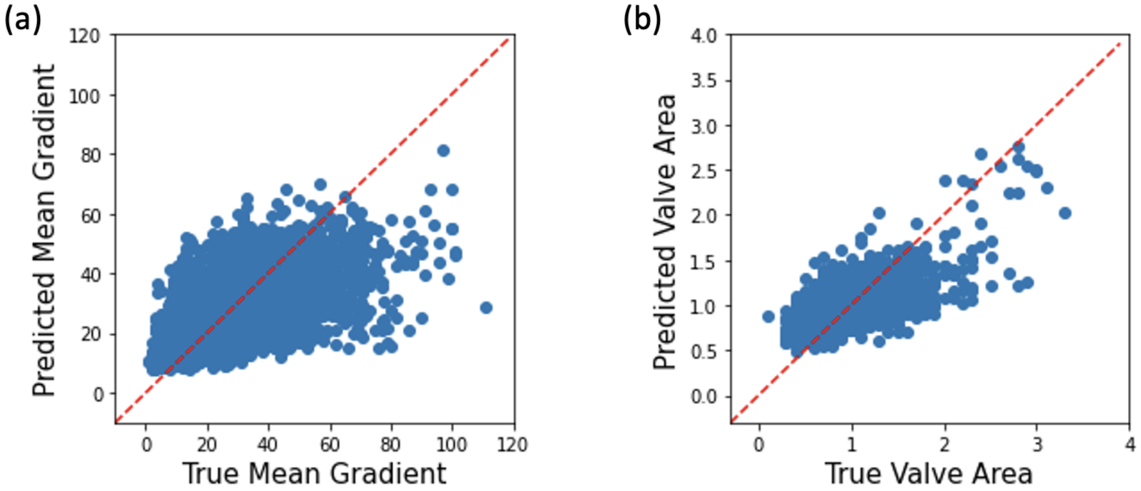


Figure 5-14: Predictions vs. true values of the regression model for mean gradient and valve area.

5.5 Conclusions

In this section, we present a deep learning model to identify severe AS patients using only a single PLAX video. We collected and curated a large echocardiogram dataset of $28,734$ studies from $16,066$ patients for model developing and evaluation. We developed computational tools to identify PLAX videos, de-identify the frames, isolate and extract the echo image and ECG signals within each frame. We show deep learning models can classify whether mean gradient is above $40mmHG$ and whether valve area is smaller than $1cm^2$. with a decent AUC. We also show the model gives high NPV and PPV at different prevalence levels of the disease. This indicates the usefulness of the model in different clinical settings. Finally Saliency maps analysis show the model focuses on the aortic valve area within each frame, and the ejection period of a cardiac cycle. This interpretation proves the model making decisions with reasonable and reliable mechanism to extract information from the pixel data.

5.5.1 Clinical Implications

A quick screening tool of severe aortic stenosis can reduce the cost of full echo studies for AS patients each year. Within an echo study, the model can classify level of mean gradient and valve area using the first PLAX video acquired by the echo cardiologist. This information can guide the clinical to perform further measurement and analysis for more severe patients.

5.5.2 Limitations

In this section of the thesis, we showed that deep learning models can classify severe AS with a descent AUC, and can be reasonably explained by Saliency Analysis. Nevertheless, one major limitation of the work is that the model with a regression loss cannot accurately predict the value of the mean gradient and the valve area. In many cases, AS patients need to have their gradient and valve area measurements recorded in order to track the progression of the disease. Further model training and optimization for the regression task is needed to reduce the cost of expert labeling of those values for patients record.

Another limitation of the work is that the extraction of PLAX videos from a full study is dependent on a public available view classification model. The model is not fine tuned on the AS cohort we used to develop and evaluate severe AS identification. The discrepancy of patient cohort, institution, as well as echocardiogram devices may negatively affect the classification performance and thus generate fake PLAX videos in the curated PLAX dataset. Training a tailored view classification model with expert labeled data on the AS cohort, or learning an unsupervised model to cluster different views, could potentially improve the view classification results. Moreover, the tailored view classification model can identify other important views like Parasternal Short Axis (PSAX), another commonly used view for AS diagnosis [75]. Using additional views can potentially improve the model's performance in identifying severe AS patients.

Another direction for future work is modeling the progression of AS, as understanding the progression of AS can dramatically help the management of the disease [76]. One interesting and direction to explore is to identify rapid progressive AS patients [77] using Echocardiogram videos. The diagnosis models can be used as pretraining tasks and provides model weights as initial values for the prognostic models. By fine tuning the model with patients who have multiple studies, the prognostic models can provide more insights into the progression of Aortic Stenosis.

Chapter 6

Summary and Future Work

In this chapter, we summarize the findings of the thesis and outline possible directions for future work.

6.1 Summary of Findings

6.1.1 Generative Oversampling with a Contrastive VAE

In Chapter 2 we described the challenge of class imbalance in the context of prognosis for cardiovascular diseases. We proposed an oversampling technique for extreme class imbalance problems with a contrastive VAE. The model leverages the shared and private information in the majority and minority groups by explicitly dividing the latent space into shared and private sub-spaces. We demonstrated the model's ability to capture the shared variance in a synthetic dataset in section 2.4 and compared the method with other oversampling techniques. We also applied and compared them with a real world dataset, GRACE, and found the contrastive VAE outperformed other techniques significantly for predicting extremely rare events. These experiments show the generative model can make use of the rich information shared between groups. Prior domain specific information about the underlying distributions of the two groups can therefore help to decide when this oversampling method is most applicable.

6.1.2 Quantifying Common Support for Multiple Treatments with a Contrastive VAE

In Chapter 3, we discussed the challenge of assessing the common support assumption in observational datasets for robust estimation of treatment effect. We extended the contrastive VAE model described in Chapter 2 to multiple treatment groups where each group has a private latent variable in the generative process. With a synthetic dataset, we showed the contrastive VAE can estimate and recover the overlap between groups in cases where samples are unbalanced across groups. We further proposed an effective confidence interval to illustratively quantify the common support assumption and demonstrated its implications in cases with different level of overlap. Finally, we showed the impact of the effective confidence interval on real world data where we analyzed the effectiveness of CABG on ACS patients. By considering the criteria of common support assumption, we find CABG is more effective to younger patients ($p < 0.05$). However, a similar analysis that uses all the data (i.e., not considering the common support) suggests that CABG does not derive a benefit irrespective of age ($p > 0.05$).

6.1.3 Feature Selection for Treatment Effect Estimation

In Chapter 4 we described the challenge of having irrelevant features in datasets for treatment effect estimation. We used a Contrastive VAE to model the distribution of patients data in different treatment groups and then used these distributions to find a reduced subset that only contains features that are most relevant. As a result, the probability of the outcome conditioned on the full set of patient features is equal to the probability of the outcome conditioned only on the smaller subset containing only relevant features. We use annealed importance sampling to calculate the needed conditional probabilities using the learned distributions from the CVAE. We demonstrated the method on a synthetic dataset and showed the model can successfully exclude non-relevant features in an observational dataset. By applying the algorithm to a real world dataset, we found the method can exclude features like month of birth and month of hospital admission. We further showed a reduced feature set can make the common support assumption easier to meet and enable more robust estimation of treatment effects.

6.1.4 Identifying Severe Aortic Stenosis with a Single Echocardiogram Video

In Chapter 5, we discussed the challenge of diagnosing severe Aortic Stenosis in settings where specialists are not available. We developed a deep learning model to classify severe AS by using only a single PLAX video from a study. The model achieved an AUC of 0.88 for classifying whether the mean gradient is above 40 $mmHG$ and an AUC of 0.78 for classifying whether the valve area is below 1 cm^2. In addition, we conducted a statistical analysis to evaluate the PPV and NPV of the model at different prevalence levels. For the general population where the disease prevalence is 3%, the NPV of the model to identify high mean gradient patients is above 99% at 80% sensitivity. For an AS cohort where the disease prevalence is 50%, the PPV of the model is 80% at 80% sensitivity. The results show the model can be used both as a fast screening of severe Aortic Stenosis for the general population and a tool to accurately identify severe patients in an AS cohorts. Besides the model, we also developed computational tools to automatically extract PLAX videos, de-identify data and isolate ultrasound and ECG signals from the raw echocardiogram videos.

6.2 Directions for Future Work

6.2.1 Limitations of Contrastive VAE

In the thesis, we demonstrate the shared information assumption could be a valuable prior knowledge when modeling data with different classes or different treatment groups. We implemented the model with a Variational Autoencoder and applied it to oversampling and treatment effect estimation. One limitation of our method is that we only implemented the model for tabular data. Extending the assumption to different kinds of data such as images and time series can potentially enable more applications. For example, image augmentation is widely used in supervised and self-supervised learning [78] [79] [80]. One direction of future work can be introducing convolutional layers into the generative model, or modifying other generative models, such as Generative Adversarial Networks [81], to incorporate the assumption of shared and private latent variables.

Another shortcoming of Contrastive VAE is the computational complexity to estimate the probability densities in the observational space. In chapter 3 and 4, we used contrastive VAE to model patients in different treatment groups and estimate the probability distribution of different groups with annealed importance sampling. Though we show the distribution densities can be accurately estimated, the computational complexity of annealed importance sampling is relatively high as samples need to be generated sequentially for the intermediate distributions. One potential solution to reduce the complexity is to model the distribution with normalizing flows [53] where the probability densities can be analytically expressed. Normalizing flows construct complex distribution by transforming a simple probability density through a series of invertible mappings. A normalizing flow with shared and private sectors in the latent density can potentially be a more efficient way to model and estimate probability densities across different treatment groups.

6.2.2 Deep Learning with Echocardiogram Videos

One shortcoming of the model we described in Chapter 5 is that it depends on a view classification model that is not tailored to our dataset. Though the precision for PLAX is acceptable, we inevitably included a small portion of videos that are not PLAX. Moreover, the view classification model for other views perform poorly. For example, the precision for PSAX view is only 60% with the data we collected. This limits us from using different views to improve the model or develop new models for other tasks. There are two potential solutions. We can created our dataset for view classification with expert labeling different views for the videos. The other way is to train an unsupervised or self supervised learning algorithm and cluster the videos using a low dimensional representation. Prior works have demonstrated the success of the self supervised representation with ECGs [82], CT images [83] and MRIs [84]. A general representation learning of Echocardiogram videos can potentially be useful for not only the view classification problem, but also for various downstream tasks.

Another direction for future work is modeling the prognosis of AS. In Chapter 5, we demonstrated deep learning can help the diagnosis of AS. Prognosis of AS, on the other hand, are much more challenging but also more important as understanding the progression of AS can dramatically help the management of the disease [76]. One interesting and important direction to explore is to identify rapid progressive AS patients [77] using Echocardiogram videos. The diagnosis models can be used as pre-training tasks and provides model weights as initial values for the prognostic models. By fine tuning the model with patients who have multiple studies, the prognostic models can provide more insights into the progression of Aortic Stenosis.

Chapter 7

Supplementary Information

7.1 Data Sets Information

7.1.1 GRACE Dataset

A detailed description and usage guide of GRACE data set is included in

`https://github.com/mit-ccrg/grace`

7.1.2 Echocardiogram Dataset

Description of echo data and access is available at

`https://github.com/mit-ccrg/echo_models/blob/master/data.md`

7.2 Implementation Details

7.2.1 Contrastive-VAE

Implementation of the Contrastive VAEs are available at

`https://github.com/mit-ccrg/contrastive-variational-autoencoder`

Implementation of Annealed Importance Sampling and Simulated Annealing is also included in this repo.

7.2.2 Echocardiogram Pre-processing Tools

https://github.com/mit-ccrg/echo-preprocess

Except for the pre-process tools we mentioned in Chapter 5, we also implemented an ECG extraction and segmentation tool. A standard Lead I ECG signal is usually obtained during an echo study and is recorded simultaneously in the echo videos. The ECG signal can be used to identify the cardiac circle of the echo, as segmentation of ECGs are much easier than the echo video along the time axis. Segmentation of the ECG can be applied to the echo video and enable alignment of the video, or extracting frames corresponds to certain point in the cardiac circle. Moreover, the ECG signal itself can be an additional input for downstream tasks. It can also be used for multi-modal learning between ECGs and Echos.

As part of the pre-processing tools, we implemented an algorithm to extract and segment the ECG signals. An example of this algorithm can be seen in Figure 7-1. We start from the last frame of an Echo video as the entire signal is kept in this frame. Then the box containing the ECG is identified by matching the color and position of the signal. In the next step, the pixels are converted to pseudo voltage signal according to their position in the box. We then interpolate the pseudo voltage signal as shown in step 4. R peaks of the interpolated signal can be detected with ECG segmentation tools. We implemented our algorithm using the tools provided in the open source package Neurokit [85].

7.2.3 Echocardiogram Models

Implementation of deep learning models, training and testing scripts can be find in this repo.

https://github.com/mit-ccrg/echo_models

And a detailed description of the implementation can be found here

https://github.com/mit-ccrg/echo_models/blob/master/PlaxNet.md

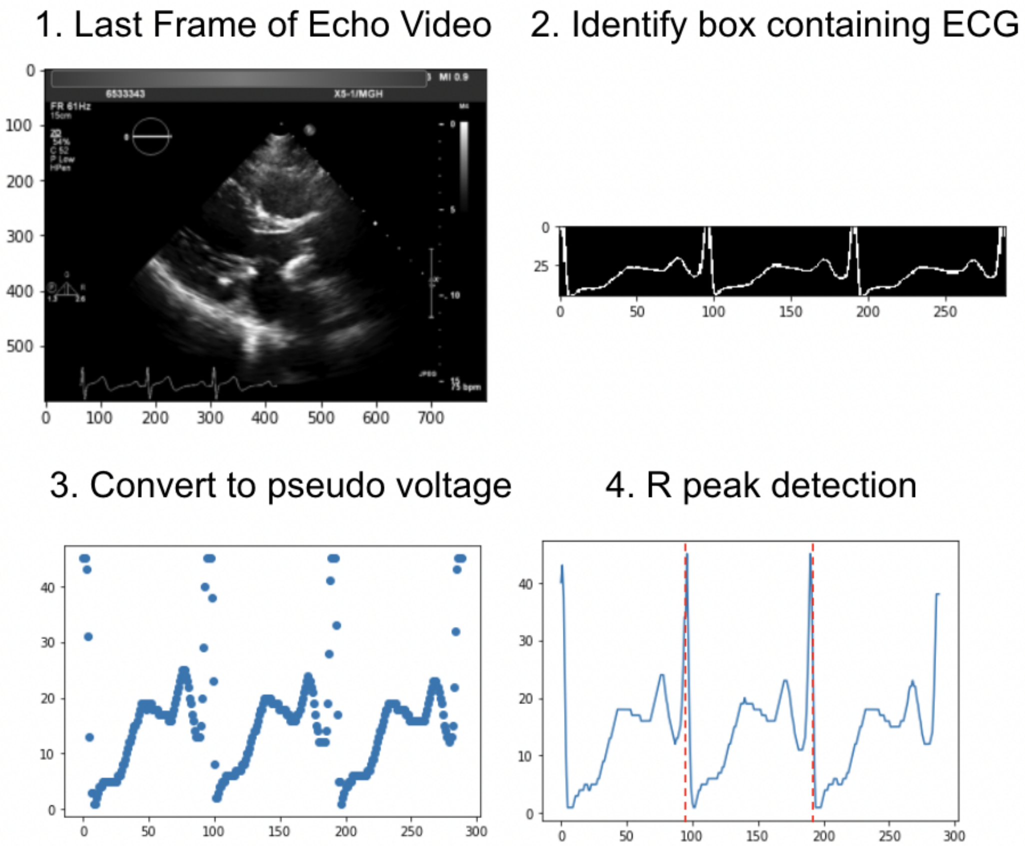


Figure 7-1: Steps to extract and segment ECG signals in Echo video

Bibliography

[1] The Heart and Vascular Centre. What is aortic stenosis. *https://www.heartvascularcentre.com/structural-heart-and-heart-valve-procedures/what-is-aortic-stenosis*, Dec 2018.

[2] Mayo Clinic. Aortic valve stenosis. *https://www.mayoclinic.org/diseases-conditions/aortic-stenosis/symptoms-causes/syc-20353139*, Feb 2021.

[3] SONOSIF. Tte: Transthoracic echocardiogram. *https://sonosif.com/clinical-apps/tte-transthoracic-echocardiogram/*, Feb 2021.

[4] Paul D Myers, Benjamin M Scirica, and Collin M Stultz. Machine learning improves risk stratification after acute coronary syndrome. *Scientific reports*, 7(1):1–12, 2017.

[5] Alec Vahanian and Catherine M Otto. Risk stratification of patients with aortic stenosis. *European heart journal*, 31(4):416–423, 2010.

[6] David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P Langlotz, Paul A Heidenreich, Robert A Harrington, David H Liang, Euan A Ashley, et al. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020.

[7] Joon-myoung Kwon, Kyung-Hee Kim, Ki-Hyun Jeon, and Jinsik Park. Deep learning for predicting in-hospital mortality among heart disease patients based on echocardiography. *Echocardiography*, 36(2):213–218, 2019.

[8] K Shailaja, B Seetharamulu, and MA Jabbar. Machine learning in healthcare: A review. In *2018 Second international conference on electronics, communication and aerospace technology (ICECA)*, pages 910–914. IEEE, 2018.

[9] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[10] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[11] Jennifer Hill and Yu-Sung Su. Assessing lack of common support in causal inference using bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *The Annals of Applied Statistics*, pages 1386–1420, 2013.

[12] Kristjan Greenewald, Karthikeyan Shanmugam, and Dmitriy Katz. High-dimensional feature selection for sample efficient treatment effect estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 2224–2232. PMLR, 2021.

[13] Elliott M Antman, Marc Cohen, Peter JLM Bernink, Carolyn H McCabe, Thomas Horacek, Gary Papuchis, Branco Mautner, Ramon Corbalan, David Radley, and Eugene Braunwald. The timi risk score for unstable angina/non–st elevation mi: a method for prognostication and therapeutic decision making. *Jama*, 284(7):835–842, 2000.

[14] Eng Wei Tang, 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):29–35, 2007.

[15] Gregory G Schwartz, Philippe Gabriel Steg, Michael Szarek, Vera A Bittner, Rafael Diaz, Shaun G Goodman, Yong-Un Kim, J Wouter Jukema, Robert Pordy, Matthew T Roe, et al. Peripheral artery disease and venous thromboembolic events after acute coronary syndrome: role of lipoprotein (a) and modification by alirocumab: prespecified analysis of the odyssey outcomes randomized clinical trial. *Circulation*, 141(20):1608–1617, 2020.

[16] SW Davies, AH Gershlick, and R Balcon. Progression of valvar aortic stenosis: a long-term retrospective study. *European heart journal*, 12(1):10–14, 1991.

[17] Kristen A Severson, Soumya Ghosh, and Kenney Ng. Unsupervised learning with contrastive latent variable models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4862–4869, 2019.

[18] Blase A Carabello and Walter J Paulus. Aortic stenosis. *The lancet*, 373(9667):956–966, 2009.

[19] Raphael Rosenhek, Ursula Klaar, Michael Schemper, Christine Scholten, Maria Heger, Harald Gabriel, Thomas Binder, Gerald Maurer, and Helmut Baumgartner. Mild and moderate aortic stenosis: natural history and risk stratification by echocardiography. *European Heart Journal*, 25(3):199–205, 2004.

[20] Clifton Phua, Damminda Alahakoon, and Vincent Lee. Minority report in fraud detection: classification of skewed data. *Acm sigkdd explorations newsletter*, 6(1):50–59, 2004.

[21] Xinjian Guo, Yilong Yin, Cailing Dong, Gongping Yang, and Guangtong Zhou. On the class imbalance problem. In *2008 Fourth international conference on natural computation*, volume 4, pages 192–201. IEEE, 2008.

[22] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[23] Yanmin Sun, Andrew KC Wong, and Mohamed S Kamel. Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, 23(04):687–719, 2009.

[24] Chris Seiffert, Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. Resampling or reweighting: A comparison of boosting implementations. In *2008 20th IEEE International Conference on Tools with Artificial Intelligence*, volume 1, pages 445–451. IEEE, 2008.

[25] Shiven Sharma, Colin Bellinger, Bartosz Krawczyk, Osmar Zaiane, and Nathalie Japkowicz. Synthetic oversampling with the majority class: A new perspective on handling extreme imbalance. In *2018 IEEE international conference on data mining (ICDM)*, pages 447–456. IEEE, 2018.

[26] Abubakar Abid, Martin J Zhang, Vivek K Bagaria, and James Zou. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nature communications*, 9(1):1–7, 2018.

[27] Abubakar Abid and James Zou. Contrastive variational autoencoder enhances salient features. *arXiv preprint arXiv:1902.04601*, 2019.

[28] Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

[29] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[30] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

[31] Takaya Saito 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):e0118432, 2015.

[32] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[33] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019.

[34] Mathias Carl Blom, Awais Ashfaq, Anita Sant'Anna, Philip D Anderson, and Markus Lingman. Training machine learning models to predict 30-day mortality in patients discharged from the emergency department: a retrospective, population-based registry study. *BMJ open*, 9(8):e028015, 2019.

[35] Akram Mohammed, Pradeep SB Podila, Robert L Davis, Kenneth I Ataga, Jane S Hankins, and Rishikesan Kamaleswaran. Using machine learning to predict early onset acute organ failure in critically ill intensive care unit patients with sickle cell disease: Retrospective study. *Journal of Medical Internet Research*, 22(5):e14693, 2020.

[36] Denis Agniel, Isaac S Kohane, and Griffin M Weber. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *Bmj*, 361, 2018.

[37] Richard Ambrosino, Bruce G Buchanan, Gregory F Cooper, and Michael J Fine. The use of misclassification costs to learn rule-based decision support models for cost-effective hospital admission strategies. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 304. American Medical Informatics Association, 1995.

[38] Donald B Rubin. *Matched sampling for causal effects.* Cambridge University Press, 2006.

[39] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[40] Xavier de Luna and Per Johansson. Testing for the unconfoundedness assumption using an instrumental assumption. *Journal of Causal Inference*, 2(2):187–199, 2014.

[41] Melissa M Garrido, Amy S Kelley, Julia Paris, Katherine Roza, Diane E Meier, R Sean Morrison, and Melissa D Aldridge. Methods for constructing and assessing propensity scores. *Health services research*, 49(5):1701–1720, 2014.

[42] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

[43] Stuart J Pocock, Susan E Assmann, Laura E Enos, and Linda E Kasten. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practiceand problems. *Statistics in medicine*, 21(19):2917–2930, 2002.

[44] Rajeev H Dehejia and Sadek Wahba. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, 84(1):151–161, 2002.

[45] Paul R Rosenbaum et al. *Design of observational studies*, volume 10. Springer, 2010.

[46] Fan Li, Kari Lock Morgan, and Alan M Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.

[47] Nathan Kallus. Generalized optimal matching methods for causal inference. *arXiv preprint arXiv:1612.08321*, 2016.

[48] Michael Oberst, Fredrik Johansson, Dennis Wei, Tian Gao, Gabriel Brat, David Sontag, and Kush Varshney. Characterization of overlap in observational studies. In *International Conference on Artificial Intelligence and Statistics*, pages 788–798. PMLR, 2020.

[49] Wangzhi Dai, Kenney Ng, Kristen Severson, Wei Huang, Fred Anderson, and Collin Stultz. Generative oversampling with a contrastive variational autoencoder. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 101–109. IEEE, 2019.

[50] Yuhuai Wu, Yuri Burda, Ruslan Salakhutdinov, and Roger Grosse. On the quantitative analysis of decoder-based generative models. *arXiv preprint arXiv:1611.04273*, 2016.

[51] Lawrence Cayton. Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep*, 12(1-17):1, 2005.

[52] Keith AA Fox, Gordon FitzGerald, Etienne Puymirat, Wei Huang, Kathryn Carruthers, Tabassome Simon, Pierre Coste, Jacques Monsegu, Philippe Gabriel Steg, Nicolas Danchin, et al. Should patients with acute coronary disease be stratified for management according to their risk? derivation, external validation and outcomes using the updated grace risk score. *BMJ open*, 4(2), 2014.

[53] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[54] Susan Athey, Guido W Imbens, et al. Machine learning for estimating heterogeneous causal effects. Technical report, 2015.

[55] Susan M Shortreed and Ashkan Ertefaie. Outcome-adaptive lasso: variable selection for causal inference. *Biometrics*, 73(4):1111–1122, 2017.

[56] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029. PMLR, 2016.

[57] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. *Advances in Neural Information Processing Systems*, 31, 2018.

[58] Wangzhi Dai and Collin M Stultz. Quantifying common support between multiple treatment groups using a contrastive-vae. In *Machine Learning for Health*, pages 41–52. PMLR, 2020.

[59] Brian H Grimard and Jan M Larson. Aortic stenosis: diagnosis and treatment. *American family physician*, 78(6):717–724, 2008.

[60] Blase A Carabello. Evaluation and management of patients with aortic stenosis. *Circulation*, 105(15):1746–1750, 2002.

[61] Graham H Bevan, David A Zidar, Richard A Josephson, and Sadeer G Al-Kindi. Mortality due to aortic stenosis in the united states, 2008-2017. *Jama*, 321(22):2236–2238, 2019.

[62] Helmut Baumgartner, Judy Hung, Javier Bermejo, John B Chambers, Thor Edvardsen, Steven Goldstein, Patrizio Lancellotti, Melissa LeFevre, Fletcher Miller Jr, Catherine M Otto, et al. Recommendations on the echocardiographic assessment of aortic valve stenosis: a focused update from the european association of cardiovascular imaging and the american society of echocardiography. *European Heart Journal-Cardiovascular Imaging*, 18(3):254–275, 2017.

[63] Jeffrey Zhang, Sravani Gajjala, Pulkit Agrawal, Geoffrey H Tison, Laura A Hallock, Lauren Beussink-Nelson, Mats H Lassen, Eugene Fan, Mandar A Aras, ChaRandle Jordan, et al. Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy. *Circulation*, 138(16):1623–1635, 2018.

[64] Ali Madani, Ramy Arnaout, Mohammad Mofrad, and Rima Arnaout. Fast and accurate view classification of echocardiograms using deep learning. *NPJ digital medicine*, 1(1):1–8, 2018.

[65] Arghavan Arafati, Daisuke Morisawa, Michael R Avendi, M Reza Amini, Ramin A Assadi, Hamid Jafarkhani, and Arash Kheradvar. Generalizable fully automated multi-label segmentation of four-chamber view echocardiograms based on deep convolutional adversarial networks. *Journal of The Royal Society Interface*, 17(169):20200267, 2020.

[66] Shakiba Moradi, Mostafa Ghelich Oghli, Azin Alizadehasl, Isaac Shiri, Niki Oveisi, Mehrdad Oveisi, Majid Maleki, and Jan Dhooge. Mfp-unet: A novel deep learning based approach for left ventricle segmentation in echocardiography. *Physica Medica*, 67:58–69, 2019.

[67] Rohan Shad, Nicolas Quach, Robyn Fong, Patpilai Kasinpila, Cayley Bowles, Miguel Castro, Ashrith Guha, Erik E Suarez, Stefan Jovinge, Sangjin Lee, et al.

Predicting post-operative right ventricular failure using video-based deep learning. *Nature communications*, 12(1):1–8, 2021.

[68] Kenya Kusunose, Takashi Abe, Akihiro Haga, Daiju Fukuda, Hirotsugu Yamada, Masafumi Harada, and Masataka Sata. A deep learning approach for assessment of regional wall motion abnormality from echocardiographic images. *Cardiovascular Imaging*, 13(2_Part_1):374–381, 2020.

[69] Zhe Huang, Gary Long, Benjamin Wessler, and Michael C Hughes. A new semi-supervised learning benchmark for classifying view and diagnosing aortic stenosis from echocardiograms. In *Machine Learning for Healthcare Conference*, pages 614–647. PMLR, 2021.

[70] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[71] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[72] David M Steinberg, Jason Fine, and Rick Chappell. Sample size for positive and negative predictive value in diagnostic research using case–control designs. *Biostatistics*, 10(1):94–105, 2009.

[73] John Chambers. Aortic stenosis, 2005.

[74] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[75] Hasan Alzahrani, Michael Y Woo, Chris Johnson, Paul Pageau, Scott Millington, and Venkatesh Thiruganasambandamoorthy. Can severe aortic stenosis be identified by emergency physicians when interpreting a simplified two-view echocardiogram obtained by trained echocardiographers? *Critical ultrasound journal*, 7(1):1–4, 2015.

[76] Steven J Lester, Brett Heilbron, Ken Gin, Arthur Dodek, and John Jue. The natural history and rate of progression of aortic stenosis. *Chest*, 113(4):1109–1114, 1998.

[77] Martin Peter, Andreas Hoffmann, Clifford Parker, Thomas Lüscher, and Dieter Burckhardt. Progression of aortic stenosis: role of age and concomitant coronary artery disease. *Chest*, 103(6):1715–1719, 1993.

[78] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision

transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[79] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[80] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

[81] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[82] Nathaniel Diamant, Erik Reinertsen, Steven Song, Aaron D Aguirre, Collin M Stultz, and Puneet Batra. Patient contrastive learning: A performant, expressive, and practical approach to electrocardiogram modeling. *PLoS computational biology*, 18(2):e1009862, 2022.

[83] Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. 3d self-supervised methods for medical imaging. *Advances in Neural Information Processing Systems*, 33:18158–18172, 2020.

[84] Aiham Taleb, Christoph Lippert, Tassilo Klein, and Moin Nabi. Multimodal self-supervised learning for medical image analysis. In *International Conference on Information Processing in Medical Imaging*, pages 661–673. Springer, 2021.

[85] Dominique Makowski, Tam Pham, Zen J Lau, Jan C Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and SH Chen. Neurokit2: A python toolbox for neurophysiological signal processing. *Behavior research methods*, 53(4):1689–1696, 2021.