Improving Clinical Prediction Models with Statistical Representation Learning

by

Zidi Xiu

Department of Biostatistics & Bioinformatics
Duke University

Date: _____

Approved:

_____
Benjamin Goldstein, Supervisor

_____
Ricardo Henao Giraldo

_____
David Page

_____
Sheng Luo

_____
Yuan Wu

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Biostatistics & Bioinformatics
in the Graduate School of Duke University
2021

Abstract

Improving Clinical Prediction Models with Statistical Representation Learning

by

Zidi Xiu

Department of Biostatistics & Bioinformatics
Duke University

Date: _____

Approved:

_____
Benjamin Goldstein, Supervisor

_____
Ricardo Henao Giraldo

_____
David Page

_____
Sheng Luo

_____
Yuan Wu

An abstract of a dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor of Philosophy
in the Department of Biostatistics & Bioinformatics
in the Graduate School of Duke University
2021

Copyright © 2021 by Zidi Xiu
All rights reserved

Abstract

This dissertation studies novel statistical machine learning approaches for healthcare risk prediction applications in the presence of challenging scenarios, such as rare-events, noisy observations, data imbalance, missingness and censoring. Such scenarios manifest frequently in practice, and they compromise the validity of standard predictive models which often expect clean and complete data. As such, alleviating the negative impacts of real-world data challenges is of great significance and constitutes the overarching goal of this dissertation, which investigates novel strategies to (*i*) account for data uncertainties and statistical characteristics of low-prevalence events; (*ii*) re-balancing and augmenting the representations for minority data under proper causal assumptions; (*iii*) dynamically assigning scores and attend to the observed units to derive robust features. By integrating the ideas from representation learning, variational Bayes, causal inference, and contrastive training, this dissertation builds tools for risk modeling frameworks that are robust to various peculiarities of real-world datasets to yield reliable individualized risk evaluations.

This dissertation starts with a systematic review of classical risk prediction models in Chapter 1, and discusses the new opportunities and challenges presented by the big data era. With the increasing availability of healthcare data and the current rapid development of machine learning models, clinical decision support systems have seen new opportunities to improve clinical practice. However, in healthcare risk prediction applications, statistical analysis is not only challenged by data incompleteness

and skewed distributions but also the complexity of the inputs. To motivate the subsequent developments, discussions on the limitations in risk minimization methods, robustness against high-dimensional data with incompleteness, and the need for individualization are provided.

As a concrete example to address a canonical problem, Chapter 2 proposes a variational disentanglement approach to semi-parametrically learn from the heavily imbalanced binary classification datasets. In this new method, which is named *Variational Inference for Extremals* (VIE), we apply an extreme value theory to enable efficient learning with few observations. By organically integrating the generalized additive model and isotonic neural nets, VIE enjoys the merits of improved robustness, interpretability, and generalizability for the accurate prediction of rare events. An analysis of the COVID-19 cohort from Duke Hospitals demonstrates that the proposed approach outperforms competing solutions.

We investigate a more generalized setting of a multi-classification problem with heavily imbalanced data in Chapter 3, from the perspective of causal machine learning to promote sample efficiency and model generalization. Our solution, named *Energy-based Causal Representation Transfer* (ECRT), posits a meta-distributional scenario, where the data generating mechanism for label-conditional features is invariant across different labels. Such causal assumption enables efficient knowledge transfer from the dominant classes to their under-represented counterparts, even if their feature distributions show apparent disparities. This allows us to leverage a causally informed data augmentation procedure based on nonlinear independent component analysis to enrich the representation of minority classes and simultaneously data whitening. The effectiveness and enhanced prediction accuracy are demonstrated through synthetic data and real-world benchmarks compared with state-of-art models.

In Chapter 4 we deal with the time-to-event prediction with censored (missing

in outcomes) and incomplete (missing in covariates) observations. Also known as survival analysis, time-to-event prediction plays a crucial role in many clinical applications, yet the classical survival solutions scale poorly w.r.t. data complexity and incompleteness. To better handle sophisticated modern health data and alleviate the impact of real-world data challenges, we introduce a self-attention based model to capture helpful information for time-to-event prediction, called *Energy-based Latent Self-Attentive Survival Analysis* (ELSSA). A key novelty of this approach is the integration of contrastive mutual information based loss that non-parametrically maximizes the informativeness of learned data representations. The effectiveness of our approaches has been intensively validated with synthetic and real-world benchmarks, showing improved performance relative to competing solutions.

In summary, this dissertation presents flexible risk prediction frameworks that acknowledge representation uncertainties, data heterogeneity and incompleteness. Altogether, it presents three contributions: improved efficient learning from imbalanced data, enhanced robustness to missing data, and better generalizability to out-of-sample subjects.

Dedicated to my beloved parents.

Contents

Abstract **iv**

List of Tables **xiii**

List of Figures **xv**

Acknowledgements **xvii**

1 Introduction **1**

1.1 Challenges in clinical risk prediction 4

1.2 Dissertation structure 6

2 Variational Disentanglement for Rare Event Modeling **7**

2.1 Introduction 7

2.2 Background 10

2.2.1 Variational inference 11

2.2.2 Extreme value theory 12

2.2.3 Additive isotonic regression 13

2.3 Variational Inference of Extremals 13

2.3.1 Decoder: additive monotonic neural network 16

2.3.2 Latent extreme prior: Gaussian-GPD mixture 17

2.3.3 Latent posterior: Inverse Autoregressive Flow 18

2.3.4 Posterior matching with Fenchel mini-max learning 19

2.4 Related Work 21

2.5 Experiments 22

2.6 Conclusions 27

3 Supercharging Imbalanced Data Learning with Energy-based Contrastive Representation Transfer 28

3.1 Introduction 28

3.2 Preliminaries 31

3.2.1 Invertible neural network 33

3.3 Energy-based Causal Representation Transfer 34

3.3.1 Model assumptions 35

3.3.2 Causal Representation Transfer 36

3.3.3 Improving causal representation transfer 38

3.3.4 Insights and remarks 42

3.4 Related Work 43

3.5 Experiments 45

3.5.1 Experimental setup 46

3.6 Conclusion 49

4 Personalized Time-to-event Estimation with Representation Learning 50

4.1 Introduction 50

4.2 Background 54

4.2.1 Statistical survival analysis 54

4.3 Preliminaries for VSI 55

4.3.1 Variational inference 55

4.4 Variational Survival Inference 56

4.4.1 Variational bound of observed events 57

4.4.2 Variational bound of censored events 58

4.4.3 Implementing VSI 59
4.4.4 Importance-Weighted estimator for likelihood evaluation . . . 60
4.4.5 Making predictions 61
4.4.6 DISSECTING VSI 61
4.5 Preliminaries for ELSSA 63
4.5.1 Attention and Self-attention 64
4.5.2 Contrastive Learning 65
4.6 Energy-based Latent Self-attentive Survival Analysis 66
4.6.1 Model Assumptions 66
4.6.2 Covariates Embedding and Self-Attention 66
4.6.3 Energy-based Contrastive Learning 69
4.6.4 Learning Objectives 69
4.7 Related Work 70
4.8 Experiments 72
4.8.1 Evaluation Metrics 73
4.8.2 Synthetic datasets 75
4.8.3 Real-World datasets 79
4.9 Conclusions 82
4.9.1 Variational Survival Inference 82
4.9.2 ELSSA 82

5 Conclusions **85**
5.1 Summary 85
5.2 Future work 87

A Supplementary Material to "Variational Disentanglement for Rare Event Modeling" **90**

A.1 Derivation of Mixed GPD tail distribution 91
A.2 Implementation Details 93
A.2.1 Model Structure. 93
A.2.2 Numerical Integration. 95
A.2.3 Discussions on Evaluation Metrics. 96
A.3 Ablation Study 96
A.3.1 Semi-synthetic Datasets 96
A.3.2 Additional Results for semi-synthetic datasets 97
A.3.3 Long-tailed Synthetic Datasets 98
A.4 Real-world datasets 100
A.4.1 Baseline Models. 100
A.4.2 COVID Dataset 101
A.4.3 InP Dataset 102
A.4.4 SEER and SLEEP Datasets 103
A.4.5 Credit Card Fraud Detection 103
A.4.6 Exploration of the feature representation 104
A.4.7 E. Generalized to Multiple-class classification 105

B Supplementary Material to "Supercharging Imbalanced Data Learning With Energy-based Contrastive Representation Transfer" 109

B.1 Theoretical Support 109
B.1.1 Nonlinear ICA with auxiliary variables 109
B.1.2 Variance and generalization bound 110
B.1.3 Speedup from shared embedding 112
B.1.4 Energy-based GCL 113
B.2 Regression for continuous labels 115
B.3 Implementation of Augmentation 115

B.4 Toy Model Experiment 117

B.4.1 Toy data demo 117

B.4.2 Extreme-classification Toy Data 117

B.5 Real-world Data Experiments 118

B.5.1 Image Data Benchmarks 118

B.5.2 Language Data Benchmark 119

C Supplementary Material to "Personalized Time-to-event Distribution with Deep Learning Solutions" 124

C.1 Implementing VSI 124

C.2 Derivation of Likelihood lowerbound for censored observations 127

C.3 Likelihood Calculation for Statistical Models 128

C.3.1 AFT Models 128

C.3.2 Random Survival Forest 129

C.3.3 Cox Models 130

C.4 Additional Results for Simulation Datasets 131

C.5 Additional Results for Real Datasets 133

Bibliography 136

List of Tables

2.1 Ablation study of VIE 19

2.2 Performance on real-world datasets 23

2.3 Summary statistics 25

3.1 Performance on real-world datasets 47

4.1 KS statistic for simulation study. 77

4.2 Performance on simulation study 77

4.3 Summary Statistics for Real Datasets. 80

4.4 Performance on real-world datasets 81

4.5 Quantile ranges for log-likelihood in real-world datasets. 81

A.1 Ablation study of VIE with 0.5% event rate in semi-synthetic settings. 98

A.2 Ablation study of VIE with 5% event rate in semi-synthetic settings. 99

A.3 Ablation study of VIE with 1% event rate in longtailed-synthetic settings 100

A.4 `COVID` dataset covariates before pre-processing. 101

A.5 5-Fold cross validation results of `COVID` dataset 102

A.6 `InP` dataset covariates before pre-processing 103

A.7 Fraud transaction classification. 104

A.8 Performance on few-shots learning dataset 107

B.1 Validation results for extreme classification 117

B.2 Summary statistics of real-world datasets 121

B.3 MNIST experiment network architecture 122

B.4 Cifar100 experiment network architecture 122
B.5 iNaturalist experiment network architecture 122
B.6 Tiny ImageNet experiment network architecture 122
B.7 Arxiv experiment network architecture 123
B.8 MNIST results with different numbers of minority categories 123
B.9 GAN network architecture. 123
B.10 Hyperparameter of datasets (ECRT) 123
B.11 Hyperparameter of datasets (GAN) 123
C.1 80% quantile range of log-likelihood in simulation study 132
C.2 Raw C-Index with confidence intervals for simulation studies 132
C.3 Comparison of point estimation in real-world datasets 133
C.4 Raw C-Index with confidence intervals for real-world datasets 135

List of Figures

2.1 Tail estimation 16

2.2 Illustration of monotonicity in the latent space 26

2.3 Results on `COVID` mortality data 26

3.1 ECRT motivation 29

3.2 Illustration of data space $\mathcal{X}$, feature space $\mathcal{Z}$ and source space $\mathcal{S}$ . . . 32

3.3 Source space estimation module of ECRT 35

3.4 Non-parametric source space augmentation 35

3.5 Comparison of different augmentation strategies 41

3.6 Extreme classification results 44

3.7 Comparison of different causal augmentations 44

3.8 ECRT source representation trained on `iNaturalist` 45

3.9 Performance comparison with and without augmentation. 47

3.10 Comparison of learning dynamics 47

3.11 Sensitivity to hyperparameter λ 47

4.1 Illustration of personalized time-to-event estimation (observed) 76

4.2 Illustration of personalized time-to-event estimation (censored) 76

4.3 Simulated time-to-event distributions with 30% event rate 76

4.4 Log-likelihood distributions on testing datasets 78

4.5 Coverage rate on testing datasets 79

4.6 log-likelihood distributions for real-world datasets 81

4.7 Coverage rate for real-world datasets 82

A.1 Feature representation mismatch at the tail parts 91

A.2 Box plot of 10 independent 1% event rate semi-synthetic analysis. . . 98

A.3 Bootstrapped performance metrics on `COVID` dataset 102

A.4 Bootstrapped performance metric on `InP` dataset 103

A.5 Bootstrapped performance metric on `SEER` dataset 104

A.6 Bootstrapped performance metric on `SLEEP` dataset 104

A.7 Learned prior and posterior distribution and monotonic predicted risks 106

A.8 The latent representation values distribution grouped by event type . 106

A.9 Latent dimensions from the `InP` dataset 106

A.10 t-SNE plots with latent representation z. 107

A.11 Illustration of multi-classification framework. 107

B.1 Comparison of correlation decreasing with GCL and FDV in the MNIST dataset 114

B.2 Feature space augmentation for MNIST 116

B.3 Sensitivity analysis (overall top 1 accuracy) of augmentation strength λ119

B.4 Class-conditional Top-1 accuracy curve for iNat2019 120

C.1 Testing datasets log-likelihood distributions for simulation dataset. . . 131

C.2 Testing datasets coverage rate for simulation datasets 133

C.3 Likelihood distributions for real datasets 134

C.4 Coverage rates for real-world datasets 134

Acknowledgements

I would like first to express my sincere appreciation to my advisors, Prof. Benjamin Goldstein and Prof. Ricardo Henao, who have provided me valuable support, both personally and academically, and given me many opportunities to learn, to practice, and to think scientifically. During the past four years, they spent countless hours guiding me to become an independent researcher. Additional thanks to Dr. Chenyang Tao, as a colleague and a mentor, for his patience, insightful feedback, and demonstration of critical reasoning during our collaborations.

I would like to express my gratitude to my committee members, Prof. David Page, Prof. Sheng Luo, and Prof. Yuan Wu, for their generous input and guidance over the years. It's challenging to complete a machine learning thesis in the Biostatics and Bioinformatics Department, and I could not achieve it without their understanding and support. I would like to thank Prof. Michael Pencina, who gave me valuable feedback. Thanks to Prof. Kouros Owzar, my master's project advisor, for providing positive research experiences during my master's studies at Duke. Additional thanks to Prof. Li Ma, I enjoyed the fantastic lectures on statistical inference, which made the statistical theories less intimidating.

I also want to acknowledge all my collaborators. Thanks to Dr. Cara O'Brien and Dr. Armando Bendoya, they gave me the valuable experience to work with first-handed clinical data and provided insightful clinical views of the research problems. I appreciate Junya, and I enjoyed her brilliant ideas and casual chats when we stayed

up all night for several projects. Thanks to Michael and Connor for the detailed feedback during our collaboration. Thanks also to Anna and Jonathan for the inspiring weekly discussions. I feel fortunate to be in a warm research environment.

The same can be said of my peers in the Biostatistics and Bioinformatics Department. Thanks to my peers for all the interesting academic discussions and homework problems. I would like to extend my sincere thanks to Duke and the B&B Department for organizing the great colloquium to learn, explore, and collaborate. Special thanks to Prof. Sean O'Brien, Prof. Huiman Barnhart, and Prof. Megan Neely for their administrative effort to improve the program. Thanks should also go to Ms. Michelle Evans, Ms. Kendall Mincey, and Ms. LaDonna Huseman for their patience and support to me.

Finally, I am eternally grateful for the love and care from my family. I thank my parents for their love and unconditional support throughout my life. I am forever indebted to my parents for giving me the opportunities and experiences that have made me who I am and giving me the strength to chase my dreams without any concerns. I would also like to share this moment with my beloved grandparents, although my grandpa is not with me anymore. I am so happy that I was born into this family. And in the end, I would like to thank my love Yanwei. He is my best friend, my reliable partner, and my irreplaceable soulmate. I could not complete my thesis without his company, together with Goki and Lychee, my lovely furry family members. They cheered me up when I was down and helpless, standing by my side when I doubted myself.

This dissertation is just a brand new start of my life.

1

Introduction

The abundance of modern health data provides many opportunities and challenges for using machine learning techniques to build better statistical models to facilitate clinical decision-making. The emergence of Electrical Health Records (EHR) data is a great example that enables personalized therapies with dynamic and structured datasets (Goldstein et al., 2017). Early detection of diseases from patient medical histories or identifying in-hospital patients at immediate risk of life-threatening events has significant value, thus calling for effective and reliable statistical models.

In risk prediction scenarios, standard statistical methods for structured datasets have been widely studied. Starting from linear specifications, $Y = g(X\beta) + \epsilon$, regression models explain the outcome Y with the linear combination of the input X and coefficients β given the link function $g(\cdot)$; set to the identity for the case of regression. The *Least Absolute Shrinkage and Selection Operator* (LASSO) (Tibshirani, 1996) imposes a variable-selection regularization term on the regression coefficients, and generalized Linear Models (GLM) (Nelder and Wedderburn, 1972) extend the identity link to nonlinear cases (logistic, log-log, *etc.*), which allow for flexibility in the type of outcomes that can be considered, *e.g.*, using the logistic link for binary

classification. However, there are three limitations of regression based prediction methods, (i) linearity in coefficients, (ii)lack of heterogeneity, (iii) the ratio of the number of predictors and the number of observations needs to be small (Steyerberg and Vergouwe, 2014; Sugiyama, 2015)

Unlike parametric linear regression methods, non-parametric models enjoy more flexibility in their formulation (Anzai, 2012), and can be powerful when adequate training samples are provided. Decision trees and their variants have been widely used in risk prediction tasks. Decision trees are in principle interpretable, can be easily understood by humans (Safavian and Landgrebe, 1991), and are optimized based on entropy gain on each attribute splitting. To improve the prediction accuracy, ensemble methods built upon decision trees are quite popular, including bagging (averaging outcomes from several decision trees grown with subsets of training data) and boosting (a sequence of weak learner decision trees optimized sequentially to reach a strong learner). Two famous algorithms based on these techniques are *Random Forest* (Pal, 2005) and *XGBoost* (Chen and Guestrin, 2016). Kernel methods are another important branch of non-parametric risk prediction. The intuition behind these methods consists on mapping the non-separable data to a hyperspace where the prediction can be made more easily Hofmann et al. (2008). A linear combination of kernel functions represents the predictions learned from the training dataset. *Support Vector Machines* (Noble, 2006) is a simple application of kernel methods. The development of neural networks further relaxes the linear combination of covariates to complex forms (Schmidhuber, 2015). Moreover, the variety of data sources has dramatically increased with the development of novel technologies. Apart from structured datasets, images, audio sequences, natural language texts, and other forms of data rely on feature engineering or, more recently, deep learning methods for preprocessing and analysis (Shah et al., 2019).

However, in most machine learning prediction models, the training loss mini-

mization strategy is the only parameter optimization tool. Without constraints, the machine tries to recklessly absorb all the correlations and information presented in the complex training dataset (Arjovsky et al., 2019), including noise and spurious correlations. Unlike clinical trials where the study cohorts are carefully selected, real-world clinical datasets like EHR suffer from poor data quality and validity (Kohane et al., 2021). Without proper regularization to the loss minimization schemes, black-box algorithms tend to be not generalizable and robust to new datasets or observations. Researchers have shown that loss minimization algorithms perform no better than traditional logistic regression in clinical datasets (Christodoulou et al., 2019). Even when the classification outcome is correct, the reasoning and interpretation gained from the model can be problematic, yielding poor performance in the real world out-of-sample cases (Zech et al., 2018; Schramowski et al., 2020). Without a consistent performance, the translation from statistical findings to clinical decisions may not be safe or feasible (Rudin, 2019; Pencina et al., 2020).

To satisfy the needs of stable and reliable prediction models, a wealth of results in the past few years suggest jumping out of black-box algorithms and instead, focus on shedding more light on the interpretation to increase accountability and transparency. Representation learning allows extracting latent features from high-dimensional and complex datasets that are useful for prediction tasks. (Bengio et al., 2013). The Information Bottleneck principle proposed by Tishby et al. (2000) showed the ability of a bottleneck shaped deep learning architecture can compress the noisy inputs into low-dimensional encoded representations, and it can further reduce the complexity (Tishby and Zaslavsky, 2015). Previous work has tried to capture representations from diagnosis code and medical procedure code in EHR datasets (Choi et al., 2016) with feed-forward neural networks. In probabilistic models with latent variables, from the Bayesian view, a good representation is often one that captures the latent posterior distribution of the underlying explanatory components from the

observed input variables (Bengio et al., 2013). For instance, topic modeling summarizes count data with a hierarchical latent variable model specified as a *latent Dirichlet allocation* (LDA) (Blei et al., 2003) or a Poisson factor model (Henao et al., 2015; Gan et al., 2015). Recently, the interaction between causal graphs and machine learning techniques is getting increased attention (Schölkopf et al., 2021). In the context of causal machine learning, Arjovsky et al. (2019) identifies the invariant causal features by penalizing environmental heterogeneity unrelated to the outcome of interest that is spurious. Interpretable machine learning (Ahmad et al., 2018) is an alternative way to balance between performance and generalizability. By including interpretable penalties which are specifically chosen to match the domain of interests, the unconstrained machine learning methods can be better understood by humans (Rudin et al., 2021).

1.1 Challenges in clinical risk prediction

Besides the interpretability and robustness of machine learning methods, clinical data analysis has other challenges (Goldstein et al., 2017). As a motivating example, considering in-patient data from 82,450 Duke Health patients collected between 2014-2016. We abstracted time-varying clinical data (*i.e.*, vital signs, laboratory tests, medications) and followed patients until Intensive Care Unit (ICU) transfer or Death. Such high-dimensional, multivariate longitudinal survival datasets with rare event rates ($\leqslant$ 5%) provide rich information and challenges for data analysis which we describe below.

Data missingness. In EHR-based studies, informative missingness is a major challenges (Goldstein et al., 2017). Missingness can occur in both covariates and the outcomes. In standard machine learning methods, for tabular datasets, it generally requires the input datasets to be preprocessed with no missing values included, *i.e.*, by deleting observations with missing values or imputation (García-Laencina

et al., 2010). The *missing at completely random* (MCAR) assumption does not hold in general (Little and Rubin, 2019), therefore it's not appropriate to use complete observations only. Under *missing at random* (MAR) assumption, *i.e.,* explainable with observed variables, imputation methods are widely used. Apart from standard EM algorithm (Rubin, 1996), machine learning based imputation methods have been introduced. Recently, attention based imputation methods showed promising results (Haneuse et al., 2021), by structural information extracted from the input covariates. Rooted from attention mechanism, we apply an attention layer on top of the input covariates to handle the missing variables, without directly impute the missing values in the prediction tasks.

Censored outcomes. Censoring is a scenario where the outcome is only partially known, and it is not uncommon in clinical datasets, especially in time-to-event prediction. There are three reasons why this can happen, (*i*) when a subject drops out during the study (lost to follow-up), (*ii*) when the study terminates before the event happens (administrative censoring), (*iii*) a subject withdraws from the study (Klein and Moeschberger, 2006). Similar to the missing data problem discussed above, if the censored occurs completely at random, it can usually be safely ignored. In the case of informative censoring, when participants are lost to follow-up or withdrawn due to reasons that are related to the study, such censoring contains crucial information that should be adequately considered (Scharfstein and Robins, 2002).

Low event rate. Early detection of rare diseases is valuable. In our motivating dataset, the patients who experienced target events (*e.g.*, ICU transfer, death) only account for a small portion of the population (*i.e.*, less than 5%). Such low prevalence makes traditional statistical modeling of risk particularly challenging, especially when the sample size of the group of interest is small and the data is complex. Flexible non-parametric machine learning techniques tend to over-fit given the small number of event observations and the disproportionally large number of model parameters

(Shah et al., 2019).

By integrating ideas from representation learning, variational Bayes, causal inference, and contrastive training, the overall objective of this dissertation is to build risk modeling frameworks that are robust to various peculiarities of real-world datasets to yield reliable individualized risk evaluations.

1.2 Dissertation structure

The rest of this dissertation is organized as follows. Chapter 2 and Chapter 3 focus on the topics of imbalanced learning in two different settings, binary classification and multi-class classification. Chapter 2 capitalizes on the extreme behavior of the rare events in latent space with an *Extreme Value Theory* (EVT) motivated prior distribution under the *Variational Inference* (VI) framework. Chapter 3 views the imbalanced learning tasks from the causal machine learning perspective. On the basis of *nonlinear independent component analysis* (NICA), augmenting the minority classes relies on the feature independence. Chapter 4 discusses at length about time-to-event estimation and provides novel frameworks for discrete approximations to individual time-to-event distribution estimation in the presence of missing data, called *Energy-based Latent Self-Attentive Survival Analysis* (ELSSA). By combining contrastive mutual information based loss with survival log-likelihood, it can non-parametrically maximizes the informativeness of learned data representations for time-to-event prediction and effectively handle censored subjects.

2

Variational Disentanglement for Rare Event Modeling

2.1 Introduction

Early identification of in-hospital patients who are at imminent risk of life-threatening events, *e.g.*, death, ventilation or intensive care unit (ICU) transfer, is a critical subject in clinical care Bedoya et al. (2019). Especially during a pandemic like COVID-19, the needs for healthcare change dramatically. With the ability to accurately predict the risk, an automated triage system will be well-positioned to help clinicians better allocate resources and attention to those patients whose adverse outcomes can be averted if early intervention efforts were in place.

Despite the great promise it holds, with the richness of modern Electronic Health Record (EHR) repositories, the construction of such a system faces practical challenges. A major obstacle is the scarcity of patients experiencing adverse outcomes of interest. In the COVID-19 scenario, which we consider in our experiments, the

This chapter is a published paper: Xiu, Zidi, Chenyang Tao, Michael Gao, Connor Davis, Benjamin Goldstein, and Ricardo Henao. "Variational disentanglement for rare event modeling." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 12, pp. 10469-10477. 2021 (Xiu et al., 2021)

mortality of patients tested positive at the Duke University Health System (DUHS) is slightly lower than 3%. Further, in another typical EHR dataset we consider, less than 5% of patients are reported to suffer adverse outcomes (ICU transfer or death). In these *low-prevalence* scenarios, commonly seen in clinical practice, standard classification models such as logistic regression suffer from *majority domination*, in which models tend to favor the prediction accuracy of majority groups. This is clearly undesirable for critical-care applications, given the high false negative rates (Type-II error), in which patients in urgent need of care could be falsely categorized.

Situations where the distribution of labels is highly skewed and the accuracy of the minority class bears particular significance Dal Pozzolo et al. (2017); Lu et al. (2020); Machado and Lopes (2020) have been associated with the name *imbalanced dataset* He and Garcia (2009), whereas the methods dealing with such cases are coined *extreme classification* Zong et al. (2013). Under such a setting, the lack of representation of minority cases severely undermines the ability of a standard learner to discriminate, relative to balanced datasets Mitchell (1999). Consequently, these solutions do not generalize well on minority classes, where the primary interest is usually focused.

To address such a dilemma, several remedies have been proposed to account for the imbalance between class representations. One of the most popular strategies is the sampling-based adjustment, where during training, a model oversamples the minority classes (or undersamples the majority classes) to create balance artificially Holte et al. (1989); Drummond et al. (2003). To overcome the biases and the lack of information that naive sampling adjustments might induce, variants have been proposed to maximally preserve the clustering structure of the original dataset Mani and Zhang (2003); Yen and Lee (2009) and to promote diversity of oversampling schemes Han et al. (2005). Alternatively, cost-sensitive weighting where minority losses are assigned larger weights provides another popular option via tuning the

relative importance of minority classes Elkan (2001); Munro et al. (1996); Zhou and Liu (2005).

While the above two strategies introduce heuristics to alleviate the issues caused by class imbalance, importance sampling (IS) offers a principled treatment that flexibly combines the merits of the two Hahn and Jeruchim (1987); Heidelberger (1995). Each example is sampled with the probability of a pre-specified importance weight, and with the weight's inverse when accounting for the relative contribution in the overall loss. This helps to flexibly tune the representation of rare events during training, without biasing the data distribution Heidelberger (1995); Shimodaira (2000); Gretton et al. (2009). It is important to note that, poor choice of importance weights may result in uncontrolled variance that destabilizes training Robert and Casella (2013); Botev and Kroese (2008), calling for adaptive Rubinstein and Kroese (2013) or variance reduction schemes Rubinstein and Kroese (2016) to protect against such degeneracy.

Apart from the above strategies that fall within the standard empirical risk minimization framework, recent developments explicitly seek better generalization for the minority classes. One such example is the *one-class classification* that aims to capture one target class from a general population Tax (2002). *Meta-learning* and *few-shot learning* strategies instead trying to transfer the knowledge learned from data-rich classes to facilitate the learning of data-scarce classes Böhning et al. (2015); Finn et al. (2017). Additionally, non-cross-entropy based losses or penalties have been proved useful to imbalanced classification tasks (Weinberger and Saul, 2009; Huang et al., 2016; Oh Song et al., 2016). For instance, the Focal loss Lin et al. (2017) up-weights the harder examples, and Cao et al. (2019) introduced a label-distribution-aware margin loss encouraging minority classes having larger margins.

In this work, we present a novel solution called *variational inference for extremals*

(VIE), capitalizing on the learning of more generalizable representations for the minority classes. Our proposal is motivated by the observation that the statistical features of "rarity" have been largely overlooked in the current literature of rare-event modeling. And the uncertainties of rare-events are often not considered. Framed under the Variational Inference framework, we formulate our model with the assumption that the extreme presentation of (unobserved) latent variables can lead to the occurrence (or the inhibition) of rare events. This encourages the accurate characterization of the tail distribution of the data representation, which has been missed by prior work to the best of our knowledge. Building upon the state-of-the-art machine learning techniques, our solution features the following contributions: (i) the model accounts for representation uncertainty based on variational inference; (ii) the adoption of mixed Generalized Pareto priors to promote the learning of heavy-tailed feature representations; and (iii) integration of additive isotonic regression to disentangle representation and facilitate generalization. We demonstrate how our framework facilitates both model generalization and interpretation, with strong empirical performance reported across a wide-range of benchmarks.

2.2 Background

To simplify our presentation, we focus on the problem of rare event classification for binary outcomes. The generalization to the multiple-class scenario is simple and presented in the Appendix A.4.7. Let $D = \{x_i, y_i\}_{i=1}^N$ be a dataset of interest, where x_i and y_i denote predictors and outcomes, respectively, and N is the sample size. Without loss of generality, we denote $y = 1$ as the minority event label (indicating the occurrence of an event of interest), and $y = 0$ as the majority label.

In the following, we will briefly review the three main techniques we used in this work, namely, *variational inference* (VI), *extreme value theory* (EVT), and *additive isotonic regression*. VI allows for approximate maximum likelihood inference

while accounting for data uncertainty. EVT provides a principled and efficient way to model extreme, heavy-tailed representations. Additive isotonic regression further introduces monotonic constraints to *disentangle* the contribution of each latent dimension to the outcome.

2.2.1 Variational inference

Consider a latent variable model $p_\theta(v, z) = p_\theta(v|z)p(z)$, where $v \in \mathbb{R}^m$ is the observable data, $z \in \mathbb{R}^p$ is the unobservable latent variable, and θ represents the parameters of the likelihood model, $p_\theta(v|z)$. The marginal likelihood $p_\theta(v) = \int p_\theta(v, z)\,\mathrm{d}z$ requires integrating out the latent z, which typically, for complex distributions, does not enjoy a closed-form expression. This intractability prevents direct maximum likelihood estimation for θ in the latent variable setup. To overcome this difficulty, Variational Inference (VI) optimizes computationally tractable variational bounds to the marginal log-likelihood Kingma and Welling (2014); Chen et al. (2018). Concretely, the most popular choice of VI optimizes the following Evidence Lower Bound (ELBO):

$$\mathrm{ELBO}(v; p_\theta(v, z), q_\phi(z|v)) \triangleq \mathbb{E}_{Z\sim q_\phi(z|v)}\left[\log \frac{p_\theta(v, Z)}{q_\phi(Z|v)}\right] \leqslant \log p_\theta(v), \tag{2.1}$$

where $q_\phi(z|v)$ is an approximation to the true (unknown) posterior $p_\theta(z|v)$, and the inequality is a direct result of Jensen's inequality. The variational gap between the ELBO and true marginal log-likelihood, *i.e.*, $\log p_\theta(v) - \mathrm{ELBO}(v; p_\theta(v, z), q_\phi(z|v))$, is given by the Kullback–Leibler (KL) divergence between posteriors, *i.e.*,

$$\mathrm{KL}(q_\phi(z|v)||p_\theta(z|v)) = \mathbb{E}_{Z\sim q_\phi(z|v)}[\log q_\phi(Z|v)] - \mathbb{E}_{Z\sim q_\phi(z|v)}[\log p_\theta(Z|v)],$$

which implies that the ELBO tightens as $q_\phi(z|v)$ approaching the true posterior $p_\theta(z|v)$. For estimation, we seek parameters θ and ϕ that maximize the ELBO in (2.1).

Given a set of observations $\{v_i\}_{i=1}^N$ sampled from data distribution $v \sim p_d(v)$, maximizing the expected ELBO is also equivalent to minimizing the KL divergence $\mathrm{KL}(p_d(v) \parallel p_\theta(v))$ between the empirical and model distributions. When $p_\theta(v|z)$ and $q_\phi(z|v)$ are specified as neural networks, the resulting architecture is commonly known as the *variational auto-encoder* (VAE) Kingma and Welling (2014), where $q_\phi(z|v)$ and $p_\theta(v|z)$ and are known as *encoder* and *decoder*, respectively. Note that $q_\phi(z|v)$ is often used for subsequent inference tasks on new data.

2.2.2 Extreme value theory

Extreme Value Theory (EVT) provides a principled probabilistic framework for describing events with extremely low probabilities, which we seek to exploit for better rare event modeling. In particular, we focus on the *exceedance* models, where we aim to capture the asymptotic statistical behavior of values surpassing an extreme threshold Davison and Smith (1990); Tao et al. (2017), which we briefly review below following the notation of Coles et al. (2001). Without loss of generality, we consider exceedance to the right, *i.e.*, values greater than a threshold u. For a random variable X, the conditional cumulative distribution of exceedance level x beyond u is given by $F_u(x) = P(X - u \leqslant x | X > u) = \frac{F(x+u)-F(u)}{1-F(u)}$, where $x > 0$ and $F(x)$ denotes the cumulative density function for X.

A major result from EVT is that under some mild regularity conditions, *e.g.*, continuity at the right end of $F(x)$ and others, $F_u(x)$ will converge to the family of *Generalized Pareto Distributions* (GPD) regardless of $F(x)$, as u approaches the right support boundary of $F(x)$ Balkema and De Haan (1974); Pickands III et al. (1975), *i.e.*, $\lim_{u\to\infty} F_u(x) \xrightarrow{L_\infty} G_{\xi,\sigma,u}(x)$ Falk et al. (2010), where $\mathrm{GPD}_{\xi,\sigma,u}(x)$ is of the form

$$G_{\xi,\sigma,u}(x) = \begin{cases} 1 - [1 + \xi(x-u)/\sigma]^{-\frac{1}{\xi}}, & \text{if } \xi \neq 0 \\ 1 - \exp[-(x-u)/\sigma], & \text{if } \xi = 0 \end{cases} \tag{2.2}$$

where σ is a positive scale parameter. When $\xi < 0$ the exceedance x has bounded support $u \leqslant x \leqslant u - \sigma/\xi$, otherwise when $\xi \geqslant 0$, x ia unbounded on the right. A major implication of this asymptotic behavior is that, for modeling extreme values, one only needs to fit extreme samples to the log-likelihood function of the GPD.

2.2.3 Additive isotonic regression

Also known as monotonic regression, isotonic regression is a non-parametric regression model that constrains the relation between predictor and outcome to be monotonic, (*e.g.*, non-decreasing $f(a) \leqslant f(b)$ for $a \leqslant b$) Barlow et al. (1972). Such monotonic constraint is a natural and flexible extension to the standard linear relation assumed by many statistical models. To accommodate multi-covariate predictors, additive isotonic regression combines isotonic models for each individual one-dimensional predictor Bacchetti (1989). Standard implementations often involve specialized algorithms, such as local scoring algorithms Hastie (2017) and the alternating conditional expectation (ACE) method of Breiman and Friedman (1985). All these approaches typically require costly iterative computations and are not scalable to large datasets. Here we consider recent advances in unconstrained monotonic neural networks, which allow for efficient and flexible end-to-end learning of monotonic relations with robust neural nets based on standard training schemes such as stochastic gradient descent Sill (1998); Wehenkel and Louppe (2019).

2.3 Variational Inference of Extremals

The proposed model is based on the hypothesis that *extreme events are driven by the extreme values of some latent factors*. Specifically, we propose to recast the learning of low-prevalence events into the learning of extreme latent representations, thus amortizing the difficulties associated with directly modeling rare events as outcomes. To allow for more efficient learning from the rare events, we make some further

assumptions to regularize the latent representation: (i) *effect disentanglement*: the contribution from each dimension of the latent representation to the event occurrence is additive; (ii) *effect monotonicity*: there is a monotonic relation between the outcome likelihood and the values of each dimension of the latent representation. The key to the proposed approach is using an additive isotonic neural network to model the one-dimensional disentangled monotonic relations from a latent representation, which is obtained via variational inference. Specifically, we impose an EVT prior to explicitly capture the information from the few minority group samples into the tail behavior of the extreme representation. Below we provide the rationale for our choices followed by a description of all model components.

Disentanglement & additive isotonic regression. Consistent with assumptions (i) and (ii), we posit a scenario in which the underlying representation of extreme events is more frequent at the far end of the representation spectrum, for which additive isotonic regression is ideal. The disentanglement consists of modeling each latent dimension individually, thus avoiding the curse of dimensionality when modeling combinatorial effects with few examples. Further, the monotonicity constraint imposed by the isotonic regression model restricts possible effect relations, thereby improving generalization error by learning with a smaller, yet still sufficiently expressive, class of models Bacchetti (1989).

EVT & VI. Note that the spread of representation of extreme events is expected to be more uncertain relative to those of the normal, more abundant events, due to a few plausible causes: (i) extreme events represent the breakdown of system normality and are expected to behave in uncertain ways; (ii) there is only a small number of examples available for the extreme events, so the learned feature encoder will tend to be unreliable. As a result, we can safely assume that the encoded features associated with the extremes events will lie outside the effective support of the Gaussian distribution assumed by the standard VI model. In other words,

the representation of the events can manifest as a heavy-tailed distribution. This will compromise the validity and generalizability of a prediction model if not dealt with appropriately. So motivated, we explicitly model the distribution of the extreme underlying representations via EVT. Using EVT, we decouple the learning of the tail end of the representation distribution. Since EVT-based estimation only requires very few parameters, it allows for accurate modeling with a small set of tail-end samples. Further, in combination with the variational inference framework, it accounts for representation uncertainty via the use of a stochastic encoder, which further strengthens model robustness.

Benefits of heavy-tailed modeling. A few other considerations further justify modeling with a heavy-tailed distribution for the extreme event representation. One obvious benefit is that it allows better model resolution along the representation axis, *i.e.*, better risk stratification. For light-tail representations, extreme examples are clustered in a narrow region where the tail vanishes, thus a standard (light-tailed) learning model will report the average risk in that region. However, if the representations are more spread out, then there is a more gradual change in risk (Figure A.1 in Appendix A), which can be better captured, as shown in Figure 2.1. Another argument for favoring heavy-tailed representations is that heavy-tailed phenomena are very common in nature Bryson (1974), and these tail samples are often encoded less robustly due to the lack of training examples. Allowing long-tail representations relieves the burden of an encoder.

Model structure. We consider latent variable model $p_\theta(y, x, z) = p_\theta(y|z)p_\theta(x|z)p(z)$, where $v = \{x, y\}$ are the observed variables. Under the VI framework, similar to (2.1) we write the $\text{ELBO}(v; p_\theta(v, z), q_\phi(z|v))$ as

$$\mathbb{E}_{Z\sim q_\phi(z|v)}[\log p_\theta(y|Z)] + \mathbb{E}_{Z\sim q_\phi(z|v)}[\log p_\theta(x|Z)] - \text{KL}(q_\phi(z|v) \parallel p(z)) \tag{2.3}$$

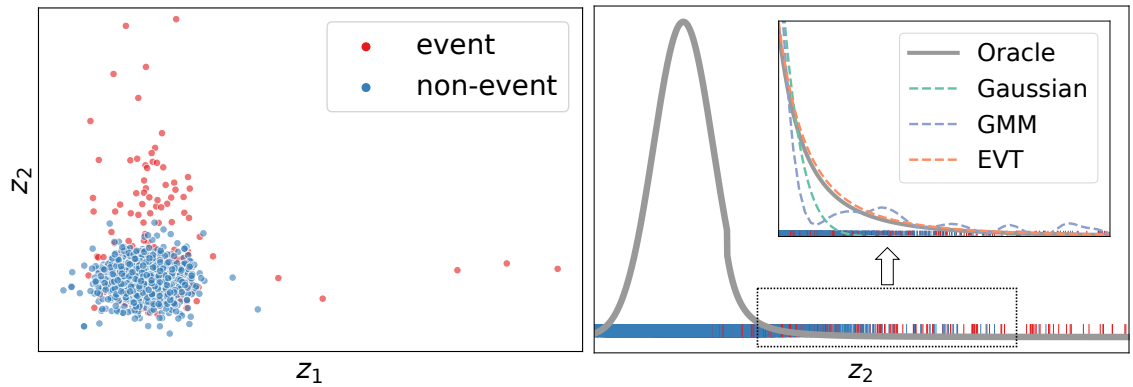

FIGURE 2.1: Left: Distribution of a two-dimensional latent space z where the long tail associates with higher risk. Right: Tail estimations with different schemes for the long-tailed data in one-dimensional space. EVT provides more accurate characterization comparing to other mechanisms.

where $p_\theta(y|z)$ is specified as an additive isotonic regression model, $p(z)$ is modeled with Gaussian GPD mixture, and the approximate posterior, $q_\phi(z|v)$, is specified as an inverse auto-regressive flow. Note that unlike in the standard ELBO in (2.3), we want to drop the term $\mathbb{E}_{Z\sim q_\phi(z|v)}[\log p_\theta(x|Z)]$ because we are not interested in modeling the covariates. Note this coincides with the *variational information bottleneck* (VIB) formulation Alemi et al. (2016). Additionally, the posterior $q_\phi(z|v)$ will not be conditioned on y, but only on x, because in practice, the labels y are not available at inference time. Specifically, we rewrite the objective in (2.3) as

$$\Psi_\beta(x, y; p_\theta(y|z), q_\phi(z|x)) = \mathbb{E}_{Z\sim q_\phi(z|x)}[\log p_\theta(y|Z)] - \beta \mathrm{KL}(q_\phi(z|x) \parallel p(z)), \tag{2.4}$$

where β is a hyperparameter controlling the relative contribution of the KL term to the objective. Below we provide details for each component of the proposed approach.

2.3.1 Decoder: additive monotonic neural network

First, let us consider the following monotone mapping $\int_l^z h(s;\theta)ds + \gamma$, consisting on integrating a non-negative function $h(s;\theta)$ specified as a neural network with one-dimensional input, s, and parameterized by θ. The choice of the lower end l is

arbitrary, and γ is a bias term. For multi-dimensional latent representation $z \in \mathbb{R}^p$, we write the additive monotonic neural network (AMNN) as

$$H(z;\theta) = \sum_{j}^{p}[\alpha_j \int_{l}^{z_j} h_j(s;\theta)ds] + \gamma, \tag{2.5}$$

where α_j serves as a weight which controlling the effect directions. In other words, when $\alpha_j > 0$, it can be interpreted as an event stimulator; otherwise it is an event blocker. To ensure $h(s;\theta)$ is non-negative, we apply exponential activation function to the network's output. The integration of z is conducted with numerical integration by the Riemann-Stieltjes method Davis and Rabinowitz (2007).

Based on (2.5), we have $\log p_\theta(y|z) = \ell_{\text{CLL}}(y, H(z;\theta))$, where

$$\ell_{\text{CLL}}(y,a) = \log\{\mathbb{1}_{y=1}(y)(1-\exp(-\exp(a))) + \mathbb{1}_{y=0}(y)\exp(-\exp(a))\},$$

is the complementary log-log (CLL) link, where $\mathbb{1}_{(\cdot)}$ is the indicator function. We prefer CLL over the standard logistic link since the CLL link is more sensitive at the tail end Aranda-Ordaz (1981).

2.3.2 Latent extreme prior: Gaussian-GPD mixture

To better capture the tail behavior of the latent representation, we assume random variable $Z \sim p(z)$ is a mixture of a standard Gaussian distribution truncated at u and a GPD for modeling the tail end thresholded at u, *i.e.*, $F(z) = \Phi(z)$ when $z \leqslant u$ and $F(z) = \Phi(u) + (1-\Phi(u))G_{\xi,\sigma}(z-u)$ when $z > u$, where $\Phi(z)$ denotes the CDF of a standard Gaussian distribution. Note that for $z > u$, $F(z)$ can be expressed as a GPD with parameters $(\tilde{\xi}, \tilde{\sigma}, \tilde{u})$ McNeil (1997), where $\tilde{\xi} = \xi$ and if $\xi \neq 0$, $\tilde{\sigma} = \sigma(1-\Phi(u))^{\xi}$ and $\tilde{u} = u - \tilde{\sigma}((1-\Phi(u))^{-\xi} - 1)/\xi$. Otherwise, when $\xi = 0$, $\tilde{\sigma} = \sigma$ and $\tilde{u} = u + \tilde{\sigma}\log(1-\Phi(u))$. Consequently, the CDF for the mixed GPD is given by

$$F(z) = \mathbb{1}_{(-\infty,u]}(z)\Phi(z) + \mathbb{1}_{(u,\infty)}(z)G_{\tilde{u},\xi,\tilde{\sigma}}(z). \tag{2.6}$$

For simplicity, we denote the set of parameters in GPD as ψ=$\{\xi_{\text{GPD}},\sigma_{\text{GPD}}\}$ and the threshold u is a user-defined parameter. In the experiments we set u to $\Phi^{-1}(0.99)$.

2.3.3 Latent posterior: Inverse Autoregressive Flow

Considering we have adopted a long-tailed GPD prior, we seek a posterior approximation $q_\phi(z|x)$ that is: (i) a flexible parameterization to approximate arbitrary distributions; and (ii) with a tractable likelihood to be able to evaluate the $\text{KL}(q_\phi(z|x) \parallel p(z))$ exactly. We need (i) because the true posterior is likely to exhibit heavy-tailed behavior due to the extended coverage of the GPD prior, and (ii) is to ensure accurate and low-variance Monte Carlo estimation of the KL-divergence at the tail end of the prior. These requirements invalidate some popular choices, *e.g.*, a standard Gaussian posterior is light-tailed, and the implicit neural-sampler-based posterior typical in the work of adversarial variational Bayes Mescheder et al. (2017), does not have a tractable likelihood.

One model family satisfying the above two requirements is known as the generative flow Rezende and Mohamed (2015), where simple invertible transformations with tractable log Jacobian determinants are stacked together, transforming a simple base distribution into a complex one, while still having closed-form expressions for the likelihood. In this work, we consider the *inverse autoregressive flow* (IAF) model Kingma et al. (2016). The flow chain is built as

$$z_t = \mu_t + \sigma_t \odot z_{t-1}, \text{ for } 1 \leqslant t \leqslant T, \tag{2.7}$$

where $\mu_t \in \mathbb{R}^p$ and $\sigma_t \in \mathbb{R}^p$ are learnable parameters, $\odot$ denotes the element-wise product, z_0 is typically drawn from a p-dimensional Gaussian distribution, $z_0 \sim \mathcal{N}(\mu_0, \text{Diag}(\sigma_0^2))$ where μ_0 and σ_0 are obtained from an initial encoder defined by a neural network given input x with parameter ϕ. A sample from the posterior $q_\phi(z|x)$ is given by z_T, obtained by "flowing" z_0 through (2.7). Provided the Jacobians $\frac{d\mu_t}{dz_{t-1}}$

and $\frac{d\sigma_t}{dz_{t-1}}$ are strictly upper triangular Papamakarios et al. (2017), we obtain the following closed-form expression for the log posterior

$$\log q(z|x) = \log q(z_0|x) - \sum_{t=1}^{T} \log\det \left| \frac{dz_t}{dz_{t-1}} \right| \tag{2.8}$$

$$= -\sum_{j=1}^{p} \left(\frac{1}{2} e_j^2 + \frac{1}{2}\log(2\pi) + \sum_{t=0}^{T} \log \sigma_{t,j} \right),$$

where $e_j = (x_j - \mu_{0,j})/\sigma_{0,j}$ for the jth dimension.

Table 2.1: Ablation study of VIE with different combinations of architectures on realistic synthetic datasets with 1% event rate. The oracle model has used the ground-truth model parameters to predict.

					Average AUC (standard deviation)		Average AUPRC (standard deviation)	
	Prior	Encoder	Decoder	Prior Match	n=5k	n=20k	n=5k	n=20k
VAE	Gaussian	Gaussian	MLP	True	0.552 (0.092)	0.674 (0.020)	0.026 (0.010)	0.061 (0.017)
VAE-GPD	mixed GPD	Gaussian	AMNN	False	0.569 (0.062)	0.653 (0.027)	0.021 (0.003)	0.035 (0.013)
IAF-GPD	mixed GPD	IAF	AMNN	False	0.511(0.021)	0.665 (0.029)	0.017(0.002)	0.025 (0.008)
Fenchel-GPD	mixed GPD	Implicit	AMNN	True	0.623 (0.036)	0.694 (0.021)	0.037 (0.010)	0.062 (0.026)
VIE	mixed GPD	IAF	AMNN	True	**0.684** (0.031)	**0.701** (0.017)	**0.050** (0.009)	**0.079** (0.025)
Oracle (with 90% confidence interval)					0.704 [0.662, 0.751]		0.092 [0.058, 0.141]	

2.3.4 Posterior matching with Fenchel mini-max learning

We consider an additional modification that explicitly encourages the match of the aggregated posterior $q_\phi(z) = \int q_\phi(z|x)p_d(x)\,\mathrm{d}x$ to the prior $p(z)$, which has been reported to be vastly successful at improving VAE learning Mescheder et al. (2017). In our case, $q_\phi(z)$ does not have a closed-form expression for the likelihood ratio of the KL formulation, which motivates us to use a sample-based estimator. We consider the mini-max KL estimator based on the Fenchel duality Tao et al. (2019); Dai et al. (2018). Concretely, recall the KL can be expressed in its Fenchel dual form[1]

$$\Gamma(p, q_\phi, \nu) = \mathbb{E}_{Z\sim q_\phi(z)}[\nu(Z)] - \mathbb{E}_{Z'\sim p(z)}[\exp(\nu(Z'))]$$

$$\mathrm{KL}(q_\phi(z) \parallel p(z)) = \max_{\nu\in\mathcal{F}} \Gamma(p, q, \nu), \tag{2.9}$$

[1] We have removed the constant term for notational clarity.

where $\nu(z)$ is commonly known as the critic function in the adversarial learning literature, and we maximize wrt $\nu(z)$ in the space of all functions $\mathcal{F}$, modeled with a deep neural network. We use (2.4) and (2.9) to derive an augmented ELBO that further penalizes the discrepancy between the aggregated posterior and the prior, *i.e.*, $\Psi_\beta(x, y; p_\theta(y|z), q_\phi(z|x)) - \lambda \mathrm{KL}(q_\phi(z) \parallel p(z))$, where λ is a regularization hyperparameter Chen et al. (2018). Solving for this objective results in the following mini-max game

$$\max_{\theta,\phi} \min_{\nu} \Psi_\beta(x, y; p_\theta(y|z), q_\phi(z|x)) - \lambda \Gamma(p_\theta, q_\phi, \nu), \tag{2.10}$$

where β and λ are regularization hyperparameters. In a similar vein to β-VAE and adversarial variational Bayes (AVB), our objective leverages β, $\lambda > 0$ to balance the prediction accuracy and the complexity of the latent representation via KL regularization. Further, from $\Psi_\beta(x, y; p_\theta(y|z), q_\phi(z|x))$ in (2.4), note that the decoder $p_\theta(y|z)$ is obtained from the additive neural network in (2.5), $p_\psi(z)$ is the Gaussian GPD mixture with CDF in (2.6), $q_\phi(z|x)$ is the autoregressive flow implied by (2.7) and $\nu(z; \omega)$ is the critic function specified as a neural network and parameterized by ω.

To avoid collapsing to suboptimal local minima, we train the encoder arm more frequently to compensate for the detrimental posterior lagging phenomenon He et al. (2019). The pseudo-code for the proposed VIE is summarized in Algorithm 1 and

detailed architecture can be found in the Appendix A.2.

Algorithm 1: Variational Inference with Extremals.

Data: $\mathcal{D} = (x, y)$. x: inputs, y: labels
Networks and parameters:
· $\texttt{Init-Encoder}(x, \epsilon; \phi)$: Initial encoder network;
· $\texttt{IAF}(z; \phi)$: recursive autoregressive neural network;
· $\nu(z; \omega)$: critic neural network;
· $\texttt{AMNN}(z; \theta)$: additive monotonic neural net;
· prior: $p_\psi(z) = \texttt{MixedGPD}(z;\psi,u)$, $\psi = \{\xi_{\text{GPD}}, \sigma_{\text{GPD}}\}$
Initialize: $\texttt{Init-Encoder}$, $\texttt{IAF}$, ν, $\texttt{AMNN}$, ψ
for *iteration* $k \in \{1, \ldots, K\}$ **do**
  Sample $\{(x_i, y_i)\}_{i=1}^m$ from $\mathcal{D}$, $\{\epsilon_i\}_{i=1}^m$ from $p(\epsilon)$
  $[\mu_0, \sigma_0] = \texttt{Init-Encoder}(x, \epsilon; \phi)$
  Sample $z_{\texttt{pr}}$ from $p_\psi(z)$, $\boldsymbol{z}_0$ from $\mathcal{N}(\mu_0, \Sigma_0)$
  Compute $l_{\texttt{post}} := \log q_\phi(z_0|x)$
  for *step* $t \in \{1, \ldots, T\}$ **do**
    $[\mu_t, \sigma_t] = \texttt{IAF}(z_{t-1}; \phi)$, $z_t = \mu_t + \sigma_t \odot z_{t-1}$
    $l_{\texttt{post}} = l_{\texttt{post}} - \sum(\log \sigma_t)$
  end
  $\log p_\theta(y|z_T) = \ell_{\text{CLL}}(y, \text{AMNN}(z_T; \theta))$
  Descend ω by $\nabla_\omega \frac{1}{m} \sum[\nu_\omega(z_{\texttt{pr}}) - \log \nu_\omega(z_T)]$
  Ascend $\Omega = \{\phi, \psi, \theta\}$ by
  $\nabla_\Omega \frac{1}{m} \sum[\log p_\theta(y|z_T) - \log \nu_\omega(z_T) - \text{KL}]$, where $\text{KL} = l_{\texttt{post}} - \log p_\psi(z_T)$
end

2.4 Related Work

Rare-event modeling with regression. Initiated by King and Zeng (2001), the discussion on how to handle the unique challenges presented by rare-event data for regression models has attracted extensive research attention. The statistical literature has mainly focused on bias correction for sampling Fithian and Hastie (2014) and estimation Firth (1993), driven by theoretical considerations in maximum likelihood estimation. However, their assumptions are often violated in the face of modern datasets Sur and Candès (2019), characterized by high-dimensionality and complex interactions. Our proposal approaches a solution from a representation learning perspective Bengio et al. (2013), by explicitly exploiting the statistical regularities of

extreme values to better capture extreme representations associated with rare events.

Re-sampling and loss correction. Applying statistical adjustments during model training is a straightforward solution to re-establish balance, but often associated with obvious caveats. For example, the popular down-sampling and up-sampling He and Garcia (2009) discard useful information or introduce artificial bias, exacerbating the chances of capturing spurious features that may harm generalization Drummond et al. (2003); Cao et al. (2019), and their performance gains may be limited Byrd and Lipton (2019). While traditionally tuned by trial and error, recent works have explored automated weight adjustments Lin et al. (2017); Zhang et al. (2020) , and principled loss correction that factored in class-size differences Cui et al. (2019). Our contribution is orthogonal to these developments and promises additional gains when used in synergy.

Transferring knowledge from the majority classes. Adapting the knowledge learned from data-rich classes to their under-represented counterparts has shown success in few-shot learning, especially in the visual recognition field Wang et al. (2017); Xiu et al. (2020) , and also in the clinical settings Böhning et al. (2015). However, their success often critically depends on strong assumptions, the violation of which typically severely undermines performance Wang et al. (2020). Related are the one-class classification (OCC) models Tax (2002), assuming stable patterns for the majority over the minority classes. Our assumptions are weaker than those made in these model categories, and empirical results also suggest the proposed VIE works more favorably in practice (see experiments).

2.5 Experiments

We carefully evaluate the proposed VIE on a diverse set of realistic synthetic data and real-world datasets with different degrees of imbalance. Our implementation is based on PyTorch, and code to replicate our experiments are available from `https:`

	Average AUC						Average AUPRC					
	COVID		InP		SEER		COVID		InP		SEER	
Event category	Mortality	Combined	12h	168h	3mo	11mo	Mortality	Combined	12h	168h	3mo	11mo
LASSO	0.856	0.853	0.822	0.760	0.888	0.845	0.235	**0.542**	0.092	0.216	0.140	0.309
MLP	0.862	0.854	0.824	0.768	0.885	0.856	0.225	0.531	0.093	0.221	0.169	0.322
DeepSVDD	NA	NA	0.633	0.551	0.592	0.572	NA	NA	0.020	0.063	0.026	0.068
IW	0.856	0.860	0.776	0.728	0.798	0.832	0.193	0.511	0.073	0.165	0.123	0.274
Focal	0.829	0.854	0.750	0.705	0.868	0.835	0.238	0.484	0.044	0.149	0.141	0.263
LDAM	0.857	0.843	0.819	0.774	0.893	0.755	0.202	0.535	0.086	0.197	0.177	0.332
VIE	**0.883**	**0.867**	**0.840**	**0.780**	**0.895**	**0.862**	**0.268**	0.535	**0.100**	**0.240**	**0.189**	**0.345**

Table 2.2: Average AUC and AUPRC from real-world datasets.

`//github.com/ZidiXiu/VIE/`. We provide additional experiments and analyses in the Appendix A.4.

Baseline Models We consider the following set of competing baselines to compare the proposed solution: LASSO regression Tibshirani (1996), MLP with re-sampling and re-weighting (MLP), Importance-Weighting model (IW) Byrd and Lipton (2019), FOCAL loss Lin et al. (2017), Label-Distribution-Aware Margin loss (LDAM) Cao et al. (2019), and SVD based one-class classification model (Deep-SVDD) Ruff et al. (2018). We evaluate baseline models performance on test set with the best performed hyper-parameters on validation dataset. For detailed settings please refer to the Appendix A.4.

Evaluation Metrics To quantify model performance, we consider AUC and AUPRC. AUC is the area under the Receiver Operating Characteristic (ROC) curve, which provides a threshold-free evaluation metric for classification model performance. AUC summarizes the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR). AUPRC evaluates the area under Precision-Recall (PR) curve and summarizes the trade-off between TPR and True Predictive Rate. We discuss other metrics in the Appendix A.2. In simulation studies, we repeat simulation ten times to obtain empirical AUC and AUPRC confidence intervals. For real world datasets, we applied bootstrapping to estimate the confidence intervals.

Ablation study for VIE

VIE applies a few state-of-art techniques in variational inference in order to achieve optimal performance. In this section, we decouple their contributions via an ablation study, to justify the necessity of including those techniques in our final model. To this end, we synthesize a semi-synthetic dataset based on the Framingham study Mitchell et al. (2010), a long-term cardiovascular survival cohort study. We use a realistic model to synthesize data from the real-world covariates under varying conditions, *i.e.*, different event rates, sample size, non-linearity, *etc.* More specifically, we use the CoxPH-Weibull model Bender et al. (2005) to simulate the survival times of patients $T = \{\frac{-\log U}{\lambda \exp(g(x))}\}^{1/\nu}$, where $g(x)$ is either a linear function or a randomly initialized neural net. Our goal is to predict whether the subject will decease within a pre-specified time frame, *i.e.*, $T < t_0$. Via adjusting the cut-off threshold t_0, we can simulate different event rates. A detailed description of the simulation strategy is in Appendix A.3.

We experiment with different combinations of advanced VI techniques, as summarized in Table 2.1. Limited by space, we report results at 1% event rate with $g(\cdot)$ set to a randomly initialized neural network under various sample sizes. Additional results on linear models and other synthetic datasets are consistent and can be found in Appendix A.3. IAF and GPD only variants perform poorly, even compared to the vanilla VAE solution. This is possibly due to the fact that priors are mismatched. Explicitly matching to the prior via Fenchel mini-max learning technique improves performance. However, without using an encoder with a tractable likelihood, the model cannot directly leverage knowledge from the GPD prior likelihood. Stacked together (mixed GPD+IAF+Fenchel), our full proposal of VIE consistently outperforms its variants, approaching oracle performance in the large sample regime.

Table 2.3: Summary statistics for real-world datasets.

	COVID	InP	SEER
sample size	25,315	67,655	68,082
dimension	1268(668)	73(39)	789(771)
event rate (%)	2.6%, 8%	$1 \sim 5\%$	$1 \sim 5\%$

Real-world datasets

To extensively evaluate real-world performance, we consider a wide range of real-world datasets, briefly summarized below: (*i*) `COVID`: A dataset of patients admitted to the DUHS with positive COVID-19 testing, to predict death or use of a ventilator. (*ii*) `InP` O'Brien et al. (2020): An in-patient data from DUHS, to predict the risk of death or ICR transfers. (*iii*) `SEER` Ries et al. (2007): A public dataset studying cancer survival among adults curated by the U.S. Surveillance, Epidemiology, and End Results (SEER) Program, here we use a 10-year follow-up breast cancer subcohort. Summary statistics of these three real-world datasets are given in Table 2.3. Note that `InP` and `SEER` are survival datasets, specifically `SEER` includes censored subjects. We follow the data pre-processing steps in Xiu et al. (2020). To create outcome labels, we set a cut-off time to define an event of interest the same as in the ablation study, and exclude subjects censored before the cut-off time (less than 0.2%). Datasets have been randomly split into training, validation, and testing datasets with ratio 6:2:2. See Appendix A.4 for details on data pre-processing.

Table 2.2 compares VIE to its counterparts, where the numbers are averaged over the bootstrap samples. We see the proposed VIE yields the best performance in almost all cases, and the lead is more significant with low event rates. Note that the one-class classification based DeepSVDD performs poorly, which implies treating rare events as outliers are inappropriate. Re-weighting and resampling based methods (IW, Focal) are less stable compared to those simple baselines (LASSO,

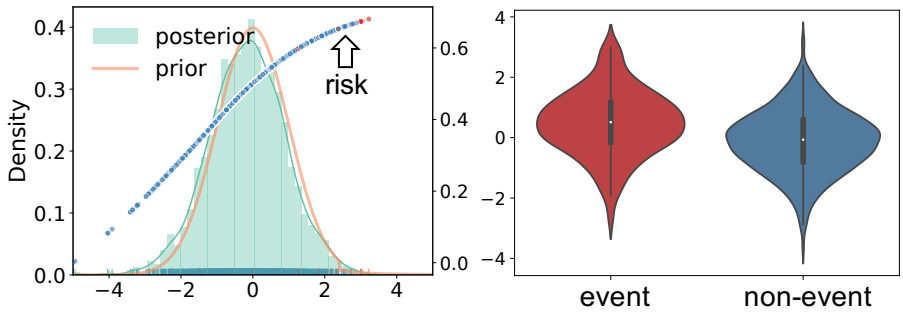

FIGURE 2.2: First latent dimension from the `InP` dataset (1% event rate). Left: Learned prior and posterior distribution, and monotonic predicted risks (right axis). Right: The latent representation values distribution grouped by event type.

MLP). The theoretically optimal LDAM works well in general, second only to VIE in most settings. To further demonstrate the stability of our method, we visualize the bootstrapped evaluation scores for the COVID dataset in Figure 2.3, and defer the additional cross-validation results to the Appendix A.3. We see that VIE leads consistently.

We also verify empirically that the estimated GPD shape parameters ξ_{GPD} are mostly positive (see the Appendix), indicating heavier than Gaussian tails as we have hypothesized. In Figure 2.2, we visualize one such latent dimension from the `InP` dataset, along with the associated risk learned by AMNN. In this example, the tail part is heavier than Gaussian and is associated with elevated risk. See our Appendix A.4 for examples where the extended tail contributes to prohibit the event.

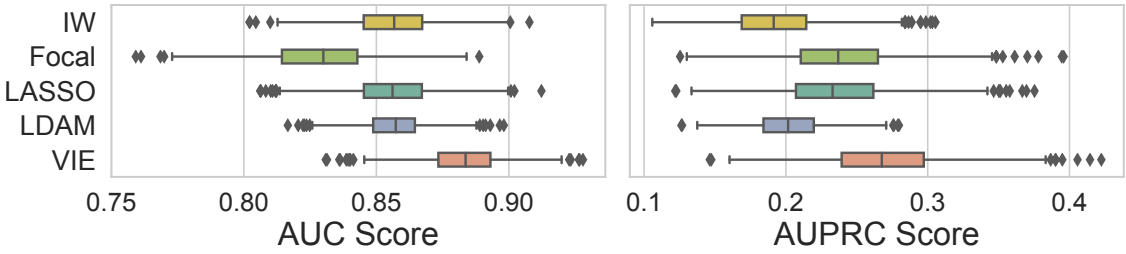

FIGURE 2.3: Bootstrapped AUC (left) and AUPRC (right) distributions for the `COVID` mortality data (2.6% event rate).

2.6 Conclusions

Motivated by the challenges of rare-event prediction in clinical settings, we presented Variational Inference with Extremals (VIE), a novel extreme representation learning-based variational solution to the problem. In this model we leveraged GPD to learn the extreme distributions with few samples and applied additive monotonic neural networks to disentangle the latent dimensions' effects on the outcome. VIE featured better generalization and interpretability, evidenced by strong performance on synthetic and real datasets.

Acknowledgements

The authors would like to thank the Duke Institute for Health Innovation (DIHI) for providing access to curated COVID-19 data and outcomes. This research was supported in part by NIH/NIDDK R01-DK123062, NIH/NIBIB R01-EB025020, NIH/NINDS 1R61NS120246-01, DARPA, DOE, ONR and NSF.

3

Supercharging Imbalanced Data Learning with Energy-based Contrastive Representation Transfer

3.1 Introduction

Learning with imbalanced datasets is a common yet still very challenging scenario in many machine learning applications. Typical scenarios include: (*i*) rare events, where the event prevalence is extremely low while their implications are of high cost, *e.g.*, severe risks that people seek to avert (King and Zeng, 2001); (*ii*) emerging objects in a dynamic environment, which call for quick adaptation of an agent to identify new cases with only a handful target examples and plentiful past experience (Fei-Fei et al., 2006). A typical scenario in natural datasets is that the occurrence of different objects follows a power law distribution. And in many situations, the accurate identification of those rarer instances bears more significant social-economic values, *e.g.*, fraud detection (Dal Pozzolo et al., 2017), driving safety (Dingus et al., 2016), nature conservation (Kaggle, 2017), social fairness (Chouldechova and Roth, 2018), and public health (Machado and Lopes, 2020).

Notably, severe class imbalance and lack of minority labels are the two major difficulties in this setting, which render standard learning strategies unsuitable (Kot-

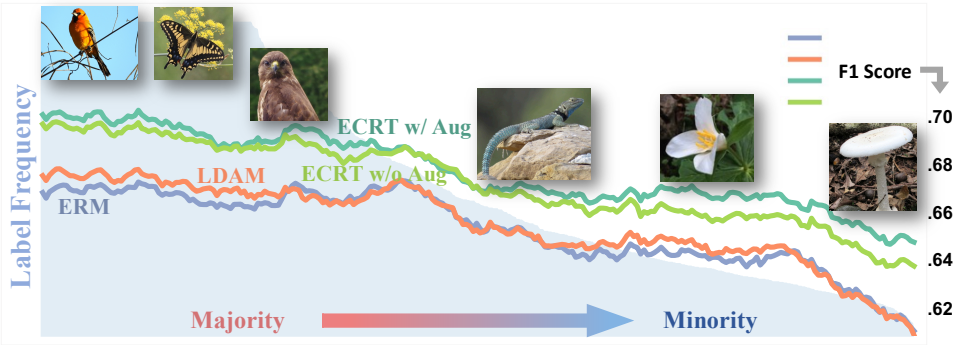

FIGURE 3.1: Energy-based causal representation transfer alleviates tail imbalance for natural image classification. The shaded region denotes label frequency in the `iNaturalist` data, with some representative images shown. Solid curves are for label-conditional F1 scores (higher is better) for the proposed *energy-based causal representation transfer* (ECRT), LDAM (state of the art) and the standard ERM. ECRT consistently outperforms the others, especially in the low-sample regime.

siantis et al., 2006). Without explicit statistical adjustments, the imbalance induces bias towards the majority classes. On the other hand, the lack of minority representations prevents the identification of stable correlations that generalize in predictive settings. In addition, due to technological advancements, more and higher dimensional data are routinely collected for analysis. This also inadvertently exacerbates the issue of minority modeling as there is an excess of predictors relative to the limited occurrence of minority samples.

Various research efforts have been directed to address the above issues, with class re-balancing being the most popular heuristic. The two most prominent directions in the category include statistical *resampling* and *reweighting*. Resampling alters the exposure frequency during training, *i.e.*, more for the minorities and less so for the majorities (Drummond et al., 2003). Alternatively, reweighting directly amends the relative importance of each sample based on their class (Cui et al., 2019), sampling frequency (Botev and Kroese, 2008) and associated cost for misclassification (Elkan, 2001). Recent developments also have considered class-sensitive and data-adaptive losses to more flexibly offset the imbalance (Cao et al., 2019; Lin et al., 2017). While being intuitive and working reasonably well in practice, an important is that these

approaches offer no protection against the over-fitting of minority instances (see Figure 3.1), a fundamental obstacle towards better generalization.

To circumvent this key limitation, inductive biases are often solicited to impose strong constraints that suppress spurious correlations that hinder out-of-sample generalization. A classical strategy is few-shot learning (Fei-Fei et al., 2006; Wang et al., 2020), where the majority examples are used to train meta-predictors and transferable features, leaving only a few parameters to tune for the minority data. In anomaly detection, methods such as one-class classification instead regard minority classes as outliers that do not always associate with stable, recognizable patterns (Moya and Hush, 1996; Ruff et al., 2018). Despite their relatively strong assumptions, these methods capitalize on their superior ability in generalizing in the low sample regime in empirical settings.

More recently, establishing causal invariance has emerged as a new, powerful learning principle for better generalization ability even under apparent distribution shifts (Rojas-Carulla et al., 2018). In contrast to standard empirical risk minimization (ERM) schemes, where the generalization to similar data distributions is considered, causally-inspired learning instead embraces robustness against potential perturbations (Arjovsky et al., 2019). This is achieved via only attending to causally relevant features and associations postulated to be invariant under different settings (Peters et al., 2016). Specifically, contributions from spurious, unstable features are effectively blocked or attenuated. Interestingly, compromise of performance can be expected in those models (Rothenhäusler et al., 2018), as a direct consequence of discarding useful (but non-causal) correlations in exchange for better causal generalization. Recently, (Hyvarinen et al., 2019) proposed non-parametric causal disentanglement of data representation to identify invariant relations.

This work explores the advancement of imbalanced data learning via adopting causal perspectives, with the insight that the key algorithm can be reformulated as

an energy-based contrastive learning procedure to drastically improve efficiency and flexibility. In recognition of the limitations discussed above, we present *energy-based causal representation transfer* (ECRT): a novel imbalanced learning scheme that brings together ideas from causality, contrastive learning, energy modeling, data-augmentation and weakly-supervised learning, to address the identified weakness of existing solutions. Our key contributions are: (i) a causal representation encoder informed by an invariant generative mechanism based on generalized contrastive learning; (ii) integrated data-augmentation and source representation regularization techniques exploiting feature independence to enrich minority representations that better balance the trade-off between utility and invariance; (iii) a key novelty is the derivation of an energy-based contrastive learning algorithm that greatly enhance model parallelism for large-label settings and extends generalized contrastive learning; and (iv) insightful discussions on the justifications for the use of proposed approach. Our claims are supported by strong experimental evidence.

3.2 Preliminaries

Notation and problem definition. We use $\boldsymbol{x} \in \mathbb{R}^p$ to denote the input data and $\boldsymbol{z} \in \mathbb{R}^d$ for the predictive features extracted from $\boldsymbol{x}$. Let $y \in \{1, \ldots, M\}$ be the class label. The number of training samples and those with label m are denoted as n and n_m, respectively. We use $\mathbb{E}[\cdot]$ to denote the expectation (average) of an empirical distribution, $\boldsymbol{a}_i$ to denote a vector associated with the i-th sample, and $[\boldsymbol{a}]_b$ to denote the b-th entry of vector $\boldsymbol{a}$. For simplicity, we assume class M is the minority class, *i.e.*, $n_m \gg n_M$, for $m \in \{1, \ldots, M-1\}$. Throughout, we refer to $\mathcal{X}, \mathcal{Z}$ and $\mathcal{S}$ as data, feature, and source spaces, respectively, as shown in Figure 3.2, and defined below. Our goal is to accurately predict the label of minority instances with very limited training examples of it. Generalization to multiple minority categories is straightforward.

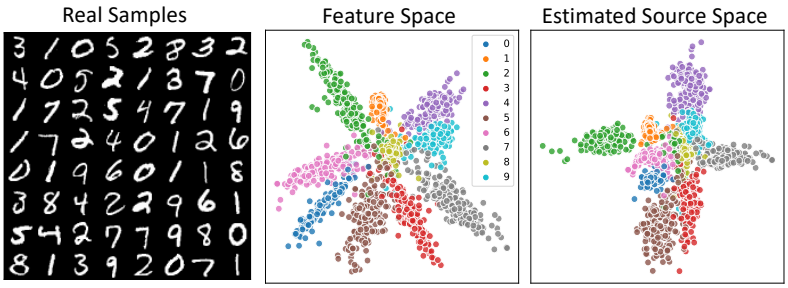

FIGURE 3.2: Illustration of data space $\mathcal{X}$, feature space $\mathcal{Z}$ (predictive but entangled) and source space $\mathcal{S}$ (independent/disentangled) representations identified by ECRT for `MNIST`.

Generalized contrastive learning and ICA The proposed approach is based on a generalized form of *independent component analysis* (ICA), which addresses the inverse problem of signal *disentanglement* (Hyvärinen and Oja, 2000). Specifically, ICA decorrelates features $\boldsymbol{Z}$ of the observed signal $\boldsymbol{X}$ into a source signal representation $\boldsymbol{S} = f_\psi(\boldsymbol{Z})$, where $f_\psi(\boldsymbol{z})$ is a smooth and invertible mapping known as the *de-mixing* function, and while assuming that the components of $\boldsymbol{S}$ are statistically independent, *i.e.*, with density $q(\boldsymbol{s}) = \prod_j q_j([\boldsymbol{s}]_j)$. Notationally, we call $[\boldsymbol{s}]_j$ the j-th *independent component* (IC) of $\boldsymbol{Z}$. While *nonlinear ICA* (NICA) is generally infeasible (Comon, 1994; Hyvärinen and Pajunen, 1999), (Hyvarinen et al., 2019) has recently proposed a setting in which the identification of NICA can be achieved, by requiring an additional auxiliary label y. Specifically, NICA assumes that source signals are conditionally independent given y, *i.e.*, $q(\boldsymbol{s}|y) = \prod_j q_j([\boldsymbol{s}]_j|y)$, then $f_\psi(\boldsymbol{z})$ can be identified using *generalized contrastive learning* (GCL), whose implementation is detailed below.

GCL solves a generalized regression problem in which a *critic function* predicts whether label y and representation $\boldsymbol{z}$ are correctly paired, *i.e.*, *congruent*. We call $(y_i, \boldsymbol{z}_i)$ a congruent pair and $(y_j, \boldsymbol{z}_i)$ an incongruent pair if $i \neq j$. Specifically, the critic is defined as $r_\nu(y, \boldsymbol{z}) = \sum_{a=1}^{d} \gamma_\nu^a(y, [\boldsymbol{s}]_a)$, where $\gamma_\nu^a(\cdot, \cdot)$ is a neural network with parameters ν, whose inputs are the label y and the a-th coordinate of $\boldsymbol{s} = f_\psi(\boldsymbol{z})$,

denoted as $[\boldsymbol{s}]_a$. Then, GCL optimizes the following objective:

$$\arg\min_{f_\psi, r_\nu} \underbrace{\mathbb{E}_i[h(-r_\nu(y_i, \boldsymbol{z}_i))] + \mathbb{E}_{j\neq i}[h(r_\nu(y_j, \boldsymbol{z}_i))]}_{\mathcal{L}_{\mathrm{GCL}}(f_\psi, r_\nu)}, \tag{3.1}$$

where $h(r) = \log(1 + \exp(r))$ is the softplus function. (3.1) seeks to optimize $f_\psi(\cdot)$ and $r_\nu(\cdot, \cdot)$ by maximizing the discriminative power to tell apart congruent and incongruent pairs.

In fact, an interesting result showed by (Hyvarinen et al., 2019) revealed that maximizing the ability of the critic function $r_\nu(\cdot, \cdot)$ for correctly identifying matching pairs $(\boldsymbol{z}, y)$ leads to the identification (up to univariate transformation) of $f_\psi(\cdot)$, such that the components of $\boldsymbol{s}$ are conditionally independent given y. See the Appendix B.1 for a formal exposition.

3.2.1 Invertible neural network

The recent interest in generative modeling has popularized the use of *invertible neural networks* (INN) in machine learning, with prominent examples such as normalizing flows (Rezende and Mohamed, 2015) and neural ordinary differential equations (ODEs) (Chen et al., 2018). Unlike standard neural networks, an INN seeks to establish a one-to-one mapping between the input and output domains, *i.e.*, the forward map $\boldsymbol{s} = f_\psi(\boldsymbol{z})$ as well as the corresponding inverse map $\boldsymbol{z} = f_\psi^{-1}(\boldsymbol{s})$. Standard constructions of INN achieve representational flexibility by stacking simple invertible transformations. In practice, the efficiency of the forward or inverse passes are often trade-off depending on the application needs (Papamakarios et al., 2017). Here we aim for fast forward computations, thus adopt the *masked auto-regressive flow* (MAF) design for our INN (Papamakarios et al., 2017), which allows efficient parallelization of the autoregressive architecture via *causal masking*.

Let $\{\boldsymbol{z}^t\}_{t=0}^T$ be a flow of length T, in which $\boldsymbol{z}^0 = \boldsymbol{z}$, $\boldsymbol{z}^{t+1} = F_t(\boldsymbol{z}^t)$, $\boldsymbol{s} = \boldsymbol{z}^T$, and we let $f_\psi = F_1 \circ F_2 \cdots \circ F_{T-1}$. Specifically, the MAF is constructed as a series of

shift and scale transformations $\boldsymbol{z}^{t+1} = F_t(\boldsymbol{z}^t) \triangleq \boldsymbol{a}_t(\boldsymbol{z}^t) \odot \boldsymbol{z}^t + \boldsymbol{b}_t(\boldsymbol{z}^t)$, where $\odot$ is the element-wise product, and $\boldsymbol{a}_t(\cdot)$ and $\boldsymbol{b}_t(\cdot)$ are vector transformations known as scale and shift, respectively, that follow a causal autoregressive structure, *i.e.*, that $[\boldsymbol{z}^{t+1}]_k$ only depends on $[\boldsymbol{z}^t]_{<k}$. As a direct consequence of this structure, the MAF-based INN results in a tractable Jacobian $J(\boldsymbol{z}) \triangleq |\det(\nabla_{\boldsymbol{z}} f_\psi)|$ of $f_\psi(\boldsymbol{z})$ that facilitates likelihood computations. In fact, assuming the sources $\boldsymbol{S}$ have prior density $p(\boldsymbol{s})$, the likelihood of the features $p(\boldsymbol{z})$ of the MAF specification is given by (Papamakarios et al., 2017):

$$\log p(\boldsymbol{z}) = \log |\det(\nabla_{\boldsymbol{z}} f_\psi)| + \log p(\boldsymbol{s}) = \sum_t \log |\boldsymbol{a}_t(\boldsymbol{z}^t)| + \log p(\boldsymbol{s}) \triangleq \mathcal{L}_{\text{FLOW}}(f_\psi) \tag{3.2}$$

Jointly with $\mathcal{L}_{\text{GCL}}(f_\psi, r_\nu)$, $\mathcal{L}_{\text{FLOW}}(f_\psi)$ can be used to optimize the parameters of the de-mixing function $f_\psi(\boldsymbol{z})$.

3.3 Energy-based Causal Representation Transfer

Algorithm 2: Energy-based Causal Representation Transfer.

1. Pre-train encoder and predictor:
$$e_\theta, h_{\phi'} \leftarrow \arg\min\{\mathbb{E}[\ell(h(e(\boldsymbol{x})), y)]\}$$
2. NICA estimation with fixed $\boldsymbol{z} = e_\theta(\boldsymbol{x})$:
$$f_\psi, r_\nu \leftarrow \texttt{GCL}(\{(\boldsymbol{z}, y)\}) \quad \text{\% Equation (3.1)}$$
3. Source space augmentation with fixed $\boldsymbol{s} = f_\psi(\boldsymbol{z})$:
$$(\tilde{\boldsymbol{s}}, y) \leftarrow \texttt{AUG}(\{\boldsymbol{s} | y = M\}) \quad \text{\% Equation (3.3)}$$
4. Minority predictor modeling with augmented source:
$$h_\phi \leftarrow \arg\min\{\mathbb{E}[\ell(h(\boldsymbol{s}, y = M)] + \lambda \mathbb{E}[\ell(h(\tilde{\boldsymbol{s}}, y = M)]\}$$

In this section, we describe the construction of *energy-based causal representation transfer* (ECRT): a causally informed data transformation and augmentation procedure to improve learning with imbalanced datasets. Our model assumes a shared causal data-generation procedure, which can be accurately identified by learning

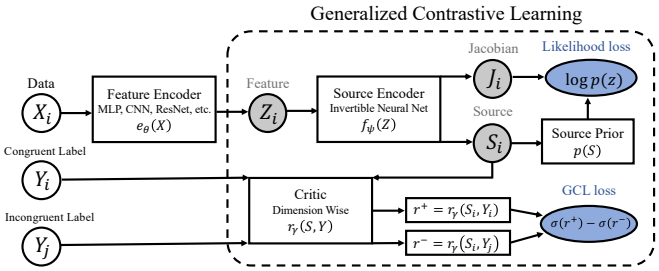

FIGURE 3.3: Source space estimation module of ECRT. We use GCL to identify the demixing function $\boldsymbol{s} = f_\psi(\boldsymbol{z})$ via telling apart congruent & incongruent pairs.

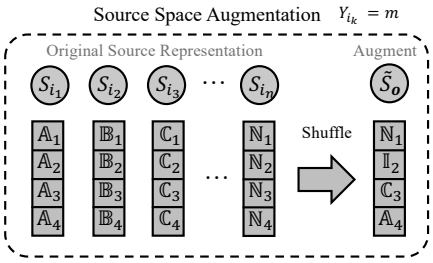

FIGURE 3.4: Non-parametric source space augmentation based on shuffling.

with the majority classes under assumed class-conditional representation independence. We obtain decorrelated representations that facilitates data augmentation and efficient learning with minority classes.

The proposed model consists of the following components: (i) a feature encoder module $\boldsymbol{z} = e_\theta(\boldsymbol{x})$; (ii) two classification modules, $h_{\phi'}(\boldsymbol{z})$ and $h_\phi(\boldsymbol{s})$, for predicting label y from features $\boldsymbol{z}$ and sources $\boldsymbol{s}$, respectively; (iii) a nonlinear ICA module for the de-mixing function $\boldsymbol{s} = f_\psi(\boldsymbol{z})$; (iv) a critic function $r_\nu(y, \boldsymbol{z})$ for GCL; and (v) a data augmentation module. Further, $(\theta, \phi', \phi, \psi, \nu)$ denote the parameters of all the neural-network-based modules. Algorithm 2 outlines a general workflow, and below, we elaborate on our assumptions and detail the implementation of ECRT.

3.3.1 Model assumptions

To enable knowledge transfer across classes, we make the following assumptions:

Assumption 3.3.1. Let $\boldsymbol{z}$ be a sufficient statistic (features) of $\boldsymbol{x}$ for predicting label y, all class conditional feature distributions $p(\boldsymbol{z}|y)$ share a common ICA de-mixing function $f_\psi(\boldsymbol{z})$.

This implies there exists a smooth invertible function $f_\psi : \mathcal{Z} \to \mathcal{S}$, and a set of IC distributions $\{q(\boldsymbol{s}|y = m)\}_{m=1}^M$, that are linked to the conditional feature distributions $p(\boldsymbol{z}|y = m)$ via $\boldsymbol{S}^m = f_\psi(\boldsymbol{Z}^m)$, where $\boldsymbol{S}^m \sim q(\boldsymbol{s}|y = m)$ and $\boldsymbol{Z}^m \sim p(\boldsymbol{z}|y = m)$.

The subscript m in $\boldsymbol{S}^m$ and $\boldsymbol{Z}^m$ indicates $y = m$, in a slight abuse of notation. Importantly, $f_\psi^{-1}(\boldsymbol{s}) : \mathcal{S} \to \mathcal{Z}$ is the *invariant causal mechanism* underlying the features of observed data. This specification, which is consistent with Assumption 4.6.1 enables the identification of the shared generating process, namely, the de-mixing function $f_\psi(\boldsymbol{z})$ that connects the likely dissimilar conditionals $\{p(\boldsymbol{z}|y = m)\}_{m=1}^M$ via the source conditionals $\{q(\boldsymbol{s}|y = m)\}_{m=1}^M$, as demonstrated in the context of NICA by (Hyvarinen et al., 2019)

A shortcoming of Assumption 4.6.1 is that the hypothesis supporting it is somewhat strong, untestable and it may not hold in practice. However, we argue that the structural constrains imposed to the model by the assumption via the source conditionals $q(\boldsymbol{s}|y = m)$ and the invertibility of $f_\psi(\boldsymbol{z})$, restrict the search space of the otherwise over-flexible model space powered by neural networks. Further, the causal mechanism implied by $f_\psi^{-1}(\boldsymbol{s})$ enables an effective knowledge transfer mechanism across classes via $q(\boldsymbol{s}|y = m)$.

3.3.2 Causal Representation Transfer

Encoder pre-training. To implement ECRT, we first find a good (reduced) feature representation of $\boldsymbol{x}$ highly predictive of label y. This can be achieved via supervised representation learning (see Figure 3.3), which optimizes an encoder and predictor pair $(e_\theta(\boldsymbol{x}), h_{\phi'}(\boldsymbol{z}))$ to minimize the label prediction risk $\mathcal{L}(\phi') \triangleq \mathbb{E}[\ell(h_{\phi'}(e_\theta(\boldsymbol{x})), y)]$, where $\ell(\cdot, \cdot)$ is a suitable loss function, *e.g.*, cross-entropy, hinge loss, *etc.* To avoid capturing *spurious* (non-generalizable) features that overfit the minority class, we advocate training only with majority samples at this stage; assuming that $M > 2$. Alternatively, one could also consider unsupervised feature extraction schemes, such as auto-encoders. Further, we also recommend using statistical adjustments such as importance weighting, to reduce the impact of data imbalance.

De-mixing representation with GCL. After obtaining a good feature representation $\boldsymbol{Z} = e_\theta(\boldsymbol{X})$, we proceed to learn the de-mixing function $f_\psi(\boldsymbol{z})$, such that

the coordinates of the source representation $\boldsymbol{S} = f_\psi(\boldsymbol{Z})$ are (approximately) independent given the label y. This can be done by optimizing the GCL objective in (3.1) wrt the feature representation and label pairings $(y, \boldsymbol{z})$, adopting the *masked auto-regressive flow* (MAF) (Papamakarios et al., 2017) to model the smooth, invertible transformation $f_\psi(\boldsymbol{z})$, which allows efficient parallelization of the autoregressive architecture via *causal masking*. The procedure is outlined in Figure 3.3.

Augmenting the minority. Inspired by (Teshima et al., 2020), we artificially augment the minority feature representations $\boldsymbol{z}$ via random permutations in the source space $\mathcal{S}$, as shown in Figure 3.4. Provided the assumed conditional independence of the sources, the features $\boldsymbol{Z}^m$ corresponding to label $y = m$ are generated by $\boldsymbol{Z}^m = f_\psi^{-1}(\boldsymbol{S}^m)$, where each dimension in the source representation $[\boldsymbol{S}^m]_j \sim q([\boldsymbol{s}]_j | y = m)$ are independently and *implicitly* sampled as described below. Specifically, using the (estimated) de-mixing function $f_\psi(\boldsymbol{z})$, we can obtain an approximate empirical source distribution for each $y = m$, *i.e.*, $S^m \triangleq \{\boldsymbol{s}_i = f_\psi(e_\theta(\boldsymbol{x}_i)) | y_i = m\} = \{\boldsymbol{s}_i\}_{i=1}^{n_m}$, where S^m is a collection of n_m samples of $q(\boldsymbol{s}|y = m)$. Then, we can draw new artificial samples $\tilde{\boldsymbol{s}}^m \sim q(\boldsymbol{s}|y = m)$ by randomly permuting the coordinates within elements S^m independently via

$$\tilde{\boldsymbol{s}}^m_{\boldsymbol{o}} = ([\boldsymbol{s}^m_{o_1}]_1, [\boldsymbol{s}^m_{o_2}]_2, \cdots, [\boldsymbol{s}^m_{o_d}]_d), \tag{3.3}$$

where $\boldsymbol{o} = (o_1, \cdots, o_d)$ is a random permutation of $(1, \ldots, n_m)$. Note we have used $\tilde{\boldsymbol{s}}^m_{\boldsymbol{o}}$ to emphasize that the source point is artificially created via permutation $\boldsymbol{o}$. We call this procedure *nonparametric augmentation* because it does not make distributional assumptions for $q(\boldsymbol{s}|y = m)$. However, below we will discuss its limitations and consider an alternative where a parametric form is assumed. While it is tempting to refine the predictor $h_{\phi'}(\boldsymbol{z})$ with artificial features augmented via $\tilde{\boldsymbol{z}}^m = f_\psi^{-1}(\tilde{\boldsymbol{s}}^m)$, in Section 3.3.3 we will argue that it stands to benefit more from training a new predictor directly based on source representations, *i.e.*, $h_\phi(\boldsymbol{s})$, without the need for

inverting $f_\psi(\boldsymbol{z})$.

Model refinement. Now we can leverage the augmented data to refine the prediction model. For minority class $y = M$, we optimize the following objective

$$\mathcal{L}_{\text{AUG}}(\phi') = \mathcal{L}(\phi') + \lambda(\mathbb{E}_{\tilde{\boldsymbol{Z}}^M}[\ell(h_{\phi'}(\tilde{\boldsymbol{z}}^M), M)] - \mathbb{E}_{\boldsymbol{Z}^M}[\ell(h_{\phi'}(\boldsymbol{z}), M)]), \tag{3.4}$$

where $\mathcal{L}(\phi')$ is the loss used for pre-training. Conceptually, (3.4) replaces a portion of the minority samples with augmentations. The trade-off parameter $\lambda \in [0, 1]$ encodes the relative confidence for trusting the artificially generated representations $\tilde{\boldsymbol{Z}}^M$ obtained from S^M for the minority label $y = M$. Further, at this stage we found it's beneficial to fix the encoder module to prevent the de-mixing function to accommodate the changes in the encoder which in practice may cause instability during training.

Challenges with naïve implementation. We identify three major issues with naïvely implemented ECRT, to be addressed in the section below: (*i*) *Representation conflict*: since the GCL solution is not unique, we do observe naïve GCL training drifts among viable source representations whose performance differ considerably, causing stability concerns; (*ii*) *Costly augmentation*: MAF inversions dominate the computation load during training, which becomes prohibitive in high dimensions; and (*iii*) *Gridding artifact*: a small minority sample size leads to pronounced augmentation bias when sampling nonparametrically via (3.3), manifested as a rectangular-shaped grid (see Figure 3.5).

3.3.3 Improving causal representation transfer

Energy-based GCL. Our key insight to improve GCL comes from the fact that Equation (3.1) is essentially learning the density ratio between the joint and product of marginals, *i.e.*, $\frac{p(x,y)}{p(x)p(y)}$. This immediately reminds us the recent literature on contrastive *mutual information* (MI) estimators, such as InfoNCE (Poole et al., 2019). In such works, a variational lower bound of MI is derived, and the algorithm optimizes a

critic function using the positive samples from the joint distribution, and the negative samples from product of the marginals. At their optimal value, these critics recover the density ratio or a transformation of it. Our development is based on the recent work of Guo et al. (2021), where using an energy-perspective to improve contrastive learning. Specifically, we will be using a variant of the celebrated *Donsker-Varadhan* (DV) estimator Donsker and Varadhan (1983), and applied Fenchel duality trick to compute a solution Fenchel (1949). Specifically, the *Fenchel-Donsker-Varadhan* (FDV) estiamtor takes the following form:

$$
\begin{aligned}
I_{\mathrm{FDV}} &\triangleq \hat{I}^K_{\mathrm{DV}}(\{\boldsymbol{x}_i, \boldsymbol{y}_i\}) + \frac{\sum_j \exp[(g_\theta(\boldsymbol{x}_i, \boldsymbol{y}_j) - g_\theta(\boldsymbol{x}_i, \boldsymbol{y}_i))/\tau]}{\sum_j \exp[(\hat{g}_\theta(\boldsymbol{x}_i, \boldsymbol{y}_j) - \hat{g}_\theta(\boldsymbol{x}_i, \boldsymbol{y}_i))/\tau]} + 1 \\
\hat{I}^K_{\mathrm{DV}} &= g(\boldsymbol{x}_1, \boldsymbol{y}_1) - \log(\sum_{k'=1}^{K} \exp(g(\boldsymbol{x}_1, \boldsymbol{y}'_k))/K)
\end{aligned}
\tag{3.5}
$$

where $g_\theta(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is our critic of interest and $\hat{I}^K_{\mathrm{DV}}$ is the *Donsker-Varadhan* (DV) estimator Donsker and Varadhan (1983) for the MI. Using the same parameterization used in GCL recovers the same causal identification property (see Appendix for details). Compared to the original GCL formulation, we are now using multiple negative samples instead of one, which greatly boosts learning efficiency Gutmann and Hirayama (2012). And this can be efficiently implemented with the *bilinear* critic trick (Chen et al., 2020; Guo et al., 2021) so that all in-batch data can be used as negatives. In our context, it greatly boosts training efficiency when dealing with a large number of different classes. See Algorithm S1 in Appendix.

Regularizing the data likelihood. Recall GCL solutions can only be identified up to an invertible transformation of each dimension (Hyvarinen et al., 2019), and the predictive performance of different valid GCL solutions can vary significantly. Empirically, we observe that a naïve implementation of GCL often leads to source representations that are densely packed (see Figure B.2 in the Appendix). This is undesirable, when decoding back to the feature space and making useful predictions,

the neural network predictor will need to be expansive, *i.e.*, requiring a large Lipschitz constant, thereby sacrificing optimization stability and model generalization according to existing learning theory (Cortes and Vapnik, 1995; Vapnik, 2013).

To encourage source representations that are less condensed, we consider a simple, intuitive strategy consisting on regularizing the source representation with the log-likelihood in feature space. This likelihood can be easily obtained with MAF using $\ell_{\text{FLOW}}(f_\psi)$ defined in the (3.2). Following common practice, we set source prior $p(\boldsymbol{s})$ to the standard Gaussian, and optimize the following likelihood-regularized GCL objective:

$$\tilde{\mathcal{L}}_{\text{GCL}}(f_\psi, r_\nu) = \mathcal{L}_{\text{GCL}}(f_\psi, r_\nu) + \rho \mathcal{L}_{\text{FLOW}}(f_\psi), \tag{3.6}$$

where $\rho > 0$ is the regularization strength. Naturally, this regularization will encourage the global source representation identified by GCL to be more consistent with a Gaussian-shaped distribution.

An alternative interpretation for the likelihood-regularized objective in (3.6) is that it can be understood as a relaxation to the conditional independence (Assumption 4.6.1). To see this, recall the likelihood objective obtained from the *invertible neural networks* (INN) alone attempts to map the source representations to be *unconditionally* independent, as opposed to the *conditional* independence assumed by NICA. The regularized formulation (3.6) provides a safe *"fall-back"* mode in case Assumption 4.6.1 is violated. This also motivates us to consider an important variant: making multiple class-dependent source priors, *i.e.*, $p^m(\boldsymbol{s})$ for each class label $m \in \{1, \cdots, M\}$ in (3.6), whose parameters (*i.e.*, mean and variance) are jointly learned with other model parameters. We refer to this variant as ECRT-MULTI. Compared to the fully non-parametric objective (3.1), this strategy further encourages the source representations to be independent given the class labels, and it enables *parametric* data augmentation, *i.e.*, sampling from the parametric label priors

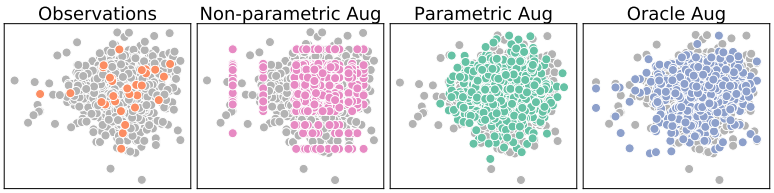

FIGURE 3.5: Comparison of different source augmentations overlaid on the ground-truth distribution (gray dots). A severe *gridding artifact* is observed in the low-sample regime for the nonparametric scheme, whereas the parametric augmentation closely matches the oracle in distribution.

instead of permuting the indices. **Modeling in the source space.** Rather than modeling the predictor $h_{\phi'}(\boldsymbol{z})$ in the feature space $\mathcal{Z}$, we advocate instead for building the predictor directly in the source space $\mathcal{S}$, *i.e.*, modeling with $h_{\phi}(\boldsymbol{s})$. This practice enjoys several benefits: (i) *Easy & robust augmentation*: many designs of high-dimensional flows are asymmetric computationally, and inverting a MAF is not only d times more costly than a forward pass, it is also numerically unstable at the boundary. Direct modeling in the source space circumvents the difficulties associated with MAF inversions during data augmentation; (ii) *Feature whitening*: the source representation identified by GCL is component-wise independent, and literature documents abundant empirical evidence that similar de-correlation based pre-processing, commonly known as *whitening*, benefits learning (Huang et al., 2018; Bansal et al., 2018; Jia et al., 2019).

Parametric augmentation. When the number of minority observations are scarce, the above non-parametric indices-shuffling augmentation suffers from the *gridding artifact* shown in Figure 3.5. This artifact amplifies the augmentation bias in the low-sample regime. To overcome this limitation, we empirically observe that the estimated class-conditional source distributions are usually Gaussian-like after the likelihood regularization (especially so when label conditional priors are used). In these situations, parametric augmentation that draws synthetic source samples from a Gaussian distribution matched to the empirical mean and variance of minority source representations is more efficient.

3.3.4 Insights and remarks

To better appreciate the gains and limitations expected from ECRT, we compile a few complementary arguments below, through the lens of very different perspectives.

Why causal augmentation works. It is helpful to understand the gains from ECRT's causal augmentation beyond the heuristic that permuting the ICs provides more training samples for the minority class. (Teshima et al., 2020) considered a similar causal augmentation procedure for few-shot learning, and provided two major theoretical arguments: (i) the risk estimator based on the augmented examples is the uniformly minimum variance unbiased estimator given the accurate estimation of f_ψ (see Theorem 1, (Teshima et al., 2020)); and (ii) with high probability, the generalization gap can be bounded by the approximation error of f_ψ (see Theorem 2, (Teshima et al., 2020)). In Appendix B.1, we give arguments that our causal augments give the 'best' label-conditional distribution estimate.

Speedup from shared embedding. While for typical supervised learning tasks the generalization bound scale as $\mathcal{O}(n^{-\frac{1}{2}})$, a superior rate of $\mathcal{O}(n^{-\eta})$ where $\eta \in [\frac{1}{2}, 1]$ is possible, if there exists abundant data for an alternative, yet related task that shares the same feature embedding (see Theorem 3, (Robinson et al., 2020)). Note n refers to the size of labeled data directly related to the *strong* task of interest, in our case, prediction of minority labels. Our ECRT employs GCL to identify one such common embedding, *i.e.*, the source space, using the majority examples, and consequently, improves predictions on the minority class.

Representation whitening. Our ECRT causally disentangles representation (Siddharth et al., 2017; Tokmakov et al., 2019) via de-correlating the representations conditionally. Extensive empirical evidence has pointed to the fact that such representation de-correlation, more commonly known as data whitening (Kessy et al., 2018), is expected to considerably improve learning efficiency (Cogswell et al., 2015). This benefit has been attributed to the better conditioning of the Fisher informa-

tion matrix for gradient-based optimization (Desjardins et al., 2015), thus rectifying second-order curvatures to accelerate convergence. Our source space modeling explicitly separates the task of representation disentanglement, and in turn, helps the prediction network to focus on its primary goal.

Potential limitations. The setting considered by ECRT is restrictive in that it precludes the learning of useful, yet non-transferable features predictive of the minority labels. For instance, there might be a feature unique to the minority class. However, since the de-mixer $f_\psi(\boldsymbol{z})$ is only trained on the majority domains absent of this feature, it can not be accounted for by the ECRT model. This is a key limitation of causally inspired models, in that they are often too conservative for only retaining the invariant features, promoting cross-domain generalization at the cost of within-domain performance degradation (Rothenhäusler et al., 2018).

3.4 Related Work

Causal invariance and representation learning. A major school of considerations for building robust machine learning models is to stipulate invariant causal relations across environments (Schölkopf et al., 2012), such that one hopes to safely extrapolate beyond the training scenario (Bühlmann, 2018). We broadly categorize such efforts into two streams, namely predictive causal models and generative causal models. Prominent examples from the first category include ICP (Peters et al., 2016; Heinze-Deml et al., 2018) and IRM (Arjovsky et al., 2019), which highlight the identification of invariant representation and causal relations via penalizing environmental heterogeneity. Our solution pertains to the second category, where the data distribution shifts across environments can be tethered by an invariant generation procedure (Hyvarinen et al., 2019; Khemakhem et al., 2020). This work exploits the recovered causally invariant source representation to improve learning efficiency and mitigate the sample scarcity of minority labels in imbalanced sets.

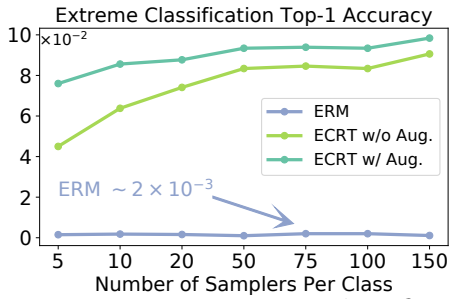

FIGURE 3.6: Extreme classification of $1k$ labels. ERM performs slightly better than random guess, while ECRT works significantly better.

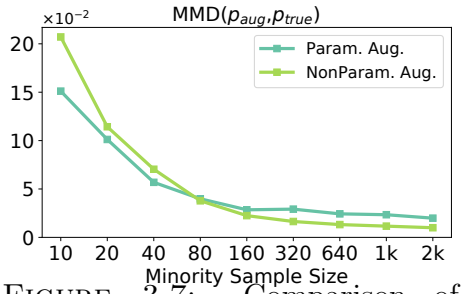

FIGURE 3.7: Comparison of different causal augmentations, lower is better. Parametric augmentation is more efficient with small samples.

Learning with imbalanced data. Resolving data imbalance is a heavily investigated topic (Liu et al., 2019). Standard sampling and weight adjustments suffer from caveats such as introducing bias and information loss. Due to these concerns, recent literature has actively explored adaptive strategies, such as redundancy-adjusted balancing weights (Cui et al., 2019), and the theoretically grounded class-size adapted margins(Cao et al., 2019). Similar to *boosting* (Freund and Schapire, 1997), adaptive weights are designed to prioritize the learning of less well-classified examples (Lin et al., 2017) while excluding apparent outliers (Li et al., 2019). Much related to our setting are the *meta-learning* scenarios (Vinyals et al., 2016; Finn et al., 2017; Wang et al., 2017), where the model tries to generalize & repurpose the knowledge learned from majority classes to minority predictions; and also the augmentation-based schemes to restore class balance (Mariani et al., 2018; Mullick et al., 2019).

Data augmentation. Due to its exceptional effectiveness, augmentation schemes are widely adopted in practical applications (Shorten and Khoshgoftaar, 2019). These augmentation strategies are built on known invariant transformations, *e.g.*, rotation, scaling, noise corruption, *etc.* (Szegedy et al., 2016), simple interpolation heuristics (Chawla et al., 2002; He et al., 2008), and more recently towards fully auto-mated procedures (Cubuk et al., 2019). Notably, recent trend in data augmentation high-

lights robust learning against adversarially crafted inputs (Goodfellow et al., 2015) and generative augmentation procedures that compose realistic artificial samples (Antoniou et al., 2017).

Domain adaption and causal mechanism transfer. Closest to our contribution is *causal mechanism transfer* (CMT) (Teshima et al., 2020), which focused on addressing few-shot learning for continuous regression. We note a few key differences to our work: (i) CMT focused on domain adaptation and does not address classification; and (ii) it bundles $(\boldsymbol{x}, y)$ for NICA which necessitates the flow inversion for sample augmentation. Grounded in the setting of imbalanced data learning, our ECRT extends applications and advocates source space modeling to simplify and improve the causal augmentation. It also features likelihood regularization to enhance representation regularity. We also offer new insights to justify the use of NICA-based augmentation, complementing the analysis from CMT by (Teshima et al., 2020).

3.5 Experiments

To validate the utility of our model, we consider a wide range of (semi)-synthetic and real-world tasks experimentally. All experiments are implemented with PyTorch, and our code is available from `https://github.com/zidixiu/ECRT`. More details of our setup & additional analyses are deferred to the Appendix B.4 and B.5.

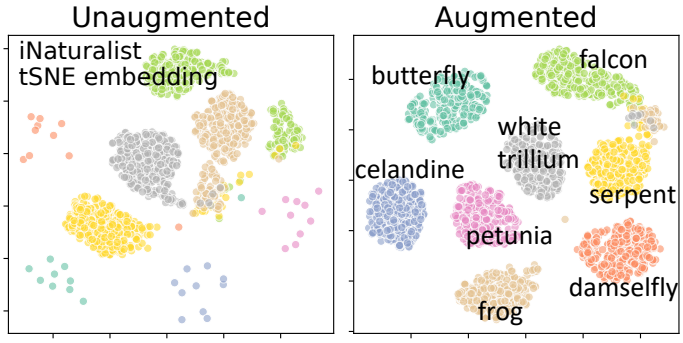

FIGURE 3.8: ECRT source representation trained on `iNaturalist`, visualized using tSNE embedding in two dimensions for eight random categories.

3.5.1 Experimental setup

Baselines. The following competing baselines are considered to benchmark the proposed solution: (*i*) Empirical risk minimization (ERM), a naïve baseline with no adjustment; (*ii*) Importance-weighting (IW) (Byrd and Lipton, 2019), a class-weight balanced training loss; (*iii*) Generative adversarial augmentation (GAN) (Antoniou et al., 2017), synthetic augmentations from adversarially-trained sampler; (*iv*) Virtual adversarial training (VAT) (Miyato et al., 2018), robustness regularization with virtual perturbations; (*v*) FOCAL loss (Lin et al., 2017), cost-sensitive adaptive weighting; (*vi*) Label-distribution-aware margin loss (LDAM) (Cao et al., 2019), margin-optimal class weights.

Evaluation metrics & setup. We consider the following metrics to quantitatively assess performance: (*i*) *negative* log-*likelihood* (NLL); (*ii*) F1 score; (*iii*) *Top-k* accuracy ($k = 1, 5$). Following the classical evaluation setup for imbalanced data learning, we learn on an imbalanced training set and report performance on a balanced validation set.

Toy model. We sample 2D standard Gaussians with different means and variances as source representation $\boldsymbol{s}$ for each label class, which is then distorted by random affine and Hénon transformations $z_1 = 1 - 1.4\,\tilde{s}_1^2 + \tilde{s}_2$, $z_2 = 0.3\,\tilde{s}_1$, to induce real-world-like complex association structures. See the Appendix B.5 for a detailed description and visualizations.

Real-world datasets. We consider the following semi-synthetic and real datasets: (*i*) *Imbalanced* `MNIST` *and* `CIFAR100`*:* standard image classification tasks with artificially created step-imbalanced following (Cao et al., 2019); (*ii*) *Imbalanced* `TinyImageNet` (adversarial training: a regularization method for supervised and semi-supervised learning, tin): a scaled-down version of the classic natural image dataset `ImageNet`, comprised of 200 classes, 500 samples per class and $10k$ validation images, with different simulated imbalances applied; (*iii*) *iNaturalist 2019* (Van Horn et al., 2018):

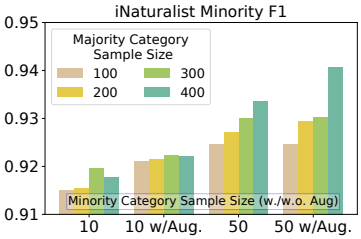

FIGURE 3.9: Performance comparison with and without augmentation. Augmented solutions improve the effective sample size.

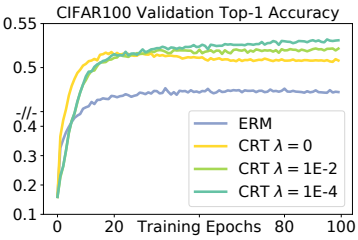

FIGURE 3.10: Comparison of learning dynamics. ECRT enables both faster learning and better model predictions.

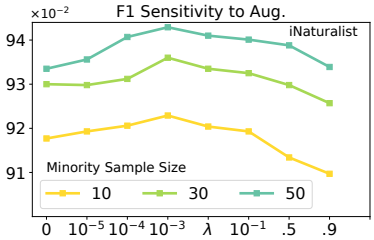

FIGURE 3.11: Sensitivity analysis of augmentation strength λ. Smaller minority sample sizes are more sensitive to the augmentation.

Table 3.1: Comparison of performance on real-world datasets (↑ higher is better, ↓ lower is better).

	CIFAR100			iNaturalist			TinyImagenet			ArXiv	
	Top-1↑	Top-5↑	NLL↓	Top-1↑	Top-5↑	NLL↓	Top-1↑	Top-5↑	NLL↓	Acc↑	NLL↓
ERM	49.29	78.22	2.95	66.73	87.86	1.70	58.52	79.01	3.22	44.64	**0.0407**
IW	43.97	68.96	3.89	67.63	88.94	1.66	60.50	80.23	2.92	46.02	0.0477
GAN	47.64	78.53	2.69	67.40	87.00	1.82	60.69	80.91	2.33	45.42	**0.0407**
VAT	46.47	74.23	3.19	67.06	87.39	1.90	59.69	82.05	2.42	45.82	0.0474
FOCAL	43.32	74.39	2.89	66.63	88.45	1.59	58.27	79.39	2.59	46.01	0.0416
LDAM	50.46	74.39	2.18	67.39	87.13	4.00	58.18	82.51	2.15	45.04	0.0450
ECRT	52.31	81.14	**1.98**	68.38	88.13	1.48	62.46	83.50	**1.79**	**48.33**	0.0434
ECRT MULTI	**53.00**	**81.99**	2.31	**69.01**	**90.01**	**1.23**	**64.40**	**84.54**	1.94		

a challenging task for image classification in the wild comprised of $26k$ training and $3k$ validation images labeled with $1k$ classes, with a skewed distribution in label frequency; (*iv*) ***arXiv*** *abstracts*, imbalanced multi-label prediction of paper categories with $160k$ samples and 152 classes.

Preprocessing, model architecture, and tuning. We use pre-trained models to extract vectorized representations for big complex datasets: ResNet (He et al., 2016) for image models of `iNaturalist`, `TinyImageNet`) and BERT (Devlin et al., 2019) for the `arXiv` (training of Deep Bidirectional Transformers for Language Understanding, arx) language model. These pre-trained representations are used as raw feature inputs, subsequently fed into a fully-connected multi-layer perceptrons (MLP) for source space encoding. For small image models (*i.e.*, MNIST, CIFAR100) we directly train a CNN from scratch for feature extraction. We used a random 8/2

split for training and validation, and applied Adam optimizers for training. We rely on the best out-of-sample cross-entropy and GCL loss for hyperparameter tuning.

ECRT efficiency. We first examine the dynamics of efficiency gains from ECRT using a toy model, with the results summarized in Figure 3.9. Consistent with the shared embedding perspective, increasing majority sample size improves the minority accuracy. Causal augmentation consistently improves performance, and it is most effective in the low-sample regime, offering more than $4\times$ boost in effective minority sample size. In Figure 3.10, we compare modeling with feature and source representations, respectively. Source space modeling speeds up learning, in addition to the modest accuracy gains. Augmented training is slightly slower, but eventually converges to a better solution. Interestingly, the major performance gain originates from the GCL-based source space modeling we proposed, and to a lesser extent, from the augmentation perspective presented by (Teshima et al., 2020).

Ablations on augmentation. We examined the contributions from augmentation as we vary the augmentation strength λ (Figure 3.11). A relatively small λ already yields improvement, but when the minority size is small, stronger augmentation degrades performance. This implies the augmentation gains are more likely to be originated from the exposure to the diversity of synthetic augmenting, and the accuracy of augmented distribution is limited by the imperfectness of empirical estimation (*e.g.*, limited sample size). In Figure 3.7 we compare the MMD distance (Gretton et al., 2012) between the parametric and nonparametric augmentation outputs and the ground-truth, confirming the improved efficiency from the ECRT parametric augmentation in the low-sample regime.

Extreme classification. A related challenge of interest is extreme classification (Choromanska et al., 2013), where the label size is huge but there is only a handful of samples pertaining to each label class, This means there is no clear majority class. In Figure 3.6, we show the proposed ECRT also fares very well in this scenario, while

standard ERM struggles (slightly better than random guess). This result evidences that ECRT also efficiently extracts generalizable information from an abundance of different class labels.

Imbalanced data learning. In Table 3.1 we compare the performance of ECRT to other competing solutions, with each baseline carefully tuned to ensure fairness. Minor discrepancies compared to results reported by prior literature are explained in the Appendix B.5, and in general our implementation performs slightly better. Overall, ECRT-based solutions consistently lead the performance chart, with the label-prior variant solidly outperforming vanilla ECRT and other competitors in most categories. While showing the varying degrees of success, FOCAL and LDAM failed to establish dominance comparing to ERM. GAN and VAT also verified the effectiveness of augmentation and adversarial perturbations in the imbalanced setting. Figure 3.1 compares performance of different learning schemes condition on different label sizes using the F1 score on the `iNat` data. In Figure 3.8 we show the ECRT learned representations, with and without augmentation, using tSNE embeddings.

3.6 Conclusion

This paper developed a novel learning scheme for imbalanced data learning. Leveraging a causal perspective, our solution cuts through the apparent data heterogeneity and identifies a shared, invariant, disentangled source representation, using only majority samples. We demonstrate this source representation modeling enables more efficient learning and allows principled augmentation. Importantly, we bridge the research between contrastive representation learning and energy modeling to causal machine learning, which construct promising directions for future research.

Acknowledgements

This research was supported in part by NIH/NIDDK R01-DK123062, NIH/NIBIB R01-EB025020, NIH/NINDS 1R61NS120246-01, DARPA, DOE, ONR and NSF.

4

Personalized Time-to-event Estimation with Representation Learning

4.1 Introduction

Prediction of event times, also known as *survival analysis* in the clinical context, is one of the most extensively studied topics in the statistical literature, largely due to its significance in a wide range of clinical and population health applications. It provides a fundamental set of tools to statistically analyze the future behavior of a system, or an individual. In the classical setup, the primary goal of time-to-event modeling is to either characterize the distribution of the occurrence of an event of interest on a population level (Kaplan and Meier, 1958; Kalbfleisch and Prentice, 2011), or more specifically, to estimate a risk score on a subject level (Cox, 1972). In recent years, there has been a surge of interest in the prediction of individualized event time distributions (Yu et al., 2011).

A characteristic feature in the study of time-to-event distributions is the pres-

Part of this chapter is a published paper: Xiu, Zidi, Chenyang Tao, and Ricardo Henao. "Variational learning of individual survival distributions." In Proceedings of the ACM Conference on Health, Inference, and Learning, pp. 10-18. 2020. (Xiu et al., 2020)

ence of censored instances, which refer to an event that is not reported during the follow-up period of a subject. This can happen, for instance, when a subject drops out during the study (right censoring), including when the study terminates before the event happens (administrative censoring). Unlike many conventional predictive models, where incomplete observations are usually safely ignored, censored observations contain crucial information that should be adequately considered. To efficiently leverage the censored observations, together with the complete observations, a classical treatment is to work with the notion of a *hazard function*, formally defined as the instantaneous event risk at time t, which can be computed by contrasting the event population to the population at risk at a specific time. Estimates can be derived, for instance by optimizing the *partial likelihood* defined by the relative hazards in the case of the Cox Proportional Hazard model (CoxPH) (Cox, 1972). Alternatively, other work follows the standard *Maximal Likelihood Estimation* (MLE) framework, where the individual event distribution is a deformed version of some baseline distribution. For example, in the *Accelerated Failure Time* model (AFT) (Kalbfleisch and Prentice, 2011), covariate effects are assumed to rescale the *temporal index* of event-time distributions, *i.e.*, they either accelerate or delay event progression. For censored events, their likelihoods are given as the cumulative density after the censoring time (Aalen et al., 2008).

While vastly popular among practitioners, these models have been criticized for a number of reasons, in particular for the assumptions they make, that consequently render them unfit for many modern applications (Wang et al., 2019). For instance, most survival models, including CoxPH and the proportional odds model (Murphy et al., 1997), work under the premise of fixed covariate effects, overlooking individual uncertainty. However, it has been widely recognized that, individual heterogeneity and other sources of variation are common and often time-dependent (Aalen, 1994). In real-world scenarios, these random factors are typically costly to measure, if not

impossible to observe. Unfortunately, many models are known to be sensitive to the violation of this fixed effect assumption, raising seriously concerns when deployed in actual practice (Hougaard, 1995).

Alternatively, machine learning techniques have been leveraged to overcome the limitations of standard statistical survival modeling schemes, especially in terms of model flexibility to address the complexity of data. For example, survival trees employed special node-splitting strategies to stratify the population and derive covariate-based survival curves (Bou-Hamad et al., 2011), support vector machines (Khan and Zubek, 2008) and neural networks (Faraggi and Simon, 1995) have been used for more expressive predictors and LASSO-type variants (Zhang and Lu, 2007) simultaneously execute variable selection to boost statistical efficiency. Bayesian statistics has also been explored in the context of model selection (Lisboa et al., 2003), averaging (Raftery, 1995) and imposing prior beliefs (Fard et al., 2016). Recent advances in modern machine learning bring extra traction to the concept of data-driven survival models, an important step toward precision medicine. Prominent examples include direct deep learning extensions of CoxPH (Katzman et al., 2016; Li et al., 2019), accelerated failure time (Chapfuwa et al., 2018) and Bayesian exponential family models (Ranganath et al., 2016). Other efforts include the use of Gaussian Process to capture complex interactions between covariates in relation to event times (Fernández et al., 2016) and competing risks (Alaa and van der Schaar, 2017). It has been argued that direct modeling of the event distribution might be beneficial (Yu et al., 2011), and more recently, adversarial distribution matching has also been considered for survival applications (Chapfuwa et al., 2018) with promising results reported.

In this chapter we present two approaches to address the challenges of non-parametric modeling of time-to-event distributions in the presence of censored instances. The first approach, named *Variational Survival Inference* (VSI) (Xiu et al.,

2020), builds upon recent developments in black-box variational inference (Ranganath et al., 2014). By explicitly accounting for latent variables in its formulation as multivariate Gaussian, VSI accommodates for individual uncertainty. The proposed VSI is a highly scalable and flexible framework without strong assumptions, featuring easy implementation, stable learning, and importantly, it does not rely on *ad-hoc* regularizers.

Though VSI fills in the gap for non-parametric time-to-event modeling with variational inference framework, it fails to account for data incompleteness and the black-box prediction framework can not be interpreted. Moreover, VSI restrains the form of latent representation as a multivariate Gaussian, which limits its flexibility.

In recognition of the limitations discussed in Section 4.1 about current methods for time-to-event estimation and VSI, we present *energy-based self-attentive survival analysis* (ELSSA): a novel time-to-event prediction model that combines ideas from representation learning and energy modeling [1] , to address the identified weakness of existing solutions to time-to-event estimations. A key novelty of this approach is the integration of contrastive mutual information based loss that non-parametrically maximizes the informativeness of learned data representations. Our key contributions include: (*i*) the novel application of the state-of-the-art contrastive learning and self-attention techniques to address the challenges of data incompleteness; (*ii*) the derivation of an energy-based mutual information objective to embed temporal information for representations optimization; (*iii*) efficiently handled censored objectives with a likelihood based objective; (*iv*) insightful discussions on the generalization to longitudinal data.

Both methods directly targets the estimation of individualized event-time distributions, rather than a risk score that correlates with event ordering.

[1] distribution free modeling without closed-form expressions.

4.2 Background

A dataset for survival analysis is typically composed of a collection of triplets $D = \{Y_i = (t_i, \delta_i, X_i)\}_{i=1}^N$, where i indexes the subjects involved in the study. For each triplet, $X_i \in \mathbb{R}^p$ denotes the set of explanatory variables, t_i is the observation time and δ_i is the event indicator. To simplify our discussion, we only consider the standard survival setup. This means δ_i is binary with $\delta_i = 1$ indicating the event of interest happened at t_i, otherwise $\delta_i = 0$ corresponds to a censoring event, *i.e.*, no event occurs until t_i and the subject is unobserved thereafter. This distinction creates a natural partition of the dataset $D = D_c \bigcup D_e$, with $D_c = \{Y_i : \delta_i = 0\}$ and $D_e = \{Y_i : \delta_i = 1\}$ representing the censored and event groups, respectively.

4.2.1 Statistical survival analysis

In survival analysis, one is interested in characterizing the survival function $S(t)$, defined as the probability that any given subject survives until time t. The basic descriptors involved in the discussion of survival analysis are: the cumulative survival density $F(t) = 1 - S(t)$, the survival density $f(t) = \partial_t F(t)$, the hazard function $h(t) = \lim_{\Delta t \to 0} \frac{P(t \leqslant T < t + \Delta t | T \geqslant t)}{\Delta t}$ and the cumulative hazard function $H(t) = \int_0^t h(s)\,\mathrm{d}s$. The following expressions are fundamental to survival analysis (Aalen et al., 2008): $S(t) = \exp(-H(t))$ and $f(t) = h(t)S(t)$. Further, we use $S(t|x)$, $f(t|x)$, $F(t|x)$, $h(t|x)$, $H(t|x)$) to denote their individualized (subject-level) counterparts given explanatory variables x. All survival models leverage these definitions to derive population-level estimators or subject-level predictive functions, *e.g.*, of risk, $S(t|x)$, or event time, $f(t|x)$.

4.3 Preliminaries for VSI

4.3.1 Variational inference

For a latent variable model $p_\theta(x, z) = p_\theta(x|z)p(z)$, we consider $x \in \mathbb{R}^p$ as an observation, *i.e.*, data, and $z \in \mathbb{R}^m$ as latent variable. The marginal likelihood, given by $p_\theta(x) = \int p_\theta(x, z)\,\mathrm{d}z$, typically does not enjoy a closed form expression. To avoid direct numerical estimation of $p_\theta(x)$, Variational Inference (VI) optimizes a variational bound to the marginal log-likelihood. The most popular choice is known as the Evidence Lower Bound (ELBO) (Wainwright and Jordan, 2008), given by

$$\mathrm{ELBO}(x) \triangleq \mathbb{E}_{Z\sim q_\phi(z|x)}\left[\log \frac{p_\theta(x, Z)}{q_\phi(Z|x)}\right] \leqslant \log p_\theta(x), \tag{4.1}$$

where $q_\phi(z|x)$ is an approximation to the true (unknown) posterior $p_\theta(z|x)$, and the inequality is a direct result of Jensen's inequality. The variational gap between the ELBO and true log-likelihood is the KL-divergence between posteriors, *i.e.*, $\mathrm{KL}(q_\phi(z|x) \parallel p_\theta(z|x)) = \mathbb{E}_{q_\phi(z|x)}[\log q_\phi(z|x) - \log p_\theta(z|x)]$, which implies the ELBO tightens as $q_\phi(z|x)$ approaches the true posterior $p_\theta(z|x)$. For estimation, we seek parameters θ and ϕ that maximize the ELBO. At test time, $q_\phi(z|x)$ is used for subsequent inference tasks on new data. Given a set of observations $\{x_i\}_{i=1}^N$ sampled from data distribution $x \sim p_d$, maximizing the expected ELBO is also equivalent to minimizing the KL-divergence $\mathrm{KL}(p_d \parallel p_\theta)$ between the empirical and model distributions. When $p_\theta(x|z)$ and $q_\phi(z|x)$ are specified as neural networks, the resulting architecture is more commonly known as the Variational Auto-Encoder (VAE) (Kingma and Welling, 2014) in the context of computational vision and natural language processing.

4.4 Variational Survival Inference

In this work we present a principled approach to address the challenges of nonparametric modeling of time-to-event distributions in the presence of censored instances. Our approach, named *Variational Survival Inference* (VSI), builds upon recent developments in black-box variational inference (Ranganath et al., 2014). It directly targets the estimation of individualized event-time distributions, rather than a risk score that correlates with event ordering. By explicitly accounting for latent variables in its formulation, VSI better accommodates for individual uncertainty. The proposed VSI is a highly scalable and flexible framework without strong assumptions, featuring easy implementation, stable learning, and importantly, it does not rely on *ad-hoc* regularizers. Our key contributions include: (i) a variational formulation of nonparametric time-to-event distribution modeling conditioned on explanatory variables; (ii) a cost-effective treatment of censored observations; (iii) a thorough discussion on how our modeling choices impact VSI performance, and (iv) an empirical validation confirming that the proposed VSI compares favorably to its counterparts on an extensive set of tasks, covering representative synthetic and real-world datasets.

Below we detail the construction of the Variational Survival Inference (VSI) model, which results in predictions of the time-to-event distribution $p_\theta(t|x)$ given attribute x, with the individual uncertainty accounted in the form of a latent variable z whose distribution is estimated under the VI framework. Unlike classical survival models, we do not need to specify a parametric form for the baseline distribution, *e.g.*, the base hazard $h_0(t)$ in CoxPH (Cox, 1972) or the base density $p_0(t)$ in AFT (Kalbfleisch and Prentice, 2011). Instead, we leverage the power of deep neural networks to amortize the learning of the event time and survival distributions, allowing arbitrary (high-order) interactions between the predictors and survival time to be captured. This overcomes the limitations caused by the restrictive assumptions

made in the classical statistical survival analysis frameworks, thus allowing flexible inference of time-to-event distributions.

4.4.1 Variational bound of observed events

We start the discussion with the simplest scenario, that for which there are no censoring events. Our goal is to maximize the expected log-likelihood $1/N \sum_i \log p_\theta(t_i|X_i)$. To model the conditional likelihood, we consider a latent variable model of the form $p_\theta(t, z|x)$. The unconditional formulation of the ELBO in (4.1) can be readily generalized to case conditional on event times as

$$\text{ELBO}(t|x) = \mathbb{E}_{Z\sim q_\phi(z|x,t)}\left[\log \frac{p_\theta(t, Z|x)}{q_\phi(Z|x,t)}\right], \tag{4.2}$$

where $q_\phi(z|x,t)$ denotes the conditional posterior approximation to the true (unknown) $p_\theta(z|x,t)$.

In particular, we assume a model distribution with the following decomposition

$$p_\theta(t, z|x) = p_\theta(t|z, x)p_\theta(z|x) = p_\theta(t|z)p_\theta(z|x), \tag{4.3}$$

which posits that z is a sufficient statistics of x w.r.t. survival time t. Another key assumption we make is that, unlike in the standard variational inference model, we have used a learnable inhomogeneous prior $p_\theta(z|x)$ for the latent z to replace the standard fixed homogeneous prior $p(z)$. Such covariate-dependent prior formulation allows the model to account for individual variation, thus further helping to close the variational gap (Tomczak and Welling, 2018). Replacing (4.3) into the ELBO expression in (4.2) results in the usual likelihood and KL decomposition pair

$$\text{ELBO}(t|x) = \mathbb{E}_{Z\sim q_\phi(z|x,t)}\left[\log p_\theta(t|Z)\right] - \text{KL}(q_\phi(z|x,t) \parallel p_\theta(z|x)), \tag{4.4}$$

from which we can see that maximizing the ELBO is equivalent to estimate the parameters of a probabilistic time-to-event model $p_\theta(t|z)p_\theta(z|x)$ with maximum likelihood such that the inhomogeneous prior $p_\theta(z|x)$ matches as well as possible a conditional posterior that explicitly accounts for event times, $q_\phi(z|x,t)$. At test time,

only $p_\theta(z|x)$ will be used to make predictions provided that t is not available during inference.

More specifically, $p_\theta(t|z)$, $p_\theta(z|x)$ and $q_\phi(z|x,t)$ are defined as neural networks

$$\begin{aligned} p_\theta(t|z) &= \text{Softmax}(g(z;\theta)), \\ p_\theta(z|x) &= \mathcal{N}(\mu_p(x;\theta), \Sigma_p(x;\theta)), \\ q_\phi(z|x,t) &= \mathcal{N}(\mu_q(x,t;\phi), \Sigma_q(x,t;\phi)), \end{aligned} \tag{4.5}$$

where $p_\theta(t|z)$ is represented on a discretized time line (see below for details), $g(z;\theta)$, $\mu_p(x;\theta)$, $\Sigma_p(x;\theta)$ and $\mu_q(x,t;\phi)$, $\Sigma_q(x,t;\phi)$ are deep neural nets parameterized by model parameters θ and variational parameters ϕ, and $\mathcal{N}(\mu,\Sigma)$ denotes the multivariate Gaussian with mean μ and (diagonal) covariance Σ. For standard tabular data, we use Multi Layer Perceptrons (MLPs) to specify these functions.

4.4.2 Variational bound of censored events

Addressing censoring in the formulation is more challenging as this type of *partial observation* is not subsumed in the conventional VI framework. To address this difficulty, we recall that in likelihood-based survival analysis, the likelihood function for censored observations is given by $\log S_\theta(t|x)$, where $S_\theta(t|x)$ is the survival function and t is the censoring time. For censored observations Y with $\delta = 0$, we do not have the exact event time t. This means that we only have partial information of the events, in that the event should happen only after the censoring time t.

To derive a tractable objective for censored observations, we first expand $\mathcal{L}_c(x,t) = \log S_\theta(t|x)$ based on its definition and an application of Fubini's theorem (Resnick,

2003) and Jensen's inequality, *i.e.*,

$$\begin{aligned}
\mathcal{L}_c(x,t) &= \log\, S_\theta(t|x) = \log \int_t^\infty p_\theta(t|x)dt \\
&\geqslant \mathbb{E}_{q_\phi(z|t,x)}\left[\log \frac{p_\theta(z|x)}{q_\phi(z|t,x)} + \log\, S_\theta(t|z)\right] \\
&= \mathbb{E}_{q_\phi(z|t,x)}[\log\, S_\theta(t|z)] - \mathrm{KL}(q_\phi(z|t,x)||p_\theta(z|x)) \\
&\triangleq \mathrm{ELBO}_c(t|x)
\end{aligned}$$

where the censored log-likelihood bound $\mathrm{ELBO}_c(t|x)$ is only evaluated on D_c, *i.e.*, the subset of censored observations. See Supplementary Materials for the full derivation of $\mathrm{ELBO}_c(t|x)$.

4.4.3 Implementing VSI

In the current instantiation of the model, we discretize time into M bins spanning the time horizon of the (training) data. This means that (at inference) t is only known up to the time bin it falls into. We note this is not a restrictive assumption as many survival data is only known up to certain temporal accuracy. That said, generalization to continuous observations is fairly straightforward. For datasets that do have a natural discretization, we leave the choice to the user. In this study, we partition the temporal index based on the percentiles of observed event time, while also allowing for an artificial $(M+1)$-th bin to account for event times beyond the full observation window, *i.e.*, events happening after the end-of-study as observed in the training cohort.

Since both $p_\theta(z|x)$ and $q_\phi(z|x,t)$ are assumed to be Gaussian, the following closed-form expression can be used in the computation of the KL terms above

$$\mathrm{KL}(q_\phi(z|x,t) \parallel p_\theta(z|x)) = \frac{1}{2}\left\{\mathrm{tr}\left(\Sigma_p^{-1}\Sigma_q\right) + (\mu_p - \mu_q)^T \Sigma_p^{-1} (\mu_p - \mu_q) - m + \log \frac{\det(\Sigma_p)}{\det(\Sigma_q)}\right\}. \tag{4.6}$$

Following Ranganath et al. (2014), we use diagonal covariance matrices and apply the reparameterization trick to facilitate stable differatiable learning.

In order to compute the term $S_\theta(t|x)$, we use discretized time scheme as previously described, and sum up all predicted probabilities subsequent to bin t. Note that this can be readily generalized to continuous time models. So long as the cumulative distribution of $p_\theta(t|z)$ enjoys a closed form expression, a numerical integration scheme is not necessary to implement VSI.

4.4.4 Importance-Weighted estimator for likelihood evaluation

For evaluation purposes, we need to be able to compute the model's log-likelihood for an observation $Y = (x_i, t_i, \delta_i)$, *i.e.*,

$$\mathcal{L}_{\text{VSI}}(x_i, t_i; \theta) = \delta_i \log p_\theta(t_i|x_i) + (1 - \delta_i) \log S_\theta(t_i|x_i). \tag{4.7}$$

In this study, we use the importance-weighted (IW) estimator (Burda et al., 2015), which provides a tighter bound to the log-likelihood. While more sophisticated alternatives might provide sharper estimates (Neal, 2001), we deem IW estimator sufficient for the scope of this study. Additionally, while the tighter bound can be repurposed for training, it does not necessarily result in improved performance (Rainforth et al., 2018), which we find to be the case in this study.

To obtain a more accurate value of the likelihood, we use the approximate posterior as our proposal, and use the following finite sample estimate

$$\hat{p}_\theta(t_i|x_i) = \int \frac{p_\theta(t_i|z)p_\theta(z|x)}{q_\phi(z|t_i, x)} q_\phi(z|t_i, x)dz \approx \frac{1}{L}\sum_{l=1}^{L} \frac{p_\theta(t_i|z_l)p_\theta(z_l|x)}{q_\phi(z_l|t_i, x)},$$

where L is the number of samples. The log-likelihood for the corresponding conditional survival function is

$$\hat{S}_\theta(t_i|x_i) = \int_{t>t_i} \int \frac{p_\theta(t_i|z)p_\theta(z|x)}{q_\phi(z|t_i, x)} q_\phi(z|t_i, x)dzdt \approx \frac{1}{L}\sum_{l=1}^{L} \frac{\int_{t>t_i} p_\theta(t, z_l|x)dt}{q_\phi(z_l|t_i, x)}$$

Note that by nature of Jensen's inequality, the resultant estimand will be an underestimation of the true log-likelihood. As L goes to infinity, the approximated lower bound will converge to the true log-likelihood.

4.4.5 Making predictions

Predictive time-to-event distribution During inference, given a new data point with x_*, according to the generative model $p_\theta(t|x_*) = \int p_\theta(t, z|x_*)dz = \int p_\theta(t|z)p_\theta(z|x_*)dz$, where the integration is conducted numerically by Monte Carlo sampling.

Point estimation of time-to-event To better exploit the learned approximated posterior $q_\phi(z|x,t)$, we generalize the importance sampling idea and provide a weighted average as time-to-event summary, rather than for instance using a summary statistic such as median or mean. Specifically, consider multiple samples of $t_*^{(l)} \sim p_\theta(t|x_*)$, then calculate a weighted average as

$$\begin{aligned} t_* &= \frac{\sum_{l=1}^{L} w_*^{(l)} t_*^{(l)}}{\sum_{l=1}^{L} w_*^{(l)}},\ w_*^{(l)} = \frac{p_\theta(z_l|x_*)}{q_\phi(z_l|t_*^{(l)}, x_*)}, \\ t_*^{(l)} &\sim p_\theta(t|x_*),\ \ z_l \sim q_\phi(z|t_*^{(l)}, x_*). \end{aligned} \tag{4.8}$$

In the Appendix C.4 we show that (4.8) gives better model performance for point-estimate-based evaluation metrics, Concordance Index in particular, compared to other popular summary statistic such as the median of $t_* \sim p_\theta(t|x_*)$ with L empirical samples.

4.4.6 DISSECTING VSI

In the experiments, we show the effectiveness of the proposed VSI model in recovering underlying time-to-event distributions. To provide additional insight into the differentiating components of the VSI model, we consider two baseline models that partially adopt a VSI design, as detailed below.

VSI without a q_ϕ arm (VSI-NoQ) In VSI, we use the variational lower bound to maximize the likelihood in survival studies by implicitly forcing the unknown intractable model posterior $p_\theta(z|x)$ to be close to the tractable posterior approximation $q_\phi(z|x,t)$. Via the KL divergence minimization, such matching allows the model to better account for interactions between covariates x and event times t captured by $q_\phi(z|x,t)$ to better inform the construction of the latent representation z via isolating out the individual uncertainty encoded by $p_\theta(z|x)$. If we exclude the interaction term (x,t) in q_ϕ and only make the prediction with x, *i.e.*, with the approximate posterior given by $q_\phi(z|x)$, through the same stochastic latent representation z, then naturally the optimal solution is to equate $q_\phi(z|x)$ with the prior $p_\theta(z|x)$ [2]. This basically eliminates q_ϕ from our formulation, and therefore we call this variant VSI-NoQ.

More specifically, without a q_ϕ arm the model described in Section 4.3 essentially becomes a feed-forward model with a special stochastic hidden layer z. In this case, the model likelihood is given by $p_\theta(t|x) = \int p_\theta(t,z|x)dz = \int p_\theta(t|z)p_\theta(z|x)dz$, where $p_\theta(t|z)$ and $p_\theta(z|x)$ are defined as in (4.3). Note that the only difference with VSI is the lack of the KL divergence term to match $p_\theta(z|x)$ to $q_\phi(z|x,t)$. This baseline model (VSI-NoQ) is considered to dissect the impact of excluding complex interaction between covariates and event time when constructing the individualized priors.

Deterministic feed-forward model (MLP) To understand the importance of the stochastic latent representations z, we consider a straightforward baseline which directly predicts the event time distribution based on the input x, *i.e.*, $p_\theta(\cdot|x) = \mathrm{Softmax}(\mathrm{g}_\theta(\mathrm{x}))$, which is essentially a standard multinomial regression with censored observation. In our study, we use the MLP to implement $g_\theta(x)$. And as such, hereafter we will refer to this model as MLP. Additionally, we also considered standard randomization schemes, such as dropout (Srivastava et al., 2014), in the construction

[2] Based on a KL-vanishing argument.

of a stochastic neural net, which promises to improve performance. Such strategy also incorporates randomness, however differs principally from the modeled uncertainty exploited by our VSI scheme. In our experiment section, we report the best results from MLP with or without dropout.

These baseline approaches use feed-forward deep learning networks to learn $p_\theta(t|x)$ without incurring the notation of variational inference. In the experiments we will show that the variational inference is crucial to the accurate learning of time-to-event distributions, resulting in better performance relative to these baselines, especially when the proportion of censoring events is high.

4.5 Preliminaries for ELSSA

This work explores the use of self-attention and covariate embeddings to address the challenge of time-to-event prediction with censored and incomplete observations. Our solution is formulated such that it can easily accommodate mixed data inputs (continuous and discrete) which is common in health data analysis. A key novelty of this approach is the integration of contrastive mutual information based loss that non-parametrically maximizes the informativeness of learned data representations.

Notations and problem definition For each subject, denoted by i, we have $(\boldsymbol{x}_i, \boldsymbol{m}_i, t_i, \delta_i)$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ denotes the set of explanatory variables, $\boldsymbol{m}_i \in \{0,1\}^d$ the missing indicator mask, t_i the event time and δ_i is the event indicator. To simplify our discussion, we only consider the standard right-censoring survival setup. This means δ_i is binary with $\delta_i = 1$ indicating the event of interest happened at t_i, otherwise $\delta_i = 0$ corresponds to a censoring event, *i.e.*, no event occurs until t_i and the subject is unobserved thereafter.

4.5.1 Attention and Self-attention

Attention. The attention mechanism allows the learning agent (the model) to focus on a subset of the input (covariates) (Walther et al., 2004; Bahdanau et al., 2015), and it has been widely used in Natural Language Processing (NLP) tasks. Consider the case when sequence of inputs (*e.g.,* an English sentence) is to be translated, into a sequence of outputs (*e.g.,* a translated sentence) in another language. A widely used sequential model for this task is the Seq2Seq model (Sutskever et al., 2014). Together with attention mechanism, it views the encoded representations as a set of key-value pairs for each word, $\{(\boldsymbol{k}_1, \boldsymbol{v}_1), (\boldsymbol{k}_2, \boldsymbol{v}_2), \ldots, (\boldsymbol{k}_l, \boldsymbol{v}_l)\}$, denoted by $(\mathbf{K}, \mathbf{V})$. On the decoder side, query vectors $\mathbf{Q} = \{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n\}$ are used to find which part of the input sequence is more contributory to this position (Vaswani et al., 2017). Note that $\mathbf{K}$, $\mathbf{V}$ and $\mathbf{Q}$ are all vectors. The attention function describes mapping a query $\boldsymbol{q}$ and a set of key-value pairs $(\boldsymbol{k}, \boldsymbol{v})$ to a weighted sum based on the similarity function of the query with the corresponding key. We use the dot-product attention function,

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}, \tag{4.9}$$

where d is a scaling factor w.r.t. the dimension of $\mathbf{K}$.

Self-attention. Self-attention, sometimes called intra-attention, is a special form of the attention mechanism, where the $\mathbf{Q}$ is from the encoder side, *i.e,* $\mathbf{K}$, $\mathbf{V}$ and $\mathbf{Q}$ are vectors from the same representations. This yields a summarized representation composed by different positions of a single sequence (input covariates) (Shaw et al., 2018). Previous work has shown the success of self-attention mechanism in finding abstractive summarization (Paulus et al., 2017) and reading comprenhension tasks (Cheng et al., 2016; Parikh et al., 2016).

The attention mechanism is capable of learning structural properties of the under-

lying input covariates distribution (Bahdanau et al., 2015). With the query vector and a set of inputs, the attention mechanism credits the relevance of every input vector against the query vector, then apply a weighted sum over the input sequence based on the relevance to create a query-specific representation of the inputs (Wu et al., 2020).

4.5.2 Contrastive Learning

Contrastive learning is an important technique used in both self-supervised and supervised learning tasks. Contrastive methods aim to learn representations by enforcing similar samples (positive pairs) to be close in the latent space and dissimilar samples (negative pairs) to be distant (Chen et al., 2020). To achieve such goal, various forms of contrastive learning mechanisms have been introduced. Embedding distances (Chopra et al., 2005), triplet loss (with an anchor point) (Schroff et al., 2015), multiclass N-pair loss (Sohn, 2016), and Noise Contrastive Estimation (NCE) which is logistic loss based contrastive learning trying to differentiate positive samples and noises (Gutmann and Hyvärinen, 2010).

Oord et al. (2018) provided a critical insight that contrastive learning is closely related to *mutual information* (MI), *i.e.,* minimizing the contrastive loss is connected with maximization MI. They proposed a method called InfoNCE, and derived a variational lower-bound of MI and optimized based on a critic function using a positive sample and K negative samples (noises)(Poole et al., 2019).

MI quantifies the *amount of information* on random variable x obtained through the other random variable y.

$$I(x, y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}, \tag{4.10}$$

which is also the *Kullback-Leibler divergence* (KL) between the joint distribution $p(x, y)$ and product of the respective marginals $p(x)p(y)$. MI has an intuitive inter-

pretation of the uncertainty about X reduced from knowing the other variable Y, *i.e.,* how much does X inform Y. $I(x, y) = H(X) - H(X|Y)$, where $H(X) = -\mathbb{E} \log p(x)$ is the entropy (Shannon, 1948).

To maximize the MI in Eq. 4.10, it's equivalent to maximize the density ratio between the joint and product of marginals, *i.e.*, $\frac{p(x,y)}{p(x)p(y)}$, which is called *point-wise mutual information* (PMI) in literature. InfoNCE optimizes a critic function, with positive samples drawn from the joint distribution and negative samples from the product of marginal distributions. At the critic's optimum value, it recovers the PMI or a function of it.

In our study, we introduced a variation of InfoNCE to maximize the MI between covariate embeddings and the time-to-event to improve the learning efficiency.

4.6 Energy-based Latent Self-attentive Survival Analysis

In this section, we describe the construction of *energy-based latent self-attentive survival analysis* (ELSSA).

4.6.1 Model Assumptions

Assumption 4.6.1. Let $\boldsymbol{z}$ be a sufficient statistic (features) of $\boldsymbol{x}$ for predicting time-to-event t, then the model has the form

$$p_\theta(t, z|x) = p_\theta(t|z, x)p_\theta(z|x) = p_\theta(t|z)p_\theta(z|x), \tag{4.11}$$

4.6.2 Covariates Embedding and Self-Attention

To handle mixed data input (discrete and continuous variables), we create an embedding vector for each covariate. By self-attention mechanism, the structural representation can be learned from the embedding matrix for each patient based on relative importance of the covariates.

Algorithm 3: Energy-based Latent Self-attentive Survival Analysis.

1. Pre-train embedding and the energy-based encoder:

$$\mathbf{V} \leftarrow \text{emb}_\theta(\boldsymbol{x}), \underbrace{\mathbf{V}^\dagger \leftarrow \text{Self-Attention}(\mathbf{V})}_{\text{\% Equation (4.12)}}, \boldsymbol{z} \leftarrow h_\psi(\mathbf{V}^\dagger)$$

$$\underbrace{\arg\min\{\mathbb{E}[\overbrace{\mathcal{L}_{\text{surv}}(q(\boldsymbol{z}), t_i)}^{\text{\% Equation (4.13)}} + \overbrace{\mathcal{L}_{\text{FDV}}(\boldsymbol{z}, \boldsymbol{t}_i^{\text{emb}})}^{\text{\% Equation (4.14)}}]\}}_{\text{emb}_\theta, h_\psi, q'_\eta},$$

3. Time-to-event estimation:

$$\text{emb}_\theta, h_\psi, q_\eta \leftarrow \arg\min\{\mathbb{E}[\mathcal{L}_{\text{surv}}(q(\boldsymbol{z}^\dagger), t_i)]\}$$

4. Making predictions:

$$\hat{t} \triangleq q_\eta(h_\psi(\text{emb}_\theta^\dagger(\boldsymbol{x}^*))),$$

where $\dagger, \ddagger$ represent self-attention modules.

Covariate embedding To compute the relevance and similarities among covariates, we embed all covariates $\boldsymbol{x}_i \in \mathbb{R}^d$ into $\mathbb{R}^{d\times s}$ space, denoted by $\mathbf{V}_i \in \mathbb{R}^{d\times s}$, for simplicity we will omit subject index i from here on. Specifically, for continuous variables, we split its range into s bins based on the percentiles and let columns of $\mathbf{V}_i$ be the landmark vectors for these bins. After creating these anchor vectors, other values can be obtained by numerical interpolation. Denote the embedding function as $\text{emb}_\theta(\cdot) \in \mathbb{R}^1 \to \mathbb{R}^s$, and parametrized by θ.

For simplicity, we use linear interpolation for other values between landmarks, *i.e.,* if $x^{(b)} \leqslant x < x^{(b+1)}$, then $\text{emb}_\theta(x) = \text{emb}_\theta(x^{(b)}) + \frac{\text{emb}_\theta(x^{(b+1)}) - \text{emb}_\theta(x^{(b)})}{x^{(b+1)} - x^{(b)}}(x - x^{(b)})$. Note that this can also be achieved by b-splines or other suitable numerical interpolation methods.

For nominal categorical variables, we simply initialize it with a binary embedding matrix and then apply a simple linear transformation, we can get $\mathbb{R}^s$ representations. For example, for a categorical variable with K levels, $\mathbf{W}^{K\times s}$ is the embedding matrix

for this variable. $e_k^\top$ is a row vector with only k-th position equals 1, other positions are 0s. Then $e_k^\top \mathbf{W}^{K\times s}$ is the k-th level's embedding vector.

Time-to-event embedding To maximize the MI between covariate embedding and the time-to-event within the contrastive learning framework, time-to-event information (t_i, δ_i) need to be embedded into $\mathbb{R}^s$ space, *i.e.,* $\boldsymbol{t}_i^{emb} = \text{emb}_\psi(t_i) \in \mathbb{R}^s$. For subjects with observed events, *i.e.* $\delta_i = 1$, the embedding is similar to the embedding method we have discussed for continuous covariates. For censored subject, we would mark the bins after the censored time with uniformly distributed probabilities, since the exact event time is unknown.

Random masking Adding random mask in the training process can improve the prediction efficiency and interpretability Fong and Vedaldi (2017); Phillips et al. (2019), especially when missingness is a problem. In practice, during the training process, we add random masking to the input embedded representations.

Self-attention and missing data Literature has shown the privilege of adapting attention mechanisms to dealing with missing data problems (Wu et al., 2020). We assume that the relative importance of each pair of covariates are transferable between different subjects. To achieve this, we adopt a pairwise similarity for all covariates embeddings. Let's set $\mathbf{V} = \text{emb}_\theta(\boldsymbol{x})$ denotes the embedding matrix for covariate $\boldsymbol{x}$, which has $d \times s$ dimensions, $[\mathbf{V}]_k$ denotes the k-th column vector of the embedding matrix $\mathbf{V}$.

$$[\mathbf{V}^\dagger]_\mathbf{k} = \frac{\left\{\sum_\mathbf{m} \tilde{\mathbf{f}}([\mathbf{V}]_\mathbf{k}, [\mathbf{V}]_\mathbf{m})\mathbf{g}([\mathbf{V}]_\mathbf{m})\right\}}{\sum_\mathbf{m} \tilde{\mathbf{f}}([\mathbf{V}]_\mathbf{k}, [\mathbf{V}]_\mathbf{m})}, \tag{4.12}$$

where $\tilde{f}([\mathbf{V}]_k, [\mathbf{V}]_m) = \mathbf{M}_k \cdot f([\mathbf{V}]_k, [\mathbf{V}]_m)$ is the masked pair-wise similarity scores. $f([\mathbf{V}]_k, [\mathbf{V}]_m)$ characterizes the similarity between query ($[\mathbf{V}_j]_k$) and key ($[\mathbf{V}_j]_m$), which can be an embedded Gaussian, $f([\mathbf{V}]_k, [\mathbf{V}]_m) = \exp([\mathbf{V}]_k[\mathbf{V}]_m^\top)$. $g(\cdot)$ is a linear transformation.

4.6.3 Energy-based Contrastive Learning

Rooted from *mutual information* (MI), we consider to maximize the mutual information of the time embedding vector $\boldsymbol{t}_i^{emb} = \text{emb}_\psi(t_i) \in \mathbb{R}^s$ and the self-attentive representation $\mathbf{V}_\mathbf{i}^\dagger$. Define an encoding network $h_\psi(\cdot) \in \mathbb{R}^{d\times s} \to \mathbb{R}^s$ and specified with the set of parameters ψ. Let $\boldsymbol{z} = h_\psi(\mathbf{V}^\dagger) \in \mathbb{R}^\mathbf{s}$. The encoding network is optimized with the following energy-based contrastive MI estimator.

Illustrated by Guo et al. (2021), we will use a variant of the famous *Donsker-Varadhan* (DV) estimator (Donsker and Varadhan, 1983) to maximize the MI from an energy-based perspective, together with a Fenchel duality trick (Fenchel, 1949) to conclude a solution. Specifically, the *Fenchel-Donsker-Varadhan* (FDV) estiamtor takes the following form:

$$\mathcal{L}_{\text{FDV}}(\text{emb}_\theta, h_\psi) = \hat{I}^K_{\text{DV}}(\{\boldsymbol{z}(t_j), \boldsymbol{t}_j\}) + \frac{\sum\limits_{\neq j} \exp[(r(\boldsymbol{t}_{\neq j}, \boldsymbol{z}(t_j)) - r(\boldsymbol{t}_j, \boldsymbol{z}(t_j)))/\tau]}{\sum\limits_{\neq j} \exp[(\hat{r}(\boldsymbol{t}_{\neq j}, \boldsymbol{z}(t_j)) - \hat{r}(\boldsymbol{t}_j, \boldsymbol{z}(t_j)))/\tau]} - 1, \quad (4.13)$$

where $r(\boldsymbol{t}, \boldsymbol{z})$ is a similarity metric, which is defined as the dot products, i.e., $\frac{\boldsymbol{t}^\top \boldsymbol{z}}{\|\boldsymbol{t}\|\|\boldsymbol{z}\|}$, and τ denotes a learnable temperature parameter. $\hat{I}^K_{\text{DV}}(\{\boldsymbol{z}(t_j), \boldsymbol{t}_k\})$ is the *Donsker-Varadhan* (DV) estimator Donsker and Varadhan (1983) for the MI. Note that $\hat{r}, \hat{I}$ to denote evaluation-only mode for the corresponding functions (because they are the "fixed" u_ϕ and do not backpropagate parameter gradients).

4.6.4 Learning Objectives

Our objective including two parts, the negative log-likelihood (to account for censoring loss) and empirical loss minimization (to achieve better summary of the time-to-event distribution).

Likelihood model Above solution give a deterministic instead of a probabilistic solution. Rooted from Eq. 4.3, after encoding x into z with a parametric form (in

order to have expression for $\log p(z|x)$), we can use flow as decoding arm to achive a flexible representation of $\log p(t|z)$. But flow with 1-dimensional variable is tricky and may not work well. With $\log p(t|z)$, we can calculate likelihood for events and censoring respectively.

$$\mathcal{L}_{\text{Likelihood}}(\boldsymbol{z}_i^\dagger, t_i; \theta) = \delta_i \log p_\theta(t_i|\boldsymbol{z}_i^\dagger) + (1-\delta_i) \log S_\theta(t_i|\boldsymbol{z}_i^\dagger), \tag{4.14}$$

where the likelihood for the censoring is defined as,

$$\hat{S}_\theta(t_i|x_i) = \int_{t>t_i} p_\theta(t_i|\boldsymbol{z}_i^\dagger)dt$$

Empirical Loss Minimization We can compare time t to $\hat{t}$ directly, with adjustment to the censored subjects. To quantify the prediction loss, we may consider the following three loss functions: i) MSE loss, $\mathbb{E}(t_i - \hat{t}_i)^2$; ii) MAE loss, $\mathbb{E}|t_i - \hat{t}_i|$; iii) RAE loss, $\mathbb{E}[|t_i - \hat{t}_i|/t_i]$. When $e_i = 1$, we treat it as standard prediction loss, $\mathcal{L}(t_i, \hat{t}_i)$. When $e_i = 0$, we would only penalize it when $\hat{t}_i < t_i$(Chapfuwa et al., 2018). However, we do not backpropogate through this loss function in the optimization process, and this will serve as an performance metric.

4.7 Related Work

Machine learning and survival analysis Early attempts of combining machine learning techniques with statistical survival analysis, such as the Faraggi-Simon network (FS-network) (Faraggi and Simon, 1995), often failed to demonstrate a clear advantage over classical baselines (Schwarzer et al., 2000). Recent progresses in machine learning allow researchers to overcome the difficulties suffered by prior studies. For example, Katzman et al. (2018) showed that weight decay, batch normalization and dropout significantly improved the performance of FS-network. Li et al. (2019) analyzed survival curves based on clinical images using deep convolution neural net

(CNN). In addition to deep nets, Fernández et al. (2016) showed that Gaussian Process can be used to effectively capture the non-linear variations in CoxPH models, and Alaa and van der Schaar (2017) further proposed a variant that handles competing risks. Similar to these works, our VSI also draws power from recent advances in machine learning to define a flexible learner.

Bayesian survival analysis Bayesian treatment of survival models has a long history. Raftery et al. (1996) first considered modeling uncertainties for survival data, Zupan et al. (1999) reported probabilistic analysis under Bayesian setup. More recently, Fard et al. (2016) exploited the Bayesian framework to extrapolate priors, and Zhang and Zhou (2018) described a Bayesian treatment of competing risks. Closest to VSI is the work of *deep exponential family* model (DEF) survival model (Ranganath et al., 2016), where the authors introduced a Bayesian latent variable model to model both predictors x and survival time t. Unlike our VSI, DEF still imposes strong parametric assumptions on the survival distribution, and it's not clear how the censored observations are handled in DEF's actual implementation. Another key difference between DEF and VSI is the factorization of joint likelihood. As the VSI encoder will only seek to capture the latent components that are predictive of the survival time distribution, while DEF encoder also needs to summarize information required to reconstruct covariates x. We argue that our VSI factorization of joint probability is more sensible for survival time modeling, because modeling x not only adds model complexity but also introduces nuisance to the prediction of survival time t. For datasets with large covariates dimensions and noisy observations, the DEF features can be dominated by the ones predictive of x rather t, compromising the main goal of modeling the survival distribution.

Individual uncertainties and randomization The seminal work of Aalen (1994) first identified importance of accounting for the individual uncertainties, the

main culprit for the failure of classical survival models, which can be remedied by explicitly modeling the random effects (Hougaard, 1995). Alternatively, Ishwaran et al. (2008) presented *Random Survival Tree* (RST) to predict cumulative hazards using a tree ensemble, demonstrating the effectiveness of a randomization scheme for statistical survival models. Our approach differs from the above schemes by systematically account for individual uncertainty using the randomness of latent variables.

Direct modeling of survival distribution The pioneering work of Yu et al. (2011) advocated the prediction of individual survival distributions, which is learned using a generalized logistic regression scheme. This idea is further generalized in the works of Luck et al. (2017) and Fotso (2018). Recently, Chapfuwa et al. (2018) explored the use of deep *Generative Adversarial Network* (GAN) to capture the individual survival distribution, which is closest to our goal. Compared the proposed VSI, the adversarial learning of survival distribution is largely unstable, and its success crucially relies on the use of *ad-hoc* regularizers.

4.8 Experiments

To validate the effectiveness of the proposed VSI and ELSSA, we benchmarked their performances against the following representative examples from both statistical and machine learning survival analysis schemes: AFT-Weibull, CoxPH, LASSO-based CoxNet (Simon et al., 2011), Random Survival Forest (RSF) (Ishwaran et al., 2008) and deep learning based DeepSurv (Katzman et al., 2018). To fully appreciate the gains from using a variational setup, we further compared the results with the baselines discussed in Section 4.4.6, namely, the feed-forward model (MLP) and VSI model without the backward encoding arm $q_\phi(z|t,x)$ (VSI-NoQ).

For data preparation, we randomly partition data into three non-overlapping sets

for training (60%), validation (20%) and evaluation (20%) purposes respectively. All models are trained on the training set, and we tune the model hyper-parameters wrt the out-of-sample performance on the validation set. The results reported in the paper are based on the evaluation set using best-performing hyper-parameters determined by the validation set. We apply ADAM optimizer with learning rate of 5×10^{-4} during training, with mini-batches of size 100. The early stopping criteria of no improvement on the validation datasets is enforced.

To ensure fair comparisons, all deep-learning based solutions are matched for the number parameters and similar model architectures & similar hyper-parameter settings. TensorFlow code to replicate our experiments can be found at `https://github.com/ZidiXiu/VSI/`. The details of the VSI model setups are related to the Appendix C.1. ELSSA is implemented with PyTorch, the full implementation can be found at `https://github.com/ZidiXiu/ELSSA/`.

4.8.1 Evaluation Metrics

To objectively evaluate these competing survival models, we report a comprehensive set of distribution-based and point-estimate based scores to assess model performance, as detailed below.

Concordance Index (C-Index) is commonly used to evaluate the consistency between the model predicted risk scores and observed event rankings (Harrell et al., 1982). Formally, it is defined as

$$\text{C-Index} = \frac{1}{|\mathcal{E}|} \sum_{(i,j)\in\mathcal{E}} \mathbb{1}_{f(x_i)>f(x_j)}$$

, where $\mathcal{E} = \{t_i < t_j | \delta_i = 1\}$ is the set of all valid ordered pairs (event i before event j) and $f(x)$ is a scalar prediction made by the model. Higher is better.

Time-dependent Concordance Index is a distribution generalization of the scalar risk score based C-Index (Antolini et al., 2005), which is computed from the predicted survival distribution. Formally it is given by

$$\mathcal{C}^{\text{td}} = P(\hat{F}(t_i|x_i) > \hat{F}(t_i|x_j)|t_i < t_j).$$

, where $\hat{F}$ denotes the model predicted cumulative survival function. We report the results using the following empirical estimator

$$\hat{\mathcal{C}}^{\text{td}} = \frac{1}{|\mathcal{E}|} \sum_{(i,j)\in\mathcal{E}} \mathbb{1}_{\hat{F}(t_i|x_i)>\hat{F}(t_j|x_j)}$$

Kolmogorov-Smirnov (KS) distance For synthetic datasets, we also report the KS distance (Massey Jr, 1951) between the predicted distribution and the ground truth. KS computes the maximal discrepancy between two cumulative densities, *i.e.*,

$$\text{KS} = \sup_t |F_1(t) - F_2(t)|,$$

and a lower KS indicates better match of two distributions.

Test log-likelihood We also report the average log-likelihood on the held-out test set. A higher score indicates the model is better aligned with the ground-truth distribution in the sense of KL-divergence. Additionally, we also evaluate the spread of empirical likelihood wrt the models. In the case of an expected log-likelihood tie, models with the more concentrated log-likelihoods are considered better under the maximal entropy principle (Chen et al., 2018) (*i.e.*, as observed instances received more uniform/similar likelihoods, better generalization of the model is implied).

Coverage Rate To quantify the proportion of observed time covered in the predicted personalized time-to-event distributions, we calculated the coverage rate for different percentile ranges. For subjects with event observations, the coverage rate is defined as the proportion of observations fall in the percentile ranges $[l, u]$ of

the predicted distributions, where l, u respectively denotes lower and upper quantile of percentile ranges, *i.e.*,

$$\text{Cover Rate}_{\text{events}}(l, u) = \frac{1}{n_e} \sum_{y_i \in \mathcal{D}_e} \mathrm{I}(l < t_i < u)$$

In our experiments, we report coverage rates of events at percentile range $[l, u] \in \{[0.05, 0.95], [0.1, 0.9], [0.15, 0.85], [0.2, 0.8], [0.25, 0.75], [0.3, 0.7], [0.35, 0.65], [0.4, 0.6], [0.45, 0.55]\}$ of the predicted personalized distributions. For censoring, we calculate the proportion of the censoring time happened before the percentiles of predicted range, since the true time-to-event for censoring is happened after censoring time,

$$\text{Cover Rate}_{\text{censor}}(l) = \frac{1}{n_c} \sum_{y_i \in \mathcal{D}_c} \mathrm{I}(t_i \leqslant l)$$

We evaluated the coverage rate for censoring at $l \in \{0.1, 0.2, \cdots, 0.9\}$ percentiles.

For all coverage rates, a higher score implies better performance. Coverage rates for events and censoring should be considered together to evaluate model performance.

4.8.2 Synthetic datasets

Following Bender et al. (2005) we simulate a realistic survival data based on the German Uranium Miners Cohort Study in accordance with the Cox-Gompertz model

$$T = \frac{1}{\alpha} \log \left[1 - \frac{\alpha \log(U)}{\lambda \exp(\beta_{\text{age}} \times \text{AGE} + \beta_{\text{radon}} \times \text{RADON})} \right]$$

, with $U \sim \text{Unif}[0, 1]$. This model simulates the cancer mortality associated with radon exposure and age. Model parameter are derived from real data: $\alpha = 0.2138$, $\lambda = 7 \times 10^{-8}$, $\beta_{\text{age}} = 0.15$ and $\beta_{\text{radon}} = 0.001$. Covariates are generated according to

$$\text{AGE} \sim \mathcal{N}(24.3, (8.4)^2), \text{ RADON} \sim \mathcal{N}(266.8, (507.8)^2),$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. We simulate uniform censoring within a fixed time horizon c, *i.e.*, we let $C_i \sim \text{UNIF}(0, c)$, then $\delta_i = \mathbb{1}_(T_i < C_i)$ and $T_i = C_i$ if $C_i < T_i$. By setting different upper bounds c for censoring, we achieve different observed event rates, 100%($c = \infty$), 50%($c = 100$) and 30%($c = 70$). For each simulation we randomly draw $N = 50k$ iid samples.

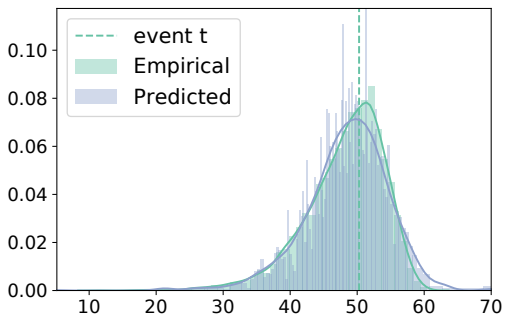

FIGURE 4.1: $\delta = 1$ (observed)

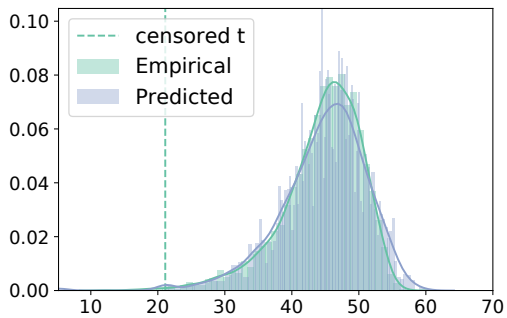

FIGURE 4.2: $\delta = 0$ (censored)

FIGURE 4.3: Two simulated time-to-event distributions with 30% event rate showing that VSI successfully predicts the underlying distributions from covariates. (left: events, right:censoring)

Prediction of subject-level distribution In practice, for each subject we only observe one t from its underlying distribution. Our goal is to accurately predict the underlying distribution from the covariates x alone (since t and δ are not observed at test time), by learning from the observed instances. Figure 4.3 compares our VSI prediction with the ground-truth for two random subjects, which accurately recovers of individual survival distribution for both observed (Figure 4.3(a)) and censored cases (Figure 4.3(b)). To systematically evaluate the consistency between the predicted and the true distributions, we compare average KS distance from models trained with various event rates in Table 4.1. Since the underlying generative process is based on CoxPH model, we consider the results from CoxPH as the oracle reference, since there is no model mis-specification. At 100% event rate (*i.e.*, complete observation), apart from the oracle CoxPH, all models perform similarly. The VSI variants give

Table 4.1: KS statistic for simulation study.

Event Rate	100%	50%	30%
CoxPH (Oracle)	0.027	0.032	0.027
AFT-Weibull	0.057	0.058	0.068
MLP	0.047	0.063	0.064
VSI-NoQ	0.049	0.068	0.066
VSI	**0.044**	0.052	**0.059**
ELSSA	0.049	**0.050**	0.066

Table 4.2: Model performance summary for simulation study based on C_{td}, C-Index and average test log-likelihood. Confidence Intervals for C-Index provided in the Appendix C.4. For NA entries, the corresponding evaluation metric can not be applied.

Models	C_{td}			C-Index Raw			log-likelihood		
	100%	50%	30%	100%	50%	30%	100%	50%	30%
CoxPH	**0.757**	0.755	0.761	0.773	0.781	0.793	NA	NA	NA
Coxnet	NA	NA	NA	0.776	0.784	0.760	NA	NA	NA
AFT-Weibull	0.742	0.750	0.768	0.773	0.781	0.793	-4.43	-2.29	-1.47
RSF	0.631	0.638	0.608	0.701	0.718	0.712	-14.12	-8.02	-5.35
DeepSurv	NA	NA	NA	0.772	0.781	0.793	NA	NA	NA
MLP	0.744	0.751	0.770	0.772	0.781	0.793	-4.15	-2.22	-1.41
VSI-NoQ	0.748	0.749	0.763	0.772	0.781	0.793	-4.16	-2.22	-1.41
VSI	0.748	0.756	0.772	0.773	0.781	0.793	-4.15	-2.22	-1.40
ELSSA	0.754	**0.758**	**0.775**	0.773	0.781	0.793	**-3.46**	**-1.89**	**-1.15**

slightly better results compared with MLP and AFT-Weibull. As the proportion of observed events decreases, VSI remains the best performing model, closely followed by the parametric AFT-Weibull. Note that neither MLP nor VSI-NoQ matches the performance of VSI, which suggests that the full VSI design better accommodates censoring observations. ELSSA performs similar to VSI in this simulation setup.

Average log-likelihood and C-Index To validate the effectiveness of VSI, we also provide a comprehensive summary of model performance against other popular or state-of-the-art alternatives in Table 4.2, under various simulation setups with different evaluation metrics. VSI consistently outperforms its counterparts in terms of the average log-likelihood and time-dependent C-Index. Together with the obser-

vation that VSI also yields better KS distance (see Table 4.1), converging evidence suggests our VSI better predicts the individual survival distributions relative to other competing solutions. ELSSA beats VSI in most evaluation metrics consistently, suggesting the non-parametric contrastive learning based mutual information estimator does shape the representation better.

We also compared the raw C-Index and the corresponding confidence intervals using the weighted average of model predicted survival time (defined in Sec 4.4.5) as the risk score, and we did not find significant differences between alternative methods, as shown in Table 4.2 and Appendix C.4. Thus ELSSA can deliver comparable performance relative to models that are compatible with the data generating mechanism. Raw C-Index quantifies the corresponding pairs without considering the time horizon, thus the distinctions among good performing models are not significant.

To provide a more informative summary, We plot the test log-likelihood distributions for selected models in Figure 4.4. We can see that VSI and ELSSA log-likelihoods estimates are tighter and higher for both observed and censored observations, especially when we have low event rates. The $(0.10, 0.90)$ percentiles range for simulation studies please refer to Appendix C.4.

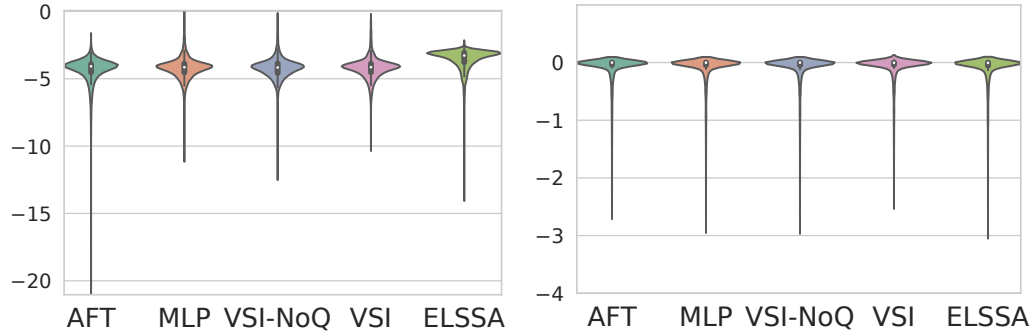

FIGURE 4.4: Test log-likelihood distributions for the 50% event rate simulation dataset. (left: events, right:censoring)

Coverage Plots In Figure 4.5, VSI achieves both relatively high coverage for

event (Figure 4.5(a)) and censored observations (Figure 4.5(b)), comparing to the oracle method CoxPH in this synthetic example. Note that while RSF performs better for the observed events, its performance on censored cases falls well below other solutions.

We refer the readers to our Supplementary Materials for additional simulations and analyses based on toy models.

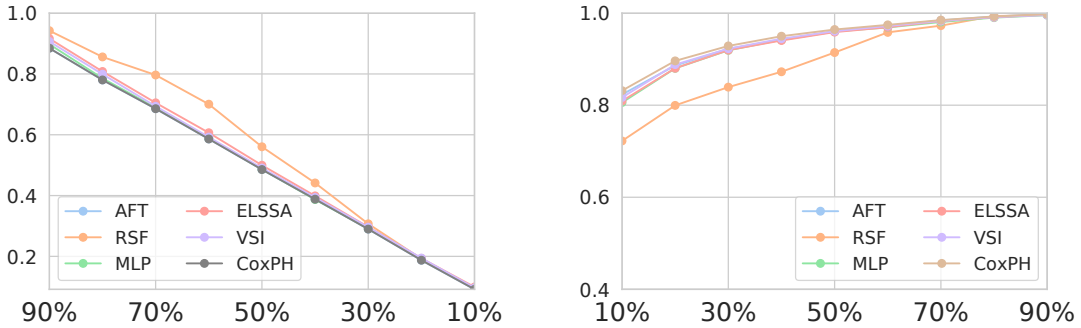

FIGURE 4.5: Test coverage rate for the 50% event rate simulation dataset. (left: events, right: censoring)

4.8.3 Real-World datasets

Moving beyond toy simulations, we further compare VSI to competing solution on the following three real-world datasets, *i*) `FLCHAIN` (Dispenzieri et al., 2012): a public dataset to determine whether the elevation in free light chain assay provides prognostic information to the general population survival, *ii*) `SUPPORT` (Knaus et al., 1995): a public dataset for a prospective cohort study to estimate survival over seriously ill hospitalized adults for 180 days period, and *iii*) `SEER` (Ries et al., 2007): a public dataset aim to study cancer survival among adults, which contains 1988 to 2001 information, provided by U.S. Surveillance, Epidemiology, and End Results (SEER) Program. In this experiments, we used 10-year follow-up breast cancer subcohort in `SEER` dataset. We follow the data pre-processing steps outlined in Chapfuwa et al.

(2018). To handle the missing values in data, we adopt the common practice of median imputation for continuous variables and mode imputation for discrete variables.

Table 4.3: Summary Statistics for Real Datasets.

	FLCHAIN	SUPPORT	SEER
N	7,894	9,105	68,082
Event rate(%)	27.5	68.1	51.0
p(cat)	26(21)	59(31)	789(771)
NaN(%)	2.1	12.6	23.4
Max event t	4998_{days}	1944_{days}	120_{months}
Loss of Info(%)	10.45	1.57	0.0

Summary statistics of the datasets are shown in Table 4.3, where N is the total number of observations, p denotes the total number of variables after one-hot-encoding, NaN(%) stands for the proportion of missingness in covariates, and loss of information stands for the proportion of censoring observations happened after the maximum event time t.

In Table 4.4 we compare the C-Indices and average log-likelihood. The advantage of ELSSA and VSI is more evident for the more challenging real datasets, especially in the cases of low observed event rates. For example, with 30% event rate, in SUPPORT dataset, VSI Confidence Interval for raw C-Index as (0.809, 0.846), while the standard CoxNet is only (0.763,0.805) and AFT (0.782,813), *i.e.*, the overlaps with that of VSI are very small. Similar results were observed for other datasets and baseline solutions. ELSSA shows remarkable robustness against data incompleteness in a real-world scenario, achieving the best results according to all three metrics. For VSI the raw C-Index is computed from the weighted average of VSI predicted distribution, please refer to Appendix C.5 for more details.

In Figure 4.6, the distribution of log-likelihood is more concentrated, in addition to a higher mean. To quantitatively evaluate the concentration, we report the difference between 10% and 90% quantiles of log-likelihood in Table 4.5. The quantile

Table 4.4: Summary for Real Datasets based on C-Index and average log-likelihood. Confidence Intervals for C-Index are provided in Appendix C.5. NA implies the corresponding evaluation metric can not be evaluated.

Models	C_{td}			C-Index Raw			log-likelihood		
	FLCHAIN	SUPPORT	SEER	FLCHAIN	SUPPORT	SEER	FLCHAIN	SUPPORT	SEER
Coxnet	NA	NA	NA	0.790	0.797	0.819	NA	NA	NA
AFT-Weibull	0.777	0.752	NA	0.792	0.797	NA	-3.09	-4.39	NA
RSF	NA	NA	NA	0.771	0.751	0.796	NA	NA	NA
DeepSurv	NA	NA	NA	0.785	0.678	NA	NA	NA	NA
MLP	0.775	0.768	0.821	0.751	0.811	0.811	-1.91	-2.86	-2.50
VSI-NoQ	0.745	0.772	0.820	0.745	0.824	0.809	-2.45	-2.79	-2.50
VSI	**0.787**	**0.775**	**0.824**	**0.792**	**0.827**	**0.826**	**-1.85**	**-2.74**	**-2.49**
ELSSA	**0.791**	**0.794**	**0.830**	**0.799**	**0.861**	**0.828**	**-1.72**	**-2.56**	**-2.32**

Table 4.5: Quantile ranges for log-likelihood in real-world datasets. Note AFT did not converge to reasonable solutions for SEER.

Models	Observed			Censored		
	FLCHAIN	SUPPORT	SEER	FLCHAIN	SUPPORT	SEER
AFT	2.491	4.706	NA	**0.468**	1.850	NA
MLP	2.970	4.273	1.780	0.518	1.540	0.623
VSI-NoQ	7.34	4.744	1.801	0.559	1.634	0.529
VSI	**2.213**	**4.143**	**1.718**	0.537	**1.354**	**0.508**

ranges of VSI are considerably smaller compared to alternative solutions under most experimental settings. This verifies VSI enjoys better model robustness compared to other popular alternatives, especially in the case of high censoring rates.

Together with the coverage plots in Figure 4.7, VSI and ELSSA has relative high coverage for both events and censoring cases which indicates better performance in

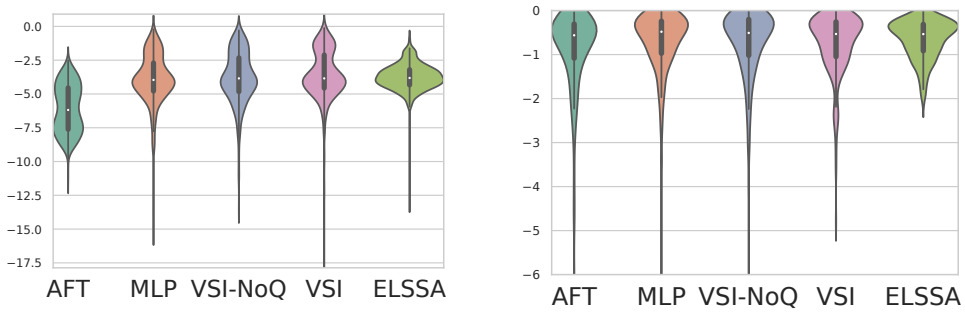

FIGURE 4.6: log-likelihood distributions for `SUPPORT` Dataset, (left: events, right:censoring)

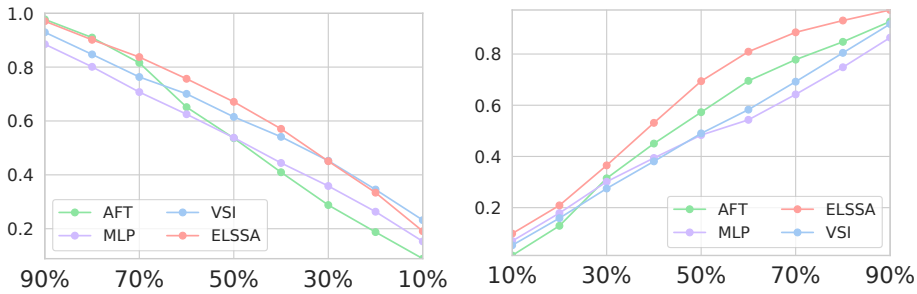

FIGURE 4.7: Coverage rate for SUPPORT Dataset, (left: events, right: censoring)

capturing the true event time in challenging real-world datasets. The consistency of those results have been verified through repeated runs on these three datasets. For more detailed results please refer to Appendix C.5.

4.9 Conclusions

4.9.1 Variational Survival Inference

We presented an approach for learning time-to-event distributions conditioned on covariates in a nonparametric fashion by leveraging a principled variational inference formulation. The proposed approach, VSI, extends the variational inference framework to survival data with censored observations. Based on synthetic and diverse real-world datasets, we demonstrated the ability of VSI to recover the underlying unobserved time-to-event distribution, as well as providing point estimations of time-to-event for subjects that yield excellent performance metrics consistently outperforming feed-forward deep learning models and traditional statistical models.

4.9.2 ELSSA

We introduced a self-attention based model to capture helpful information for time-to-event prediction. A key novelty of this approach is the integration of contrastive mutual information based loss that non-parametrically maximizes the informativeness of learned data representations. The effectiveness of our approaches has been

intensively validated with synthetic and real-world benchmarks, showing improved performance relative to competing solutions.

Generalization to longitudinal inputs

With the integration of contrastive mutual information based loss, we non-parametrically maximizes informativeness of learned data representations with time-to-event. By aggregating the temporal embeddings, we can obtain dynamic latent representations with temporal information embedded that generalizes our method to accommodate longitudinal inputs in the future.

Based on notations in Sec. 4.5, for subject i, let $\boldsymbol{x}_i(t) \in \mathbb{R}^d, \boldsymbol{m}_i(t) \in \mathbb{R}^d$ denote the covariates and missing masks at time t. Note that $\boldsymbol{t}_i$ is the observation time sequences with length l_i, $\boldsymbol{t}_i = \{t_{ij} : j \in (0, 1, \ldots, l_i)\}$. For simplicity, in the following discussions we will ignore the subject i.

Followed above notations, $\boldsymbol{z} = h_\psi(\text{emb}_\theta(\boldsymbol{x}(t_j)))$ denotes the localized latent representation at time j (subject index i omitted). Then for a single subject, the set of latent embeddings $\boldsymbol{Z}(t_l) = \{\boldsymbol{z}(t_j), \forall j \leqslant l\}$ contains the dynamic representations at all observed time points.

Attention over the horizon Give a series of history temporal embeddings up to time t_j, we summarize all informative embeddings happened before it. Again we consider self-attention which uses the pairwise influence for the previous embeddings before the time t_j and t_j itself. This generates a hidden vector which summarizes all the temporal embedding before that time.

$$\boldsymbol{z}^{\ddagger}(t_j) = \frac{\sum_k^{j-1} f(\boldsymbol{z}(t_j), \boldsymbol{z}(t_k)) g(\boldsymbol{z}(t_k))}{\sum_k^{j-1} f(\boldsymbol{z}(t_j), \boldsymbol{z}(t_k))}, \tag{4.15}$$

in attention model terminology, $\boldsymbol{z}(t_j)$ is the query, $\boldsymbol{z}(t_k)$ is the key and $g(\boldsymbol{z}(t_k))$ is the value. Similar to the self-attention step in covariate embedding steps, $f(\cdot, \cdot)$ is

an embeded Gaussian which quantifies the similarities and $g(\cdot)$ is linear transformation. The aggregated self-attentive vector can be used for predicting time-to-event dynamically.

Acknowledgements

This research was supported in part by by NIH/NIBIB R01-EB025020.

5

Conclusions

In this chapter, we conclude this dissertation by summarizing our contributions and discussing directions for future work.

5.1 Summary

Establishing reliable and robust predictions in healthcare data has been a challenging topic for decades. Over the years, extensive efforts have been made to build risk prediction models in the presence of complex inputs, missingness (Rubin, 1976) and imbalanced data (Goldstein et al., 2017) problems. In general, classical statistical prediction methods like regression lack flexibility and have constraints on the data distributions, that are not appropriate for real-world datasets. The above limitations are impeding disentangling the complex correlations in the datasets. Non-parametric statistical machine learning methods require large sample sizes of training data and are criticized for poor interpretability and robustness against new observations. Recently, representation learning has gained considerable attention in recognition of growing data complexity (Bengio et al., 2013). This dissertation investigated novel statistical machine learning approaches for risk prediction in clinical data from dif-

ferent perspectives based on the advances in representation learning techniques.

The existing clinical decision support tool, Early Warning Scores (EWS) (Smith et al., 2013; O'Brien et al., 2020), can provide clinical alerts to clinicians about patients at immediate risk of an adverse outcome. But it has been shown having no meaningful influence on patient outcomes with low classification accuracy ~ 0.74 AUC (Bedoya et al., 2019). EWS only uses seven vital signs for prediction, which does not fully take advantage of the modern EHR system. Available covariates including time-varying clinical data ,*i.e.*, vital signs, laboratory tests, medications, and their demographic information. Such high-dimensional, multivariate longitudinal survival datasets with rare event rates ($\leqslant 5\%$) provide rich information and challenges for detecting the risk of patient deterioration. To dissect the complicated prediction task, we first address the rare event classification problem in Chapter 2. Later on, we generalize the imbalanced learning framework to multi-classification situations in Chapter 3. In Chapter 4, we focus on time-to-event prediction and discuss the potential generalization to longitudinal inputs. Apart from the motivating inpatient early warning dataset we studied at length in Chapter 2, we also investigated other challenging real-world clinical datasets, *e.g.*, the `COVID` dataset for predicting mortality.

In the wake of the big data era, classical statistical models for risk prediction models become inadequate. Those methods tend to suffer from the lack of flexibility, individualization, and robustness against high-dimensional data (Lu et al., 2020). Chapter 2 presents a novel extreme representation-based variational inference framework VIE for individualized risk prediction under rare event cases. VIE differentiates events and non-event by imposing an extreme-value-theory based prior and joins the merits of additive monotonic neural networks to disentangle the latent representations. Our model acknowledges representation uncertainties while at the same time embraces improved interpretability, generalization, and robustness.

Taking a novel perspective from causal machine learning, Chapter 3 learned causally informed source representation to attenuate the spurious relationships in multi-classification with imbalanced training samples. Specifically, ECRT adopts a causal invariance generating mechanism that can be transferred from data-rich classes to data-scarce classes. With non-linear independent components, the source representations are conditionally independent. We re-balance the imbalanced tasks with augmentation on the independent feature sources.

In Chapter 4, ELSSA provides a principled non-parametric solution to time-to-event estimation with the existence of censored subjects and incomplete data. By leveraging covariate embedding and self-attention mechanisms, structural information in the covariates is extracted to alleviate the impact of missing covariates. A key novelty of this approach is the integration of contrastive mutual information-based loss that non-parametrically maximizes the informativeness of learned data representations.

In summary, this dissertation introduces flexible risk prediction frameworks that acknowledge representation uncertainties, data heterogeneity and incompleteness. Altogether, it presents three contributions: improved efficient learning from imbalanced data, enhanced robustness to missing data, and better generalizability to out-of-sample subjects.

5.2 Future work

Researches stemming from this dissertation can be extended into a number of directions. Below is a brief discussion on some of the most promising and significant fields to improve clinical prediction w.r.t. modern EHR data.

Dynamic prediction based on longitudinal data. One notable feature of EHR data is the longitudinal history for patients. A wealth of literature explored the prediction of clinical longitudinal datasets with recurrent neural networks (Harutyun-

yan et al., 2019; Lee et al., 2019; Ren et al., 2019). In ELSSA, we successfully learned informative representations for the time-to-event prediction. By combining the temporal embeddings via the self-attention mechanism as discussed in Section 4.9.2, we will be capable of predicting the time-to-event based on the longitudinal inputs dynamically.

Long-tailed predictions. Rooted from the imbalanced classification problem discussed in Chapter 2 and Chapter 3, it's a natural shift from the multi-classification problem to long-tail predictions with continuous outputs. There're numerous applications like predicting length-of-stay in hospital (Huntley et al., 1998). In the settings of ECRT, we can transfer the learned model from the head part where more samples are presented to the tail part of the distribution where the number of samples is scarce. The augmentation strategy for the tailed part based on feature independence can fit in this scenario as well to enhance the learning efficiency (Hyvarinen and Morioka, 2016; Wang et al., 2017).

Another promising direction is density estimations for anomaly detection. Based on VIE, the disentangled latent space enjoys long-tailed distributions, which can separate event and non-event with extreme representations. The semi-parametric long-tailed latent representation learning framework can also fall into the range of anomaly detection with density estimation naturally (Steinwart et al., 2005).

Beyond right censoring. While in Chapter 4 we have focused on right censoring (lost-to-follow-up) to simplify the discussions, we are aware that interval censoring and left censoring are also common in clinical datasets. A similar log-likelihood can be easily computed with numerical integration within the discrete approximation to the time-to-event estimation framework.

Generalize to competing risk. In this dissertation, we focused on a single type of event outcome, *i.e.,* mortality. In time-to-event analysis, usually, the occurrence of the event of interest is preventing another event from happening. Extending from

standard settings, in competing risk settings which are common in clinical studies, apart from event of interests and censoring, the existence of a competing event brings challenges to survival analysis (Lau et al., 2009). Machine learning literature views the competing risk problem as a shared sub-network with the prediction of the event of interests (Lee et al., 2018).

Appendix A

Supplementary Material to "Variational Disentanglement for Rare Event Modeling"

When the prevalence of an event is extremely low, but the event itself has substantial importance, the methods to identify such targets are called rare event modeling. Accurate and robust modeling of rare events is significant in many fields, such as identifying patients in high-risk and hopefully to prevent adverse outcomes from happening based on early intervention.

The scarcity of rare cases can cause extreme imbalanced among the dataset. Therefore, rare event modeling is challenging for most standard statistical approaches. As we discussed in the main text, careful statistical adjustments and new methodologies are required to approach such imbalance. Otherwise, the classifiers would be driven to the majority side and give misleading results. Also, the lack of representation in the minority class may cause unadjusted models to wrongly capturing spurious features that cannot generalize well to other observations. The apex of the risk curve or the mass of risk density usually overlays with the tail of the feature representation distribution, as illustrated in Figure A.1, traditional statistical meth-

ods (such as Gaussian based approaches) often ill perform at the tail end, which can lead to lack-of-fit and poor generalization ability.

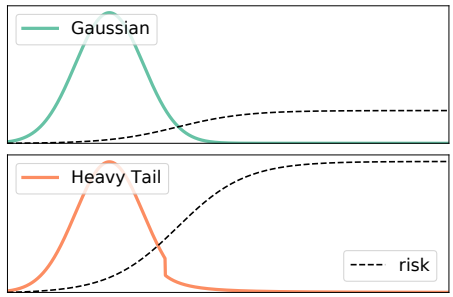

FIGURE A.1: Feature representation mismatch at the tail parts. The heavy-tailed distribution can exploit extreme behavior in the latent space.

We approach a solution to such challenges with a variational representation learning scheme that models disentangled extreme representations. Further, we design a robust, powerful prediction arm that combines the merits of a generalized additive model and isotonic neural net.

A.1 Derivation of Mixed GPD tail distribution

An important theory in *Extreme value theory* (EVT) shows that under some mild conditions, the conditional cumulative distribution of *exceedance* over a threshold u follows Generalized Pareto Distribution, GPD(u, ξ, σ) (McFadden, 1978), which has the cumulative distribution function (CDF) as:

$$G_{\xi,\sigma,u}(x) = \begin{cases} 1 - [1 + \xi(x-u)/\sigma]^{-\frac{1}{\xi}}, & \text{if } \xi \neq 0 \\ 1 - \exp[-(x-u)/\sigma], & \text{if } \xi = 0 \end{cases}$$

where σ is a positive scale parameter. According the shape parameter ξ, x could have different support. When $\xi < 0$, the exceedance x has bounded support $0 \leqslant x \leqslant u - \sigma/\xi$, otherwise x is bounded by 0 on the left. u is the location parameter. The

corresponding PDF is:

$$g_{\xi,\sigma,u}(x) = \begin{cases} \sigma^{-1}[1+\xi(x-u)/\sigma]^{-\frac{1}{\xi}-1}, & \text{if } \xi \neq 0 \\ \sigma^{-1}\exp[-(x-u)/\sigma], & \text{if } \xi = 0 \end{cases}$$

Thus the log-likelihood function is:

$$\log \text{likelihood}(x;\xi,\sigma,u) = \begin{cases} -\log\sigma - (\frac{1}{\xi}+1)\log[1+\xi(x-u)/\sigma], & \text{if } \xi \neq 0 \\ -\log\sigma - (x-u)/\sigma, & \text{if } \xi = 0 \end{cases}$$

To enable modeling of the extreme representations, we adopt the Generalized Pareto Distribution as the tail part of our new variational prior, and the regular bulk representations $z \leqslant u$ with a standard Gaussian distribution. Then mixed extreme tail distribution has the form (McNeil, 1997),

$$F(z) = P(Z \leqslant z) = P(Z \leqslant u) + (1 - P(Z \leqslant u))F_u(z-u)$$

When $z > u$, the tail estimator is,

$$\hat{F}(z) = (1 - F_n(u))G_{u,\xi,\sigma,u}(z) + F_n(u)$$

to approximate $F(z)$. Now we show that $\hat{F}(z)$ also has a GPD distribution with same ξ and the following scale and location parameters,

$$\begin{cases} \tilde{\sigma} = \sigma(1-F_n(u))^{\xi}, \tilde{u} = u - \tilde{\sigma}((1-F_n(u))^{-\xi}-1)/\xi, & \text{if } \xi \neq 0 \\ \tilde{\sigma} = \sigma, \tilde{u} = u + \tilde{\sigma}\log(1-F_n(u)), & \text{if } \xi = 0 \end{cases}$$

When $\xi = 0$,

$$\begin{aligned} \hat{F}(z) &= (1-F_n(u))(1-\exp(-(x-u)/\sigma)) + F_n(u) \\ &= 1 - (1-F_n(u))\exp(-(x-u)/\sigma) \\ &= 1 - \exp(\log(1-F_n(u)))\exp(-(x-u)/\sigma) \\ &= 1 - \exp(-\frac{1}{\sigma}(x-u-\sigma\log(1-F_n(u)))) \\ &= 1 - \exp(-\frac{1}{\tilde{\sigma}}(x-\tilde{u})) \end{aligned}$$

When $\xi \neq 0$,

$$
\begin{aligned}
\hat{F}(z) &= (1 - F_n(u))(1 - (1 + \xi(x-u)/\sigma)^{-\frac{1}{\xi}}) + F_n(u) \\
&= 1 - (1 - F_n(u))(1 + \xi(x-u)/\sigma)^{-\frac{1}{\xi}} \\
&= 1 - [(1 - F_n(u))^{-\xi}(1 + \xi(x-u)/\sigma)]^{-\frac{1}{\xi}} \\
&= (1 - F_n(u))^{-\xi} + (1 - F_n(u))^{-\xi} \cdot \xi(x-u)/\sigma \\
&= \frac{1}{(1 - F_n(u))^{\xi}} + \frac{\xi(x-u)}{\sigma(1 - F_n(u))^{\xi}} \\
&= \frac{\sigma}{\tilde{\sigma}} + \frac{\xi(x-u)}{\tilde{\sigma}} \\
&= 1 + \frac{\sigma - \tilde{\sigma} + \xi(x - \tilde{u})}{\tilde{\sigma}} \\
&= 1 + \frac{\xi(x - \tilde{u} + \xi^{-1}\sigma - \xi^{-1}\tilde{\sigma})}{\tilde{\sigma}}
\end{aligned}
$$

Therefore, $\tilde{u} = \tilde{u} - \xi^{-1}\sigma - \xi^{-1}\tilde{\sigma}$.

A.2 Implementation Details

Our main algorithm was written in PyTorch (version 1.3.1) (Paszke et al., 2017). The experiments were conducted on an Intel(R) Xeon(R) and Tesla P100-PCIE-16GB GPU (except for the `COVID` dataset). The `COVID` dataset were stored and analyzed on a protected virtual network space with Inter(R) Xeon(R) Gold 6152 CPU 2.10GHz 2 Core(s).

A.2.1 Model Structure.

In VIE, we end up optimizing the following objective,

$$
\max_{\theta,\phi} \min_{\nu} \{\mathbb{E}_{x,y\sim\mathcal{D}}[\Psi_\beta(x, y; p_\theta, q_\phi) - \lambda\Gamma(p_\theta, q_\phi, \nu)]\}, \tag{A.1}
$$

where

$$\Psi_\beta(x, y; p_\theta, q_\phi) = \mathbb{E}_{Z\sim q_\phi(z|x)}[\log p_\theta(y|Z)] - \beta \mathrm{KL}[q_\phi(z|x)||p(z)],$$

Note that the GPD parameters (ξ, σ) are absorbed in ϕ, and hyperparameter u is used in the GPD prior $p(z)$. u is set to be $F_z^{-1}(0.99)$ in all experiments. When the event rate is $\geqslant 1\%$, we set $\lambda, \beta = (1 \times 10^{-3}, 1 \times 10^{-5})$, otherwise we shrink the parameters to $\lambda, \beta = (1 \times 10^{-4}, 1 \times 10^{-6})$.

More concretely, the constituting parts $p_\theta(y|z)$, $p_\theta(z|x)$, $q_\phi(z|x)$ and $\nu(z)$ are specified as follows

$$\begin{aligned}
p_\theta(y|z) &\leftarrow \Phi(H(z;\theta)) \text{ Log-Log link } (A.3), \\
H(z;\theta) &\leftarrow \text{ Additive Monotone Neural Net (A.4) with} \\
p(z) &\leftarrow \text{Mixed GPD } (u, \xi_p, \sigma_p) \ (2.6), p = 4 \\
q_\phi(z|x) &\leftarrow \text{Inverse Autoregressive Flow } (2.8), \text{nstep} = 5 \\
\nu(z) &\leftarrow \text{Standard neural network.}
\end{aligned} \tag{A.2}$$

Pseudo-code for VIE is presented in Algorithm 4. In all experiments, AMNN, IAF, $\nu(z)$ are specified in terms of two-layer MLPs of 32 hidden units with Rectified Linear Unit (ReLU) activation functions. The initial encoder `Init-Encoder` is specified as a three-layers MLPs of 32 hidden units. We set the minibatch size to $m = 200$. The critic network $\nu(x)$ uses the RMSprop optimizer with learning rate 1×10^{-3}, other parts of the algorithm used Adam optimizer with learning rate 1×10^{-4}. To avoid over-fitting, we set the dimension of latent space as 4 in all experiments.

Note we have used the *Complementary Log-Log* (CLL) link function for $p_\theta(y|z)$ in (A.3),

$$\Phi(a) = 1 - \exp(-\exp(a)), \tag{A.3}$$

Algorithm 4: Variational Inference with Extremals.

Data: $\mathcal{D} = (x, y)$. x: inputs, y: labels
Networks and parameters: `Init-Encoder`$(x, \epsilon; \phi)$: Initial encoder network; `IAF`$(z; \phi)$: recursive autoregressive neural network; $\nu(z; \omega)$: critic neural network;
`AMNN`$(z; \theta)$: additive monotonic neural net;
prior: $p_\psi(z)$ =`MixedGPD`$(z;\psi,u)$, ψ=$\{\xi_{\text{GPD}},\sigma_{\text{GPD}}\}$
Initialize: `Init-Encoder`, `IAF`, ν, `AMNN`, ψ
for *iteration* $k \in \{1, \ldots, K\}$ **do**
  Sample $\{(x_i, y_i)\}_{i=1}^m$ from $\mathcal{D}$, $\{\epsilon_i\}_{i=1}^m$ from $p(\epsilon)$
  $[\mu_0, \sigma_0]$ =`Init-Encoder`$(x, \epsilon; \phi)$
  Sample $z_{\texttt{pr}}$ from $p_\psi(z)$, $\boldsymbol{z}_0$ from $\mathcal{N}(\mu_0, \Sigma_0)$
  Compute $l_{\texttt{post}} := \log q_\phi(z_0|x)$
  for *step* $t \in \{1, \ldots, T\}$ **do**
    $[\mu_t, \sigma_t]$ =`IAF`$(z_{t-1}; \phi)$, $z_t = \mu_t + \sigma_t \odot z_{t-1}$
    $l_{\texttt{post}} = l_{\texttt{post}} - \sum(\log \sigma_t)$
  end
  $\log p_\theta(y|z_T) = \ell_{\text{CLL}}(y, \text{AMNN}(z_T; \theta))$
  Descend ω by $\nabla_\omega \frac{1}{m} \sum[\nu_\omega(z_{\texttt{pr}}) - \log \nu_\omega(z_T)]$
  Ascend $\Omega = \{\phi, \psi, \theta\}$ by
  $\nabla_\Omega \frac{1}{m} \sum[\log p_\theta(y|z_T) - \log \nu_\omega(z_T) - \text{KL}]$, where $\text{KL} = l_{\texttt{post}} - \log p_\psi(z_T)$
end

for the outcome model as opposed to the standard *Logistic* link $1/(1+\exp(-a))$. The CLL link is more sensitive at the tail end, so it is more frequently used in statistical models dealing with vanishing probabilities (Aranda-Ordaz, 1981).

To avoid collapsing to suboptimal local minimums, we train the encoder arm more frequently to compensate for the detrimental posterior lagging phenomenon (He et al., 2019). Our pseudo-code for VIE is summarized in Algorithm 1.

A.2.2 Numerical Integration.

Following (2.5), we divide region $[l, z_j]$ evenly into M bins of width $d_j = \frac{z_j - l}{M}$, with $z_{j,M} = z_j$. For the M bins, we select a random point $z_{j,k}^r$ in each bin. The integral approximation on support $[z_{j,k}, z_{j,k+1}]$ is the rectangular area $h_j(z_{j,k}^{(r)}) * d_j$. As a result, the integral $\int_0^{z_j} h(s; \theta) ds$ is approximated with $\sum_k^{M-1} h_j(z_{j,k}^{(r)}) d_j$. With this

approximation (2.5) can be written as:

$$H(z;\theta) = \sum_{j}^{p} \alpha_j d_j \sum_{k}^{M} h_j(z_{j,k}^{(r)}) + \gamma. \tag{A.4}$$

We set $M = 100$ and $l = -5$ in all the experiments.

A.2.3 Discussions on Evaluation Metrics.

In the main text, we reported AUC and AUPRC instead of single evaluation metrics, *e.g.*, overall accuracy or error rate. Standard statistical metrics like Brier Scores (BS) and Binary classification entropy (BCE) could be deceptive when the event rate is low, *e.g.*, $\leqslant 10\%$ (Schmid and Griffith, 2014). We will add BCE loss and the positive case BCE loss in the following sections in the simulation study for reference. Some poorly performed models can have relatively low BCE scores. In this case, the ground truth (Oracle) is the best reference we have.

A.3 Ablation Study

We examine model performance on two simulation strategies. The first one is the semi-synthetic dataset, which exploits the real-world covariates structures. The second one is a synthetic dataset based on our extreme representation assumptions.

A.3.1 Semi-synthetic Datasets

We synthesize a semi-synthetic dataset based on the Framingham study (Mitchell et al., 2010), a long-term cardiovascular survival cohort study. After quality control, $40,046$ subjects with nine covariates (four continuous and five categorical) are included.

We use a realistic model to synthesize data from the real-world covariates under varying conditions, *i.e.*, different event rates, sample size, nonlinearity, *etc.* More specifically, we use the coxPH-Weibull model (Bender et al., 2005) to simulate the

survival time of patients $T = \{\frac{-\log U}{\lambda \exp(g(x))}\}^{1/\nu}$, where $g(x)$ is either a linear function or a randomly initialized neural net. Our goal is to predict whether the subject will decease within a pre-specified time frame, *i.e.*, $T < t_0$. Via adjusting the cut-off threshold t_0, we can simulate different event rates. The details are shown in Algorithm 5.

Algorithm 5: Semi-synthetic Data

```
Extract covariates from Framingham Dataset ;
Set ν, λ (the parameters of cox-Weibull distribution);
Set time cut-bound t_0;
Decide g(x) : R^q → R form;
for i ∈ {1, ..., n} do
    Sample u_i from Unif(0, 1);
    Compute t_i = {−log u_i / (λ exp(g(x_i)))}^{1/ν};
    Compute y_i = 1[t_i < t_0];
    d_i = (y_i, x_i);
end
return D = {d_i; i = 1 ... n}
```

In our experiments, the performances when $g(\cdot)$ set as a randomized neural network or a linear function do not differ very much. For simplicity, we will present the results under the neural network settings. Apart from the results at 1% event rate given in the main text, we will show the results at 0.5% and 5% event rates here. The oracle results are calculated with plugging in the true $g(x_i)$ in Algorithm 5, and the randomness is from the generating scheme of the survival time t.

A.3.2 Additional Results for semi-synthetic datasets

In 1% event rate case presented in the main text, the AUC and AUPRC distributions are summarized in Figure A.2, which corresponds to the average and standard deviation values presented in Table 2.1. We further examine the cases with 0.5% and 5% event rates to evaluate our method's robustness. Results are summarized in Table A.1 and Table A.2 respectively. Apart from the threshold-free metrics AUC and

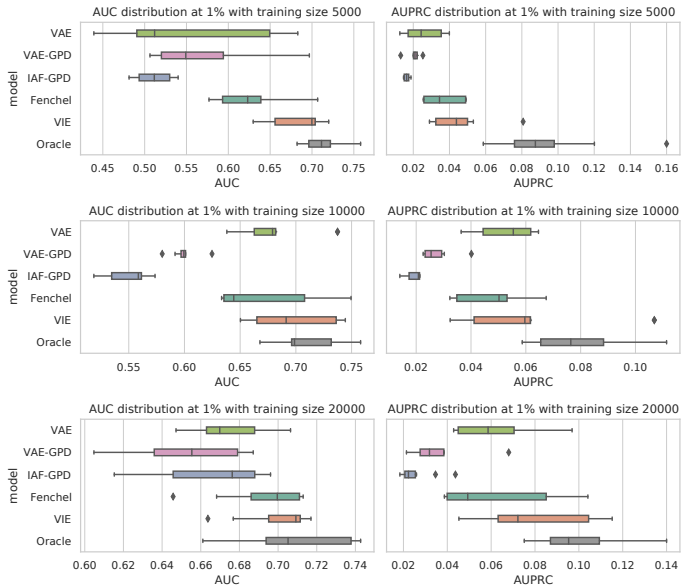

FIGURE A.2: Box plot of 10 independent 1% event rate semi-synthetic analysis.

AUPRC, we also presented Binary Cross-Entropy loss (BCE) and the BCE loss associated with true events (positive case losses). Note that for an imbalanced dataset, BCE loss can be misleading. In the model, VAE-GPD, which is poorly-behaved in AUC and AUPRC, can have relatively low BCE loss since the majority group overwhelms the minority (*more important*) group. We can refer to the BCE loss and positive case loss in the oracle results for reference. VIE performs consistently close to the oracle results, especially with low event rate and small training sample size, and Fenchel-GPD is in the second place.

Table A.1: Ablation study of VIE with 0.5% event rate in semi-synthetic settings.

	Average AUC (std) ↑			Average AUPRC (std)↑			Average BCE Loss (std) ↓			Average Positive Case Loss (std) ↓		
	n=5k	n=10k	n=20k	n=5k	n=10k	n=20k	n=5k	n=10k	n=20k	n=5k	n=10k	n=20k
VAE	0.494 (0.111)	0.623 (0.102)	0.697 (0.061)	0.007 (0.005)	0.017 (0.010)	0.020 (0.007)	0.498 (0.309)	**0.035** (0.005)	**0.031** (0.004)	**0.010** (0.010)	0.027 (0.005)	0.026 (0.005)
VAE-GPD	0.560 (0.045)	0.602 (0.044)	0.635 (0.045)	0.008 (0.002)	0.013 (0.010)	0.016 (0.003)	5.250 (2.096)	0.768 (0.258)	0.152 (0.222)	0.000 (0.000)	**0.004** (0.001)	**0.018** (0.008)
IAF-GPD	0.631 (0.039)	0.555 (0.042)	0.533 (0.061)	0.011 (0.007)	0.008 (0.002)	0.017 (0.013)	**0.032** (0.003)	0.038 (0.008)	0.043 (0.014)	0.027 (0.003)	0.024 (0.001)	0.027 (0.007)
Fenchel-GPD	0.615 (0.059)	0.652 (0.055)	0.667 (0.024)	0.022 (0.016)	0.021 (0.012)	**0.025** (0.008)	0.034 (0.004)	0.037 (0.006)	0.033 (0.002)	0.028 (0.005)	0.032 (0.006)	0.027 (0.002)
VIE	**0.654** (0.074)	**0.692** (0.076)	**0.693** (0.036)	**0.022** (0.010)	**0.027** (0.015)	0.024 (0.009)	0.041 (0.018)	0.036 (0.003)	0.032 (0.003)	0.026 (0.005)	0.030 (0.005)	0.026 (0.003)
Oracle	0.688 (0.618, 0.769)			0.043 (0.023, 0.071)			0.034 (0.028, 0.040)			0.029 (0.023, 0.035)		

A.3.3 Long-tailed Synthetic Datasets

We design the long-tailed synthetic datasets based on our proposed method, where the latent variable z enjoys a long-tailed distribution. The pseudo-code for this

Table A.2: Ablation study of VIE with 5% event rate in semi-synthetic settings.

	Average AUC (std) ↑			Average AUPRC (std)↑			Average BCE Loss (std) ↓			Average Positive Case Loss (std) ↓		
	n=5k	n=10k	n=20k	n=5k	n=10k	n=20k	n=5k	n=10k	n=20k	n=5k	n=10k	n=20k
VAE	0.594 (0.118)	0.666 (0.021)	0.693 (0.011)	0.113 (0.049)	0.144 (0.017)	0.179 (0.018)	0.308 (0.166)	0.198 (0.004)	0.198 (0.009)	0.111 (0.046)	0.147 (0.003)	0.149 (0.008)
VAE-GPD	0.583 (0.027)	0.607 (0.014)	0.663 (0.009)	0.075 (0.008)	0.087 (0.013)	0.137 (0.024)	2.106 (1.553)	0.581 (0.056)	0.195 (0.011)	**0.017** (0.015)	**0.043** (0.004)	0.143 (0.013)
IAF-GPD	0.664 (0.017)	0.554 (0.033)	0.503 (0.020)	0.113 (0.022)	0.063 (0.007)	0.057 (0.003)	0.194 (0.008)	0.208 (0.004)	0.214 (0.009)	0.144 (0.009)	0.157 (0.004)	0.160 (0.008)
Fenchel-GPD	0.666 (0.014)	0.687 (0.016)	0.681 (0.008)	**0.145** (0.011)	**0.184 (0.022)**	0.166 (0.013)	0.196 (0.008)	**0.189** (0.007)	0.190 (0.011)	0.148 (0.008)	0.141 (0.007)	0.141 (0.012)
VIE	**0.679** (0.018)	**0.693** (0.027)	**0.693** (0.015)	0.142 (0.018)	0.172 (0.032)	**0.193**(0.013)	**0.188** (0.011)	0.193 (0.005)	**0.190** (0.006)	0.139 (0.013)	0.145 (0.006)	**0.139** (0.008)
Oracle	0.694 (0.670, 0.717)			0.197 (0.179, 0.218)			0.185 (0.179, 0.198)			0.137 (0.130, 0.149)		

simulation strategy is shown in Algorithm 6, where t_0 is a pre-specified time-cut, and $H(\cdot)$ is a randomized monotone neural network to create a monotone mapping from z to the risk.

Algorithm 6: Generation of long-tailed data.

Set sample size n, latent space dimension p, number of covariates q ;
Set $\mu_p, \Sigma_p, \xi_p, \sigma_p$ (the parameters of a long-tailed distribution);
Set ν, λ (the parameters of cox-Weibull distribution);
Set time cut-bound t_0;
Initialize ψ (for MLP $g(\cdot;\psi) : \mathbb{R}^p \to \mathbb{R}^q$), θ (for AMNN $H(\cdot;\theta) : \mathbb{R}^p \to \mathbb{R}$);
for $i \in \{1, \ldots, n\}$ **do**
  Sample z_i from mixed GPD $(\mu_p, \Sigma_p, \xi_p, \sigma_p)$;
  Compute $x_i = g(z_i;\psi)$;
  Sample u_i from $\texttt{Unif}(0,1)$;
  Compute $t_i = \{\frac{-\log u_i}{\lambda \exp(H(z_i;\theta))}\}^{1/\nu}$;
  Compute $y_i = I(t_i < t_0)$;
  $d_i = (y_i, x_i)$
end
;
return $D = \{d_i; i = 1 \ldots n\}$

Table A.3 summarizes the findings with the long-tailed distributed latent space datasets, with 1% event rate. VIE can achieve relatively high AUC and AUPRC even with a small training sample size, which suggests that the proposed method can recover the long-tailed behavior in the feature representation. Among the combinations of different VI techniques, the Fenchel duality mechanism facilitates the distribution matching the best among other inference techniques.

In summary, we have tested the performance on various simulation settings (model assumptions, event rates, sample sizes, non-linearity, *etc.*) where VIE takes

Table A.3: Ablation study of VIE with 1% event rate in longtailed-synthetic settings

	Average AUC (std) ↑			Average AUPRC (std)↑			Average BCE Loss (std) ↓			Average Positive Case Loss (std) ↓		
	n=5k	n=10k	n=20k	n=5k	n=10k	n=20k	n=5k	n=10k	n=20k	n=5k	n=10k	n=20k
VAE	0.722 (0.140)	0.741 (0.099)	0.798 (0.034)	0.128 (0.076)	0.119 (0.064)	0.177 (0.034)	0.138 (0.213)	0.121 (0.200)	0.055 (0.004)	0.029 (0.010)	0.034 (0.009)	0.039 (0.005)
VAE-GPD	0.498 (0.039)	0.450 (0.021)	0.441 (0.055)	0.013 (0.002)	0.009 (0.000)	0.009 (0.002)	12.097 (5.582)	22.445 (10.247)	13.438 (7.159)	0.000 (0.000)	0.000 (0.000)	0.001 (0.002)
IAF-GPD	0.688 (0.021)	0.632 (0.037)	0.555 (0.039)	0.097 (0.019)	0.078 (0.027)	0.046 (0.014)	0.051 (0.005)	0.055 (0.003)	0.062 (0.007)	0.040 (0.005)	0.044 (0.004)	0.051 (0.007)
Fenchel-GPD	0.804 (0.028)	0.807 (0.026)	0.818 (0.020)	0.174 (0.030)	0.155 (0.052)	0.166 (0.044)	0.054 (0.007)	0.051 (0.003)	**0.047** (0.004)	0.041 (0.007)	0.040 (0.004)	0.037 (0.005)
VIE	**0.823** (0.024)	**0.810** (0.023)	**0.836** (0.026)	**0.175** (0.036)	**0.163** (0.044)	**0.202** (0.024)	**0.047** (0.005)	**0.050** (0.003)	0.049 (0.005)	**0.037** (0.004)	**0.040** (0.003)	**0.037** (0.005)
Oracle	0.829 (0.802, 0.868)			0.188 (0.153, 0.243)			0.049 (0.042, 0.055)			0.039 (0.033, 0.045)		

the lead in all cases. IAF- and GPD-only variants perform poorly, even not comparable to the vanilla VAE model. This is possibly due to the prior is not matched. Explicitly matching the prior via the Fenchel mini-max scheme improves the performance, especially in the long-tailed representation datasets. Stacked together, our full proposal of VIE consistently outperforms its variants and always approaching the oracle performance in the large sample regime.

A.4 Real-world datasets

We consider 5 real-world datasets, including 3 survival datasets in the study. Among those dataset, `COVID` and `InP` are from Duke University Health System (DUHS), which are not public at this time. `SEER`(Ries et al., 2007) and `SLEEP`(Quan et al., 1997) are two public survival datasets. Besides above clinical-based datasets, we further evaluate the model performance on `Fraud` dataset (Dal Pozzolo et al., 2017) in this supplementary material.

A.4.1 Baseline Models.

In all experiments, LDAM, FOCAL, IW, DeepSVDD and MLP are specified in terms of three-layer MLPs of 32 hidden units with ReLU activation. When tuning parameters for LASSO, based on the notation of Pedregosa et al. (2011) function `sklearn.linear_model.Lasso`, we choose α from $[10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 0.2, 0.5, 0.8]$ referred to the best performance on the validation datasets. In Focal Loss, the parameter γ is selected from the list $[0.1, 0.5, 1.0, 1.5, 2.0]$ based on the best performance on the validation datasets.

A.4.2 COVID Dataset

The dataset includes inpatient encounters to DUHS as of January 1, 2020. Vitals, administered medications, lab results, comorbidities, etc. are used as predictors to identify the risk of inpatient death, ventilation, and ICU transfer as adverse outcomes. The raw data's detailed description for each group of covariates can be found in Table A.4. The mortality rate in this dataset is 2.8%, ventilation 7.8% and ICU transfer 18%. From rare event modeling purposes, apart from the mortality prediction, we set the group of patients who experienced either death or ventilation as the combined adverse outcome group, which has 8% event rate.

Table A.4: `COVID` dataset covariates before pre-processing.

Data Name	Data Type	Number Covariates
Demographics	Numerical	1 (age)
Previous Encounters	Numerical	2
Prior Procedures w/n year	Categorical	186
Problem List w/n year	Categorical	273
Comorbidities w/n year	Categorical	545
Chief Complaint	Categorical	100
Lab Analytes Collected	Categorical	44
Lab Analytes Results	Numerical	44
Orders Placed	Categorical	32
Medications Administered	Categorical	74
Vitals Recorded	Numerical	37

In Figure A.3, we presented the comparison of VIE versus other baseline models with the two outcomes (combined and mortality). VIE shows strong performance under these metrics.

Cross-validation results To qualitatively show the superior performance of VIE, we examine the performances on a 5-fold cross validation of the `COVID` dataset for mortality prediction. VIE outperforms other baseline consistentlly on each fold. Comparing to the performance of the second-best model LDAM with a paired t test, the p-value yields $0.093 < 0.1$, with effect size 0.98, which shows the performance gap is statistically significant at $\alpha = 0.1$.

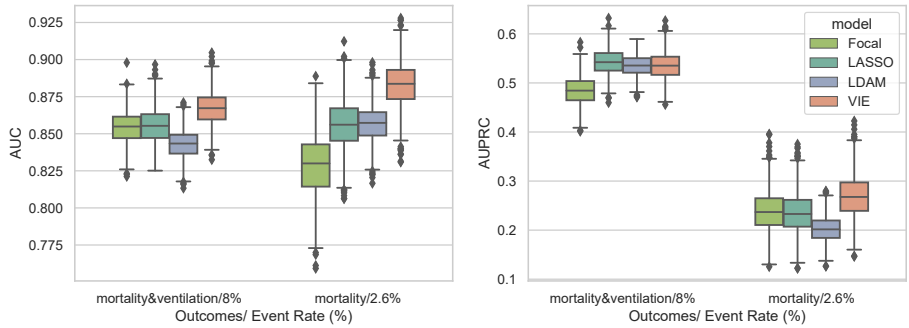

FIGURE A.3: Bootstrapped AUC (left) and AUPRC (right) Distribution of COVID dataset with different outcomes. Note that comparisons of AUPRC among event-rates groups are meaningless.

Table A.5: 5-Fold cross validation results of `COVID` dataset

	AUC					AUPRC				
k-Fold	1	2	3	4	5	1	2	3	4	5
LASSO	0.845	0.834	0.819	0.818	0.817	0.234	0.183	0.213	0.185	0.181
VAE	0.831	0.842	0.848	0.800	0.764	0.198	0.181	0.191	0.196	0.147
MLP	0.852	0.836	0.845	0.852	0.821	0.248	0.174	0.236	0.235	0.182
Focal	0.847	0.837	0.836	0.851	0.836	0.229	0.158	0.216	0.241	0.186
LDAM	0.842	0.848	0.843	0.839	0.839	0.240	0.197	0.234	0.209	0.182
VIE	**0.860**	**0.849**	**0.851**	**0.865**	**0.840**	**0.263**	**0.210**	**0.238**	**0.256**	**0.201**

A.4.3 InP Dataset

The dataset is another inpatient data of 82,450 Duke University Health System (DUHS) collected between 2014-2016. We abstracted time-varying clinical data (*i.e.*, vital signs, laboratory tests, medications) and followed patients until the Intensive Care Unit (ICU) transfer or Death. We extracted their first encounter in the system (the admission) to generate this classification study to predict the risk of the occurrence of adverse outcomes (death or ICU transfer). The descriptions of raw data can be found in Table A.6. With different sizes of the time windows, we generate four classification datasets with different event rates. As summarized in Figure A.4, VIE takes a consistent lead in both AUC and AUPRC. The advantage enlarges when the event rates drop. Note that the trend of AUPRC when event rate shrinking is not meaningful(Boyd et al., 2012).

Table A.6: `InP` dataset covariates before pre-processing

Data Name	Data Type	Number Covariates
Demographics	Numerical	3 (age, sex, race)
Admission Information	Categorical	2 (source and department)
Vitals Recorded	Numerical	10 (Diastolic, Resp, SpO2, etc.)
Lab Analytes Collected	Categorical	30
Lab Analytes Results	Numerical	30
Lab Orders Placed	Categorical	30

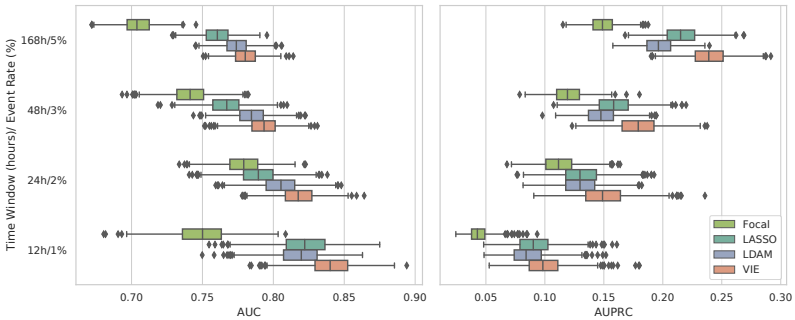

FIGURE A.4: Bootstrapped AUC (left) and AUPRC (right) Distribution of `InP` dataset with different event rate. Note that comparisons of AUPRC among event-rates groups are meaningless.

A.4.4 SEER and SLEEP Datasets

`SEER` and `SLEEP` are two public survival datasets that contain censoring (*i.e.,* an event that is not reported during the follow-up period of a subject). To create a classification dataset from a survival dataset, we deleted patients censored before the time-cut. The proportion of subjects excluded for `SEER` is less than 0.1%, for `SLEEP` dataset is less than 0.2%, which should not affect the overall credibility of the analysis. We follow the pre-processing steps provided in Chapfuwa et al. (2020).

A.4.5 Credit Card Fraud Detection

To evaluate the performances on non-clinical data, we examined the VIE model on fraud detection benchmark dataset (Dal Pozzolo et al., 2017), where fraudulent credit card transactions are coined as rare events ($\sim$ 0.2%). The dataset includes 284k records with 29 covariates. We split the original dataset into training, validation, and testing datasets with a 6:2:2 ratio to ensure fair and stable comparison.

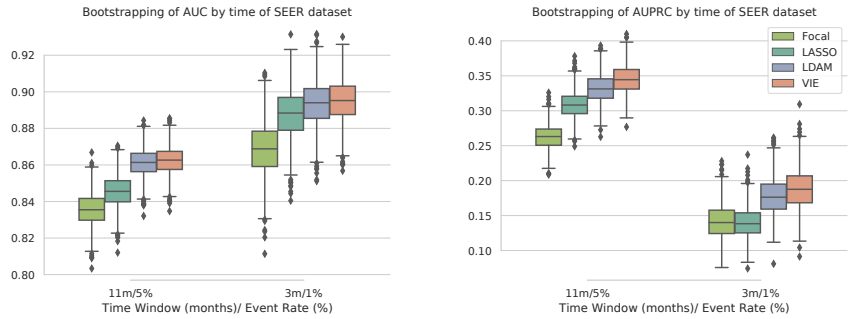

FIGURE A.5: Bootstrapped AUC (left) and AUPRC (right) Distribution of SEER dataset with different event rate.

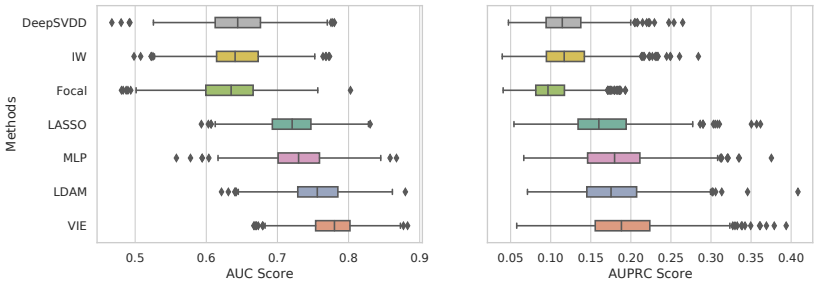

FIGURE A.6: Bootstrapped AUC (left) and AUPRC (right) Distribution of SLEEP dataset with 5% event rate.

The hyperparameters are selected based on the best performance on the validation dataset. The average and standard deviation of the metrics are presented in Table A.7. VIE outperforms other baselines and achieved an average of over 0.99 AUC in the bootstrapped samples.

Table A.7: Fraud transaction classification.

	Lasso	MLP	DeepSVDD	IW	Focal	LDAM	VIE
AUC	0.981 (0.006)	0.984 (0.007)	0.796 (0.019)	0.777 (0.026)	0.916 (0.020)	0.987 (0.005)	**0.991** (0.003)
AUPRC	0.79 (0.032)	0.80 (0.030)	0.01 (0.002)	0.57 (0.039)	0.79 (0.032)	0.78 (0.037)	**0.80** (0.032)

A.4.6 Exploration of the feature representation

We visualize the marginal relationship between latent space dimensions and risk in the real-world dataset InP dataset (1% event rate), which are shown in Figure A.9. The first dimension (top-left), the extremal behavior contributes significantly and

positively to the event risk prediction. The other three dimensions serve as inhibitors to the event risk. Empirically, all the latent dimensions have a long-tailed distribution, with learned scale parameter $\xi > 0$.

We also embed the posterior space z on a 2D plot with t-SNE, with probability contour lines, as shown in Figure A.10. The events are concentrated to one end of the latent space.

A.4.7 E. Generalized to Multiple-class classification

Our binary classification framework can be generalized to multiple-class problems easily. We will stick to the mixed-GPD distribution of the posterior z with p dimensions, and increase the number of monotone networks for each dimension of z. In the binary case, each dimension of z corresponds to one monotone network, here we can set it to k networks per-dimension. In total, we now have $p \times k$ monotone functions. Then we can apply an FC layer to the final output, with m categories, as shown in Figure A.11. In the learning object, we would replace the binary cross-entropy loss (BCE) with regular cross-entropy loss (CE) in the reconstruction term.

$$
\begin{aligned}
p_\theta(y|z) &\leftarrow \Phi(H(z;\theta)) \text{ Soft-max }, \\
H(z;\theta) &\leftarrow k \text{ Additive Monotone Neural Nets (A.4)} \\
p(z) &\leftarrow \text{Mixed GPD } (u, \xi_p, \sigma_p) \text{ (2.6)}, \\
q_\phi(z|x) &\leftarrow \text{Inverse Autoregressive Flow (2.8)}, \\
\nu(z) &\leftarrow \text{Standard neural network.}
\end{aligned}
\tag{A.5}
$$

We generate a toy dataset to illustrate VIE's performance on multiclassification problems. Based on Algorithm 5, instead of setting a binary time-cut, now we split the generated time t with a sequence of time-cuts based on the percentiles $[5\%, 15\%, 30\%, 60\%]$ of t. In this way, we have a dataset with 5 categorical outcomes with event rates $[5\%, 10\%, 15\%, 30\%, 40\%]$, respectively. To evaluate the perfor-

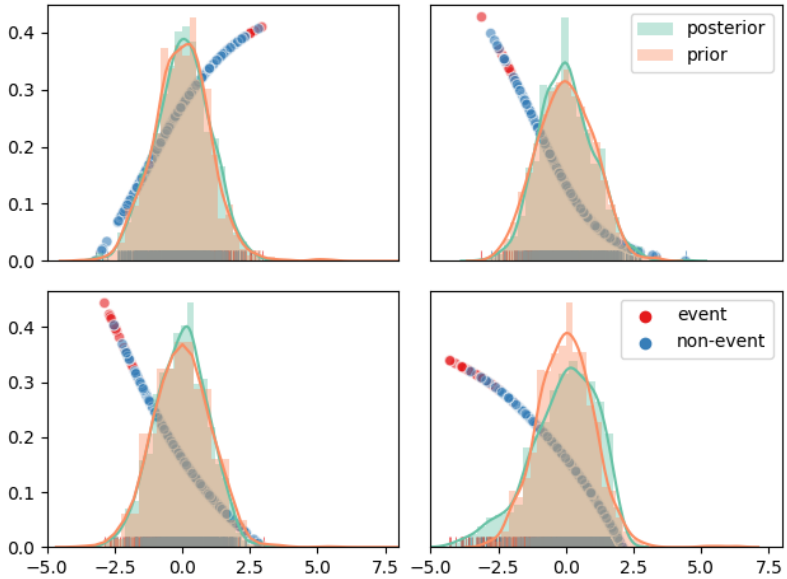

FIGURE A.7: Learned prior and posterior distribution and monotonic predicted risks

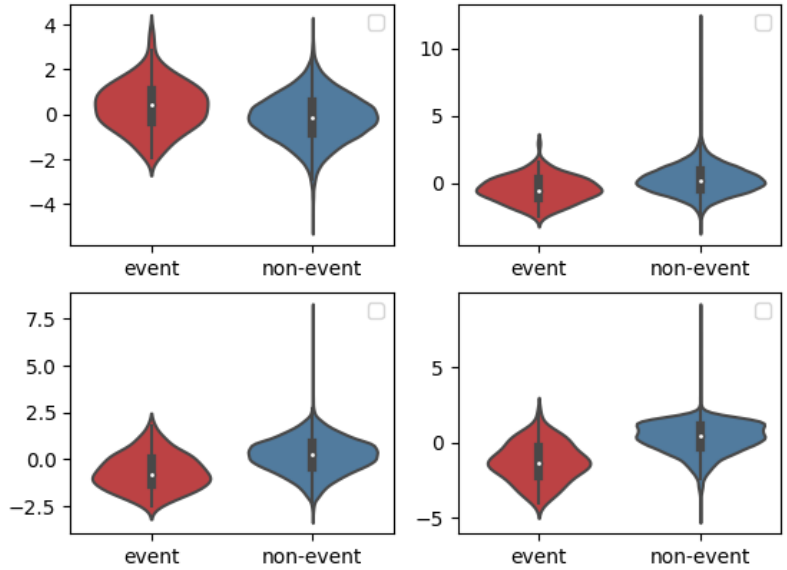

FIGURE A.8: The latent representation values distribution grouped by event type

FIGURE A.9: Four latent dimensions from the `InP` dataset (1%) event rate, where the extreme distribution in the first dimension is the simulator to the events, the other three dimensions serve as inhibitors

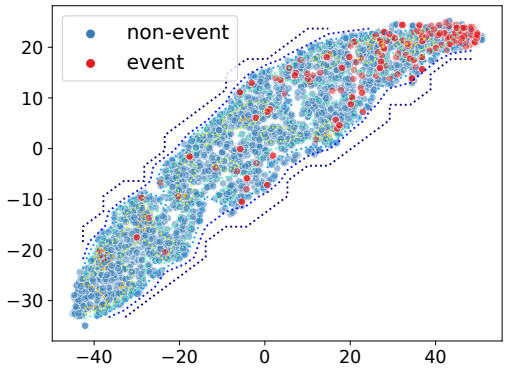

FIGURE A.10: t-SNE plots with latent representation z.

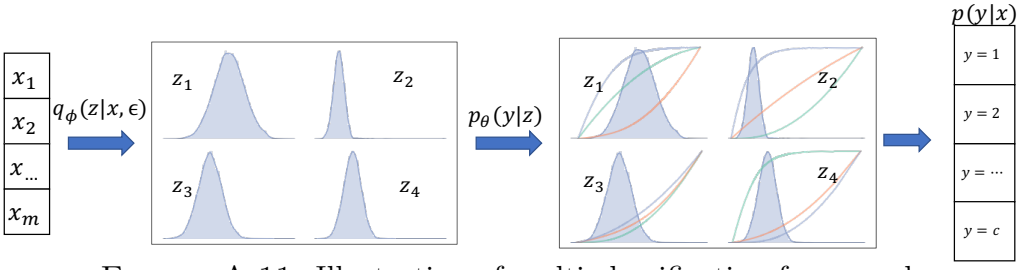

FIGURE A.11: Illustration of multi-classification framework.

mance, except for the per-class accuracy, we use $F1$ score, which is a the harmonic mean of precision (True Positives) and recall (sensitivity), $\frac{2}{\text{recall}^{-1}+\text{precision}^{-1}}$, ranges from 0 to 1, where 1 indicating better performance. We use micro-averaged F1-score (micro-F1) to calculate the overall F1 scores for all classes,

Table A.8: Performance on few-shots learning dataset

	class 1	class 2	class 3	class 4	class 5	micro-F1
event rates	5%	10%	15%	30%	40%	
Focal	**0.503**	0.202	0.095	0.139	0.089	0.1368
LDAM	0.012	**0.321**	**0.228**	0.165	0.594	0.3522
VIE	0.054	0.054	0.046	**0.372**	**0.823**	**0.4521**

Comparing to related methods: FOCAL and LDAM, the model VIE results on this toy example are comparable per class and better in terms of F1 score. FOCAL and LDAM are specified as 3-layer MLPs with 32 hidden units, and VIE uses the

previous setting, except for $k = 3$.

Appendix B

Supplementary Material to "Supercharging Imbalanced Data Learning With Energy-based Contrastive Representation Transfer"

B.1 Theoretical Support

Here we summarize some theories from literature that supports the development of this paper by making it self-contained. Attempts have been made to unify the notations, making them consistent with our paper, and also drop some contents from the original presentations that not directly relevant in this context.

B.1.1 Nonlinear ICA with auxiliary variables

The following theory lists the technical conditions required for the identification of conditional nonlinear ICA model, based on which our work was built.

Definition B.1.1 (Conditionally exponential of order k)**.** A random variable (independent component) $[\boldsymbol{s}]_i$ is conditionally exponential of order k given random vector

$\boldsymbol{c}$ if its conditional pdf can be given in the form

$$p([\boldsymbol{s}]_i|\boldsymbol{c}) = \frac{Q_i([\boldsymbol{s}]_i)}{Z_i(\boldsymbol{c})} \exp\left[\sum_{j=1}^{k} \tilde{q}_{ij}([\boldsymbol{s}]_i)\lambda_{ij}(\boldsymbol{c})\right] \tag{B.1}$$

almost everywhere in the support of $\boldsymbol{c}$, with $\tilde{q}_{ij}, \lambda_{ij}, Q_i$ and Z_i scalar-valued functions. The sufficient statistics $\tilde{q}_{ij}$ are assumed linearly independent.

Theorem B.1.2 (Theorem 3, Hyvarinen et al. (2019), identification of Nonlinear ICA)**.** *Assume (i) the data follows the nonlinear ICA model with the conditional independence $q(\boldsymbol{s}|\boldsymbol{c}) = \prod_j q_j([\boldsymbol{s}]_j|\boldsymbol{c})$; (ii) Each $[\boldsymbol{s}]_j$ is conditionally exponential given $\boldsymbol{c}$ (Definition B.1.1); (iii) There exist $nk+1$ points $\boldsymbol{c}_0, \cdots, \boldsymbol{c}_{nk}$, such that the following matrix of size $nk \times nk$*

$$\tilde{\boldsymbol{L}} = \begin{pmatrix} \lambda_{11}(\boldsymbol{c}_1) - \lambda_{11}(\boldsymbol{c}_0) & \cdots & \lambda_{11}(\boldsymbol{c}_{nk}) - \lambda_{11}(\boldsymbol{c}_0) \\ \lambda_{nk}(\boldsymbol{c}_1) - \lambda_{nk}(\boldsymbol{c}_0) & \cdots & \lambda_{nk}(\boldsymbol{c}_{nk}) - \lambda_{nk}(\boldsymbol{c}_0) \end{pmatrix} \tag{B.2}$$

is invertible; (iv) nonlinear Logistic regression system Eqn (1) is trained using functions with universal approximation capacity. Then in the limit of infinite data, $f(\boldsymbol{z})$ provides a consistent estimator of the nonlinear ICA model, up to a linear transformation of point-wise scalar functions of the independent components.

B.1.2 Variance and generalization bound

The following theories explore the consequence of training using only nonparametric causal augmentation. First we define the risk estimators.

Definition B.1.3. Let $\tilde{S}$ be the non-parametric source augmentation defined in Eqn (3) main text, $\ell(\cdot)$ be the loss function, $g(\boldsymbol{z})$ be the hypothesis function. We define the risk R and causally augmented risk estimator $\breve{R}$ wrt g respectively as

$$R(g) \triangleq \mathbb{E}_Z[\ell(g(Z))], \breve{R}(g) \triangleq \mathbb{E}_{\tilde{S}}[\ell(g(\hat{f}^{-1}(\tilde{S})))], \tag{B.3}$$

where $\hat{f}$ is the estimated causal de-mixing function.

The following theorem revealed that assuming perfect knowledge of de-mixing function f, the causally augmented risk estimator is optimal.

Proposition B.1.4. Assuming $\hat{f} = f$, and let $\ell(h(\boldsymbol{x}), y)$ be the classification loss for predictor $h \in \mathcal{H}$. Let $\hat{R}(h) = \sum_m w_m \hat{R}_m(h)$ be an estimator for $R(h) = \mathbb{E}[\ell(h(\boldsymbol{x}), y)]$, such that $\hat{R}(h)$ is an unbiased estimator for $R_m(h) = \mathbb{E}_{y=m}[\ell(h(\boldsymbol{x}), y)]$. Then $\tilde{R}(h) = \sum_m w_m \tilde{R}_m(h)$, where $\tilde{R}_m(h) \triangleq \mathbb{E}_{y_i=m}[\ell(h(\tilde{\boldsymbol{x}}_i), y_i)]$ is the minimum variance unbiased estimator among all $\hat{R}(h)$.

Proof. This is a direct consequence of Theorem B.1.5. If $\tilde{R}(h)$ is not the minimal variance estimator, then at least one of $\tilde{R}_m(h)$ is not optimal, which contradicts Theorem B.1.5. □

Theorem B.1.5 (Theorem 1, Teshima et al. (2020), minimum variance property)**.** *Assuming $\hat{f} = f$. Then for each $g \in \mathcal{G}$, the causal augmented risk estimator $\tilde{R}(g)$ is the uniformly minimum variance unbiased estimator of $R(g)$, i.e., $\mathbb{E}[\breve{R}(g)] = R(g)$ and for any unbiased estimator $\breve{R}$ of $R(g)$ (i.e., $\mathbb{E}[\breve{R}(g)] = R(g)$),*

$$Var[\tilde{R}(g)] \leqslant Var[\breve{R}(g)]. \tag{B.4}$$

Since we are bound to have estimation errors, the next theorem establishes the generalization bounds wrt such errors.

Theorem B.1.6 (Theorem 2, Teshima et al. (2020), excess risk bound)**.** *Let $\breve{g} = \arg\min \breve{R}$ and $g^* = \arg\min R(g)$, then under appropriate assumptions (Assumptionss 1-8 in Teshima et al. (2020)), for arbitrary $\delta, \delta' \in (0, 1)$, we have probability at least $1 - (\delta + \delta')$,*

$$R(\breve{g}) - R(g^*) \leqslant \underbrace{C \sum_{j=1}^{d} \|f_j - \hat{f}\|_{W^{1,1}}}_{\textit{Approximation error}} + \underbrace{4d\mathfrak{R}(\mathcal{G}) + 2dB_\ell \sqrt{\frac{\log 2/\delta}{2n}}}_{\textit{Estimation error}} + \underbrace{\kappa_1(\delta', n) + dB_\ell B_q \kappa_2(f - \hat{f})}_{\textit{Higher-order terms}}.$$

(B.5)

Here $\|\cdot\|_{W^{1,1}}$ *is the Sobolev norm and* $\mathfrak{R}(\mathcal{G})$ *is the effective Rademacher complexity defined by*

$$\mathfrak{R}(\mathcal{G}) \triangleq \frac{1}{n}\mathbb{E}_{\hat{S}}\mathbb{E}_{\sigma}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n}\sigma_i s\mathbb{E}_{S'_2,\cdots,S'_d}\tilde{\ell}(\hat{s}_i, S'_2,\cdots,S'_d)\right|\right], \tag{B.6}$$

where $\{\sigma_i\}_{i=1}^n$ *are independent sign variables,* $\mathbb{E}_{\hat{S}}$ *is the expectation wrt* $\{\hat{s}_i\}_{i=1}^n$*, the dummy variables* $S'_2,\cdots,S'_d$ *are i.i.d. copies of* $\hat{s}_1$*, and* $\tilde{\ell}$ *is defined by*

$$\tilde{\ell}(s_1,\cdots,s_d) \triangleq \frac{1}{d!}\sum_{\pi}\ell(g,\hat{f}^{-1}(s_{\pi(1)},\cdots,s_{\pi(d)})), \tag{B.7}$$

where π *denotes the permutations.* κ_1,κ_2 *are higher order terms,* B_q, B_ℓ *respectively depends on density* q *and loss* ℓ*, while* C' *depends on* (f,q,ℓ,d)*.*

B.1.3 Speedup from shared embedding

Robinson et al. (2020) built some interesting theories trying to answer the following question: "*Can large amounts of weakly labeled data provably help learn a better model than strong labels alone?*" The answer is positive, assuming there is a shared embedding between the *weak* and *strong* tasks, which respectively refers auxiliary (secondary) and main tasks of interests. We summarize its main findings below and elaborate how it lends support for ECRT.

In the setting of weakly supervised learning, we have the triplet $(\mathcal{X},\mathcal{W},\mathcal{Y})$, where $\mathcal{X}$ and $\mathcal{Y}$ respectively denote the features and labels of interest (strong task), and $\mathcal{W}$ denote weak task labels that are relevant to the prediction of $\mathcal{Y}$. It is assumed that there is this unknown good embedding $Z = f_0(X)$ that predicts W, that could

be leveraged to derive a model of the form $\hat{g}(\cdot, \hat{f}) : \mathcal{X} \to \mathcal{Y}$ that improves learning.

Algorithm 7: Weakly supervised learning.

1. Pretrain encoder with weak labels

$$\hat{f} \leftarrow \mathrm{Alg}(\mathcal{F}, \mathbb{P}_{XW})$$

2. Augment data with $z_i = \hat{f}(x_i) \Rightarrow \{(x_i, y_i, z_i)\} \sim \hat{\mathbb{P}}_{XYZ}$

3. Optimize the strong task $\hat{g} \leftarrow \mathrm{Alg}(\mathcal{G}, \hat{\mathbb{P}}_{XYZ})$

Theorem B.1.7 (Theorem 3, Robinson et al. (2020))**.** *Suppose that* $Rate_m(\mathcal{F}, \mathbb{P}_X W) = \mathcal{O}(m^{-\alpha})$ *and that* $Alg_n(\mathcal{G}, \hat{\mathbb{P}})$ *is ERM. Under suitable assumptions on* $(\ell, \mathbb{P}, \mathcal{F})$*, Algorithm 7 obtains excess risk*

$$\mathcal{O}\left(\frac{\alpha\beta \log n + \log(1/\delta)}{n} + \frac{1}{n^{\alpha\beta}}\right) \tag{B.8}$$

with probability $1 - \delta$*, when* $m = \Omega(n^{\beta})$ *for* $\mathcal{W}$ *discrete, or* $m = \Omega(n^{2\beta})$ *for* $\mathcal{W}$ *continuous.*

For concrete examples, in a typical learning scenario where $\mathrm{Alg}_m(\mathcal{F}, \mathbb{P}_{XW}) = \mathcal{O}(m^{-1/2})$, one obtains the fast rates $\mathcal{O}(1/n)$ for $m = \Omega(n^2)$.

In the context of our ECRT, we identify the learning of common causal de-mixing function $f(\boldsymbol{z})$ as the weak learning task, and the source space $\mathcal{S}$ is the common embedding space of interest. This allows us to tape into the power of weakly supervised learning to improve the main classification task. See Figure 9 in the main text for evidence.

B.1.4 Energy-based GCL

Mutual information (MI) is a popular metric to quantify the associations between random variables, and has been applied to a lot of areas like independent component

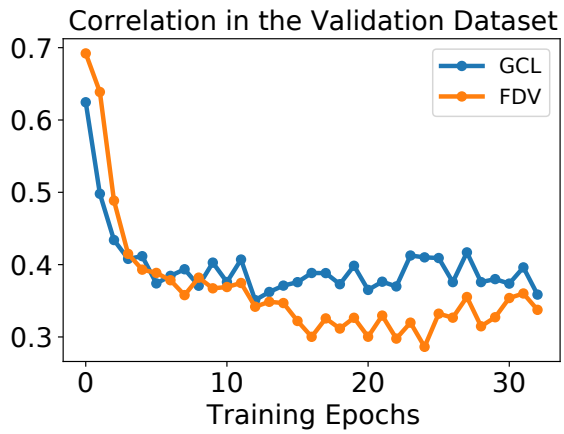

FIGURE B.1: Comparison of correlation decreasing with GCL and FDV in the MNIST dataset. FDV approaches the optimality faster, which indicates better efficiency.

analysis, fair learning and etc. Inspired by the FDV loss introduced in Guo et al. (2021), rooted from the MI between the feature z and label y. Following the Equation (4) in the main text for I_{FDV}, we have a novel MI objective, pointwise mutual information (PMI) which takes the place of logistic regression objective in GCL,

$$\mathcal{L}_{\text{FDV}}(f_\psi, g_\nu) \triangleq \hat{I}_{\text{DV}}(\{\boldsymbol{z}_i, \boldsymbol{y}_i\}) + \frac{\sum_j \exp[(g_\nu(\boldsymbol{y}_j, \boldsymbol{z}_i) - g_\nu(\boldsymbol{y}_i, \boldsymbol{z}_i))/\tau]}{\sum_j \exp[(\hat{g}_\nu(\boldsymbol{y}_j, \boldsymbol{z}_i) - \hat{g}_\nu(\boldsymbol{y}_i, \boldsymbol{z}_i))/\tau]} - 1, \quad \text{(B.9)}$$

where $\boldsymbol{y}_i$ is the embedding of label, τ is a learnable temperature parameter, $g_\nu(\boldsymbol{y}, \boldsymbol{z}) = \text{sim}(\boldsymbol{y}, \sum_{a=1}^{d} \gamma_\nu^a(y, [\boldsymbol{s}]_a))$, and $\text{sim}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^\top \cdot \boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}$, γ_ν and $\boldsymbol{s}$ are the same as Equation (1) with slightly change of dimension with linear transformations.

One notably difference between Equation (B.9) and Equation (1) from the GCL lies in the negative sample size. In energy-based GCL, the model can treat all the examples within a minibatch as negative examples, while the original GCL contrasts with only one negative example in the minibatch by permutating labels. This refinement can benefit the learning efficiency of the NICA step, the correlation between labels decrease faster than the original GCL, as illustrated in Figure B.1.

B.2 Regression for continuous labels

We can further extend the applicability of the proposed ECRT to the case of regressing continuous outcomes. While in principle, the procedures described in Sec 3 can be readily applied, we advocate coarse graining wrt label y similar to what has been practiced in *sliced inverse regression* Li (1991), especially when the feature dimension is high relative to the sample size. Specifically, we partition y into different bins, and use feed the bin label as the conditioning variable in the GCL step. We still use the regression loss for the training of encoder and predictors.

B.3 Implementation of Augmentation

Let $\hat{\boldsymbol{s}}_i^k, i = 1, \cdots, n_k$ be the estimated source representation for the k-th class.

- Non-parametric augmentation: shuffling indices as described in the main text.
- Parametric augmentation: estimate $\hat{\mu}_k = \text{mean}(\hat{\boldsymbol{s}}^k), \hat{\sigma}_k = \text{std}(\hat{\boldsymbol{s}}^k)$, then $\boldsymbol{s}^{k,aug} \sim \mathcal{N}(\hat{\mu}_k, \hat{\sigma}_k^2)$.
- Oracle augmentation: nonparametric augmentation with an abundance of class-conditional source space samples .

In Figure 9 from main text we compare the efficiency of parametric and nonparametric augmentation schemes under different minority sample size. In particular, we compute the MMD distance $\|\hat{\mu}_{aug} - \hat{\mu}_{ref}\|_\kappa$, where $\hat{\mu}_{aug} = \sum_i \kappa(\tilde{s}_i, \cdot)$ and $\hat{\mu}_{ref} = \sum_i \kappa(s_i, \cdot)$. Here $\kappa(\cdot, \cdot)$ is the Gaussian rbf kernel $\kappa(x, y) = \exp(\|x - y\|^2/2\sigma^2)$, $\|f\|_\kappa = \sqrt{\langle f, f \rangle_\kappa}$ is the RKHS norm and $\tilde{s}_i, s_i$ respectively denote augmented class-conditional samples (from few minority samples) and empirical distribution of class-conditional samples (where we use all samples from the same class that we holdout). In this example we use $2k$ samples for the ground truth and augment minority to

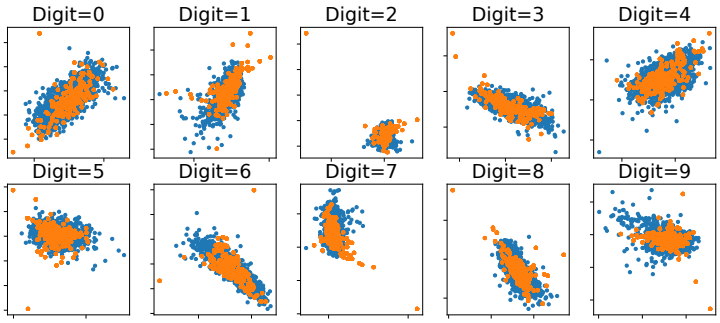

FIGURE B.2: Feature space augmentation for MNIST.

the same size. Different kernel bandwidth σ of κ yields qualitatively similar results, and in the paper we report the one with $\sigma = 0.5$.

In Figure B.2, we visualize the augmentation in feature space for the MNIST dataset. And we see for boundary points the discrepancy can be amplified by the neural network inversion, which partly explained the sub-optimal performance from feature space augmentation. In contrast, the source space augmentation advocated in this paper is more computationally efficient and robust.

Note that while in principle the majority classes can be similarly augmented, we choose not to refine our model with the augmented majorities. This decision is justified by the classical consideration for bias-variance trade-off: estimation errors of $f(\boldsymbol{z})$ is inevitable (*e.g.*, finite sample size, SGD, limited network capacity, *etc.*), and they will carry over to the augmented samples, resulting biases in the augmented estimation of our predictor. On the other hand, using augmented samples helps bring down estimation variance. For minority labels, the reduction in variance is greater than the induced bias, and consequently merits the application of ECRT to improve performance. For majority labels, this might not be the case.

B.4 Toy Model Experiment

B.4.1 Toy data demo

We sample seven groups of two-dimensional uncorrelated Gaussian of each with size 2000, with different means and variances as our real source representation $\boldsymbol{s}$. Specifically, $s_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), i = 0, \cdots, 6$, where $\boldsymbol{\mu}_0 = [-0.5, -1], \boldsymbol{\mu}_1 = [2, 1], \boldsymbol{\mu}_2 = [5, 2], \boldsymbol{\mu}_3 = [1, 3], \boldsymbol{\mu}_4 = [-2, 1], \boldsymbol{\mu}_5 = [-3.5, 4], \boldsymbol{\mu}_6 = [-4, -1], \boldsymbol{\Sigma}_0 = [0.5, 0.5], \boldsymbol{\Sigma}_1 = [3, 1], \boldsymbol{\Sigma}_2 = [1, 2], \boldsymbol{\Sigma}_3 = [0.3, 2], \boldsymbol{\Sigma}_4 = [1, 0.2], \boldsymbol{\Sigma}_5 = [1, 1], \boldsymbol{\Sigma}_6 = [2, 0.3]$. Then we perform classical Hénon transformation $z_{(1)} = 1 - 1.4\,\tilde{s}_{(1)}^2 + \tilde{s}_{(2)}, z_{(2)} = 0.3, \tilde{s}_{(1)}$ to generate the data in feature space.

B.4.2 Extreme-classification Toy Data

We sample 1000 groups of two-dimensional uncorrelated Gaussian with mean ranges uniformly sampled from range $(-4, 4)$ and standard deviation fixed to 0.1. Validation dataset is fixed with 20 samples per-class, and the sampler per class in training dataset varies with $5, 10, 20, 50, 75, 100, 150$. The summary for extreme-classification is presented in Table B.1.

Note that when the total number of categories is 10 with 20 samples per class, ERM has top-1 accuracy as high as 0.914, and the performance drops when the number of categories increasing. With 500 categories, the accuracy decreases to 0.005, and in the scenario where we presented in the main text with 1000 categories, ERM performs no better than random guessing.

Table B.1: Validation results for extreme classification.

Metric	NLL			Top 1			Top5		
Sample-size	ERM	ECRT W/O Aug	ECRT	ERM	ECRT W/O Aug	ECRT	ERM	ECRT W/O Aug	ECRT
5	6.9246	4.3884	**3.7098**	0.0015	0.0450	**0.0760**	0.0053	0.2275	**0.3220**
10	6.9242	4.0536	**3.4752**	0.0018	0.0638	**0.0856**	0.0060	0.2666	**0.3549**
20	6.9185	3.5405	**3.3657**	0.0016	0.0741	**0.0877**	0.0070	0.3126	**0.3675**
50	6.9240	3.3794	**3.2957**	0.0010	0.0834	**0.0934**	0.0045	0.3493	**0.3835**
75	6.9190	3.3499	**3.2594**	0.0020	0.0846	**0.0939**	0.0060	0.3571	**0.3950**
100	6.9182	3.3394	**3.2675**	0.0020	0.0834	**0.0934**	0.0053	0.3580	**0.3918**
150	6.9174	3.2731	**3.2597**	0.0011	0.0906	**0.0984**	0.0059	0.3777	**0.4034**

B.5 Real-world Data Experiments

B.5.1 Image Data Benchmarks

We summarized the image datasets in Table B.2 and the network architectures used for respective datasets in Tables B.3, B.4, B.5 and B.6. The hyperparameters we used in these experiments are presented in B.10. The results reported here are from our regularized non-parametric ECRT implementation, parametric ECRT implementation show a similar trend, with slightly improved performance (results now shown). We use 2 latent dims for MNIST and 32 latent dims for CIFAR100, iNaturalist and Tiny-Imagenet.

Preprocessing For iNaturalist dataset, we used pretrained Inception V3 to extract the features with dim=2048. For Tiny-imagenet we finetuned the Resnet 18 to extract the features with dim = 512.

Baselines We used the ERM, LDAM [1], Focal [2], IW [3], and VAT [4] baseline implementations. For GAN we adopted the CGAN model with architecture shown in B.9 and noise dimension listed in B.11 for each dataset. We compared all the models with their own best performance after early stopping.

Discrepancy of baseline performance. We noticed that our implementation of baseline models, especially for the ERM baseline, yields results better than what's reported in literature (LDAM in particular). Specifically, our results look better. After carefully compared our implementation to the LDAM codebase, we see that the discrepancy comes from the choice of optimizer. The use of vanilla SGD optimizer, as practiced in LDAM, results in degraded performance of baseline, and consequently a larger performance gap compared to strong solutions.

[1] `https://github.com/kaidic/LDAM-DRW`

[2] `https://github.com/artemmavrin/focal-loss`

[3] `https://github.com/idiap/importance-sampling`

[4] `https://github.com/lyakaap/VAT-pytorch`

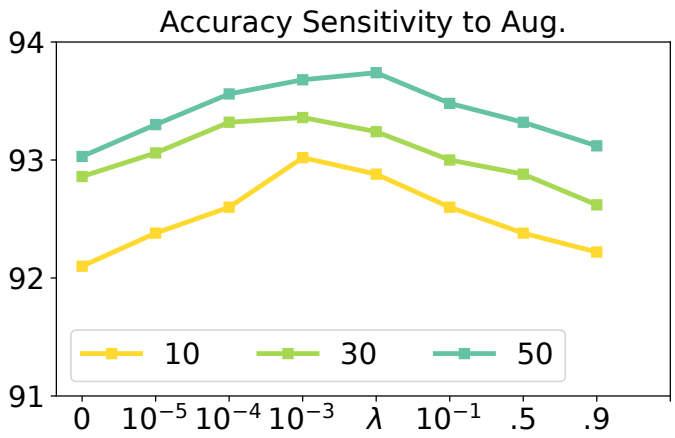

FIGURE B.3: Sensitivity analysis (overall top 1 accuracy) of augmentation strength λ. Complementing Figure 11 in main text.

Majority bias. In Figure B.4, we give the top-1 accuracy wrt different minority size on the iNaturalist dataset. This figure is complementary to the F1-label frequency plot given in Figure 14 from the main text. While the improvement at the tail part are strong under both metrics, we see clear evidence of majority bias in the Top-1 accuracy plot. ERM and LDAM show better performance in accuracy for the sample abundant majority regime, but severe performance drop in the sample deficient minority regime. This is because ERM and alike finds it more rewarding to favor the majorities during inference, which gives better sensitivity but much worse specificity for the majority samples, and consistently hurting the performance for minorities.

B.5.2 Language Data Benchmark

Dataset and preprocessing. In this experiment, have used the `arXiv` dataset hosted on `Kaggle` [5]. We use the pretrained BERT model from the `transformers` package [6] to extract sentence features. Specifically, we applied the `SciBERT` model (`allenai/scibert_scivocab_uncased`) Beltagy et al. (2019), and used the BERT

[5] `https://www.kaggle.com/Cornell-University/arxiv`

[6] `https://github.com/huggingface/transformers`

default 768-dimensional sentence embedding for each abstract. The training set includes 160k data, where class labels with more than 5k samples are identified as majority classes, with the rest assigned to minority label classes. All label classes with less than 20 samples have been excluded from our analysis. This gives us a total of 14 majority classes and 138 minority classes.

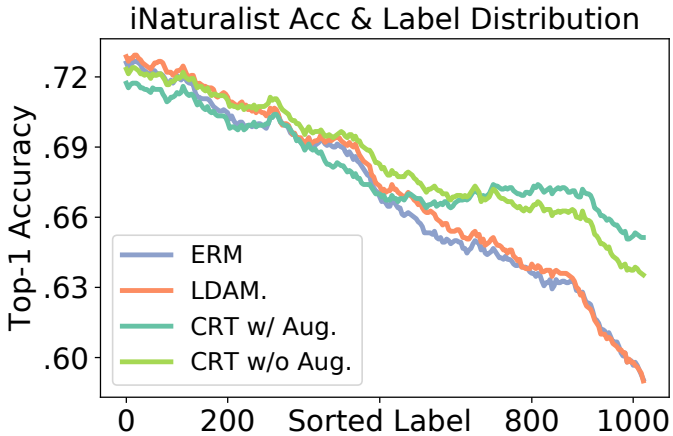

FIGURE B.4: Class-conditional Top-1 accuracy curve for iNat2019. Complementing Figure 1 (F1 score). Note that ERM and LDAM show better accuracy for sample-rich majorities, but worse F1 scores. This evidences majority bias, that a predictor has a low specificity for data rich classes.

Setup. Different from the image benchmarks, the `arXiv` data is a multi-label prediction task. Each abstract is associated with at least one, possibly multiple labels, and we make binary classifications for each label class. In the training of GCL model, we allow samples with multiple labels to be reused by different classes, as each constructs a valid source IC distribution under our hypothesis. Only standard ECRT is considered in this experiment. We set the source space dimension to 64 and use the network architecture described in Table B.7.

Evaluation. The accuracy reported for this experiment is defined as follows: say a sample is associated with k-labels, then we compared the top-k predicted labels to the true labels, and report the averaged accuracy for this sample. Like previous experiments, we target a balanced evaluation set. However, getting a perfectly

balanced evaluation set is impossible here, as samples are associated with multiple labels. We extracted a nearly-balanced evaluation set including 847 samples, where each label has 10 to 50 counts. Most of classes have 10-15 samples in our nearly-balanced evaluation set.

Table B.2: Summary statistics of real-world datasets

Name	Dim	Train (Majority)	Train (Minority)	Validation
MNIST	(28×28)	6000 $\times$(1 or 5) (cls)	1200 $\times$(1 or 5) (cls)	1000×10 (cls)
Cifar	$(32 \times 32 \times 3)$	500 $\times$50 (cls)	500 $\times$50 (cls)	100×100 (cls)
INat	(None $\times$ None$\times$3)	($\geqslant 120$) $\times$725 (cls)	(< 120) $\times$285 (cls)	3×1010 (cls)
Tiny	$(64 \times 64 \times 3)$	450 $\times$100 (cls)	45 $\times$100 (cls)	50×200 (cls)
Arxiv	(*None*)	(> 5000) $\times$14 (cls)	(< 5000) $\times$138 (cls)	12×152 (cls)

Table B.3: MNIST experiment network architecture.

Network	Architecture
Encoder	fc(unit=32)+ReLU + fc(unit=32)+ReLU + fc(unit=2)
Decoder	fc(unit=32)+ReLU + fc(unit=32)+ReLU + fc(unit=10)
FLOW	MAF($n_{blocks} = 4$, $hidden_{size} = 128$, $n_{hidden} = 2$).

Table B.4: Cifar100 experiment network architecture.

Network	Architecture
Encoder	Resnet18 [a] + fc(unit=32)
Decoder	fc(unit=256)+ReLU fc(unit=100)
FLOW	MAF($n_{blocks} = 4$, $hidden_{size} = 128$, $n_{hidden} = 2$).

[a] Resnet 18 without last layer

Table B.5: iNaturalist experiment network architecture.

Network	Architecture
Pretrain	inception(v3)
Encoder	fc(unit=1024)+ReLU +Dropout(0.1) + fc(unit=512)+ReLU +Dropout(0.1) + fc(unit=32)
Decoder	fc(unit=32)+ReLU +Dropout(0.1) + fc(unit=512)+ReLU +Dropout(0.1) + fc(unit=1010)
FLOW	MAF($n_{blocks} = 4$, $hidden_{size} = 128$, $n_{hidden} = 2$).

Table B.6: Tiny ImageNet experiment network architecture.

Network	Architecture
Pretrain	ResNet18
Encoder	fc(unit=1024)+ReLU +Dropout(0.1) + fc(unit=512)+ReLU +Dropout(0.1) + fc(unit=32)
Decoder	fc(unit=512)+ReLU +Dropout(0.1) + fc(unit=200)
FLOW	MAF($n_{blocks} = 4$, $hidden_{size} = 128$, $n_{hidden} = 2$).

Table B.7: Arxiv experiment network architecture.

Network	Architecture
Pretrain	BERT
Encoder	fc(unit=1024)+ReLU +Dropout(0.1) + fc(unit=512)+ReLU +Dropout(0.1) + fc(unit=64)
Decoder	fc(unit=64)+ReLU +Dropout(0.1) + fc(unit=512)+ReLU +Dropout(0.1) + fc(unit=152)
FLOW	MAF($n_{blocks} = 4$, $hidden_{size} = 128$, $n_{hidden} = 2$).

Table B.8: MNIST results with different numbers of minority categories.

# minority label	1		5	
	NLL	Top 1	NLL	Top 1
ERM	0.342	0.933	0.390	0.904
LDAM	1.6737	0.9609	1.50	0.940
ECRT	**0.186**	**0.972**	**0.257**	**0.950**

Table B.9: GAN network architecture.

Network	Architecture
Generator	FC(unit=256) +LeakyRelu + FC(unit=256)+Leakyrelu +FC(unit=latent$_{dim}$)
Discriminator	FC(unit=256) +LeakyRelu +dropout(0.1) + FC(unit=256)+LeakyRelu +dropout(0.1) +FC(unit=1)+sigmoid

Table B.10: Hyperparameter of datasets (ECRT).

Name	Reg Weight	Aug Strength
MNIST	1e-2	1e-3
Cifar	1e-2	1e-3
INat	5e-3	1e-3
Tiny	1e-3	1e-3
Arxiv	1e-2	1e-2

Table B.11: Hyperparameter of datasets (GAN).

Name	noise dim
MNIST	32
Cifar	64
INat	128
Tiny	128
arxiv	64

Appendix C

Supplementary Material to "Personalized Time-to-event Distribution with Deep Learning Solutions"

C.1 Implementing VSI

For model inputs, we encode each subject's time with a one-hot vector t^{OH} with dimension $(1 \times M)$ based on the observed event t_i for subject ($\delta_i = 1$). With event t has been one-hot-encoded, only one bin where the event time falls in equals 1, else equal 0. For censoring observations ($\delta_i = 0$) we have partial information about the true unknown event time, instead of missingness. We know for sure that the events would happen after the observed censored time t_i. The input vector, timepoints after the censored time t_i have been put a prior based on population time-to-event distribution by Nelson-Aalen Estimator. This re-weighting strategy informs the model that this subject didn't have event before time t_i, and also regularizes the model with respect to empirical tail distribution. One may also choose not to encode t_i but adding δ_i to the input as well. In this case, each subject in the training has input of the form (x_i, t_i, δ_i). To predict new t, the input for the prediction process is $(x_i, \delta = 1)$ where

by defaut we are predicting the time-to-event. We have shown by experiments the two form of inputs produce similar results and the encoding strategy performs better under both simulation and real-world scenarios. The KS statistic for 30% event rates has 0.073, well VSI has 0.059.

Population Survival Estimation Based on Nelson-Aalen Estimator For NA estimator, $\tilde{H}(t) = \sum_{t_j \leqslant t} \frac{d_j}{n_j}$. Therefore we could calculate $f(t)$ accordingly. Let the selected discretized time be $t_{(b)}$, where $b \in \{1, \ldots, M\}$.

$$
\begin{aligned}
\tilde{H}(t_{(b)}) &= \sum_{j \leqslant b} \frac{d_j}{n_j} \\
\tilde{S}(t_{(b)}) &= \exp(-\tilde{H}(t_{(b)})) \\
\tilde{f}(t_{(1)}) &= 1 - \tilde{S}(t_{(1)}) \\
\tilde{f}(t_{(b)}) &= \tilde{S}(t_{(b)}) - \tilde{S}(t_{(b-1)}) \\
&= \exp(-\sum_{j \leqslant b} \frac{d_j}{n_j}) - \exp(-\sum_{j \leqslant b-1} \frac{d_j}{n_j}) \\
\tilde{f}(t_{(M)}) &= 1 - \sum_{b=1}^{M-1} \tilde{f}(t_{(b)})
\end{aligned}
$$

Therefore the we have $[\tilde{f}(t_{(1)}), \tilde{f}(t_{(2)}), \ldots, \tilde{f}(t_{(M)})]$ accordingly for each timepoint. For each censored subject where the censored time falls in bin k,

$$
t_{(b)}^{\mathrm{OH}} = \begin{cases} 0, & \text{if } b \leqslant k \\ \frac{\tilde{f}(t_{(b)})}{\sum_{b>k} \tilde{f}(t_{(b)})}, & \text{if } b > k \end{cases}
$$

In this way the constructed $\sum_b t_{(b)}^{\mathrm{OH}} = 1$ as well. Now we have the soft one-hot-encoding for original censored t.

Architecture for VSI The input of VSI is $Y_i = (x_i, \delta_i, t_i^{\mathrm{OH}})$. Parameters for $p_\theta(z|x)$ and $q_\phi(z|t_i^{\mathrm{OH}}, x)$ were obtained by two MLP framework respectively, each

with hidden dimensions $[32, 32]$ with output dimension $[32]$. The decoding arm $p_\theta(t_i|z)$ has hidden dimensions $[32, 32, 32]$ with output a $(1 \times M)$ vector which is the predicted logits. Activation function is leaky ReLU, continuous variables are z-transformed before entering the model based on training datasets.

C.2 Derivation of Likelihood lowerbound for censored observations

To get the likelihood lowerbound for censored observations, we applied Fubini's theorem and Jensen's inequality.

$$
\begin{aligned}
&\mathcal{L}_c(x_i, t_i; \theta) \\
&= \log\ S_\theta(t_i|x_i) \\
&= \log \int_{t_i}^{\infty} p_\theta(t|x_i)dt \\
&= \log \int_{t_i}^{\infty} \int_z p_\theta(t, z|x_i)dzdt \\
&= \log \int_{t_i}^{\infty} \int_z \frac{p_\theta(z, t|x_i)}{q_\phi(z|t_i, x_i)} q_\phi(z|t_i, x_i)dzdt \\
&= \log \int_{t_i}^{\infty} \int_{q_\phi(z|t_i,x_i)} \frac{p_\theta(z, t|x_i)}{q_\phi(z|t_i, x_i)} dQ(z|t_i, x_i)dt \\
&= \log \int_{q_\phi(z|t_i,x_i)} \int_{t_i}^{\infty} \frac{p_\theta(z, t|x_i)}{q_\phi(z|t_i, x_i)} dtdQ(z|t_i, x_i) \\
&= \log \int_{q_\phi(z|t_i,x_i)} \int_{t_i}^{\infty} p_\theta(t|z) \frac{p_\theta(z|x_i)}{q_\phi(z|t_i, x_i)} dtdQ(z|t_i, x_i) \\
&= \log \int_{q_\phi(z|t_i,x_i)} \frac{p_\theta(z|x_i)}{q_\phi(z|t_i, x_i)} \int_{t_i}^{\infty} p_\theta(t|z)dtdQ(z|t_i, x_i) \\
&\geqslant \int_{q_\phi(z|t_i,x_i)} \log \left(\frac{p_\theta(z|x_i)}{q_\phi(z|t_i, x_i)} \int_{t_i}^{\infty} p_\theta(t|z)dt \right) dQ(z|t_i, x_i) \\
&= \mathbb{E}_{q_\phi(z|t,x)}[\log \frac{p_\theta(z|x_i)}{q_\phi(z|t_i, x_i)} + \log\ S_\theta(z, t_i|x_i)] \\
&= L_c(x_i, t_i; \theta, \phi)
\end{aligned}
$$

Likelihood for Multivariate Normal Distribution The Multivariate Normal Log-likelihood function used for ELBO has $z_l \sim N(\mu, \Sigma)$, where z_l is $m \times 1$, μ is

$m \times 1$ and Σ is $m \times m$:

$$\log \mathcal{L}(z_l; \mu, \Sigma) = -\frac{1}{2}[\log\ (|\Sigma|) + (z_l - \mu)^T \Sigma^{-1}(z_l - \mu) + m\log\ (2\pi)]$$

C.3 Likelihood Calculation for Statistical Models

C.3.1 AFT Models

Distributions of T_i and ϵ_i are corresponding to each other, with:

$$\begin{aligned} S_i(t|x_i) &= S_{\epsilon_i}(z_i) = S_0(\exp(\beta^T x_i)t) \\ z_i &= \frac{log(t) - \mu - \theta^T x_i}{\sigma} \\ h_i(t|x_i) &= \frac{1}{\sigma t_i} h_{\epsilon_i}(z_i) = h_0(\exp(\beta^T x_i)t) \end{aligned} \tag{C.1}$$

The likelihood function under general non-informative censoring has the form:

$$\begin{aligned} \mathcal{L}(\mathcal{D}; \theta) &= \prod_{i=1}^{n} h(t_i|x_i)^{\delta_i} S(t_i|x_i) \\ &= \prod_{i=1}^{n} (\sigma t_i)^{-\delta_i} \{h_{\epsilon_i}(z_i)\}^{\delta_i} S_{\epsilon_i}(z_i) \end{aligned} \tag{C.2}$$

For AFT model with Weibull distribution, ϵ_i follows the extreme value distribution and T_i follows the Weibull distribution, with scale parameter $\lambda > 0$, shape parameter

$\nu > 0$.

$$
\begin{aligned}
f_0(t) &= \lambda\nu t^{\nu-1}\exp(-\lambda t^{\nu}) \\
h_0(t) &= \lambda\nu t^{\nu-1} \\
S_0(t) &= \exp(-\lambda t^{\nu}) \\
\mathbb{E}(T) &= \frac{1}{\lambda^{1/\nu}}\Gamma(\frac{1}{\nu}+1) \\
S_\epsilon(z) &= \exp(-\exp(z)) \\
f_\epsilon(z) &= \exp(z)\exp(-\exp(z)) \\
h_\epsilon(z) &= \frac{f_\epsilon(z)}{S_\epsilon(z)} \\
&= \exp(z)
\end{aligned}
$$

Thus $\lambda = \exp(-\mu/\sigma)$ and $\nu = 1/\sigma$ in the extreme value distribution ϵ, equivalently $\mu = -\frac{\log\lambda}{\nu}$ and $\sigma = 1/\nu$. Thus the log-likelihood is:

$$
\begin{aligned}
&\log\mathcal{L}_{\text{Weibull}}(\mathcal{D};\theta) \\
&= \sum_{i=1}^{n}\left[-\delta_i\log(\sigma t_i) + \delta_i\log\{h_{\epsilon_i}(z_i)\} + \log S_{\epsilon_i}(z_i)\right] \\
&= \sum_{i=1}^{n}\left[\delta_i\log(\nu) - \delta_i\log(t_i) + \delta_i z_i - \exp(z_i)\right]
\end{aligned}
\tag{C.3}
$$

Note that for comparison log-likelihood at same scale, AFT and other parametric methods' log-likelihood were calculated at discretized event time same as VSI.

C.3.2 Random Survival Forest

Random Survival Forest could give estimates for cumulative hazard function at the pre-specifed timepoints $\hat{H}(t_{(b)}), b \in \{1, ldots, M\}$, which could let us get the estimated

survival function and time-to-event distribution.

$$\hat{S}(t_{(b)}) = \exp(-\hat{H}(t_{(b)}))$$

$$\hat{f}(t_{(1)}) = 1 - \hat{S}(t_{(1)})$$

$$\hat{f}(t_{(b)}) = \hat{S}(t_{(b)}) - \hat{S}(t_{(b-1)})$$

$$\hat{f}(t_{(M)}) = 1 - \sum_{b=1}^{M-1} \hat{f}(t_{(b)})$$

Then based on the survival likelihood function

$$\mathcal{L}(\mathcal{D};\theta) = \prod_{i=1}^{n} \hat{p}(t_i|x_i)^{\delta_i} \hat{S}(t_i|x_i)^{1-\delta_i}$$

, we could get the likelihood for RSF.

C.3.3 Cox Models

Since cox model is semiparametric, we don't know the full parameters, therefore usually we would give partial likelihood instead of full likelihood.

$$PL(\beta) = \prod_{i=1}^{n} \{\frac{\exp(\beta^T x_i)}{\sum_{l \in R(t_{(i)})} \exp(\beta^T x_i)}\}$$

$R(t_{(i)})$ is the set of subjects at risk at time $t_{(i)}$. Above partial likelihood is from multiply the probability that a subject with covariates x_i dies in $(t_{(i)}, t_{(i)} + \delta t]$ with $\delta t \to 0$ for each subject with event. $\hat{\beta}$ is obtained by Maximize $\log PL(\beta)$.

To make comparison between difference method, we would use some empirical way to calculate the full likelihood for Cox models.

$$\mathcal{L}(\mathcal{D};\theta) = \prod_{i=1}^{n} h(t_i|x_i)^{\delta_i} S(t_i|x_i)$$

$$\log \mathcal{L}(\mathcal{D};\theta) = \sum_{i=1}^{n} \delta_i [\log h_0(t_i) + \beta^T x_i] + \log S(t_i|x_i)$$

In above equations, we would use the estimated $h_0(t)$ and predicted $S(t_i|x_i)$.

C.4 Additional Results for Simulation Datasets

Distribution of log-likelihood The distributions of log-likelihoods for simulation datasets are shown in Figure C.1. In all scenarios, VSI has the most concentrated distribution of the log-likelihood, which suggests that VSI is robust under different event rate with CoxPH assumptions. MLP and VSI-NoQ performed equally good in this simple synthetic study. The percentile ranges of log-likelihood in simulation

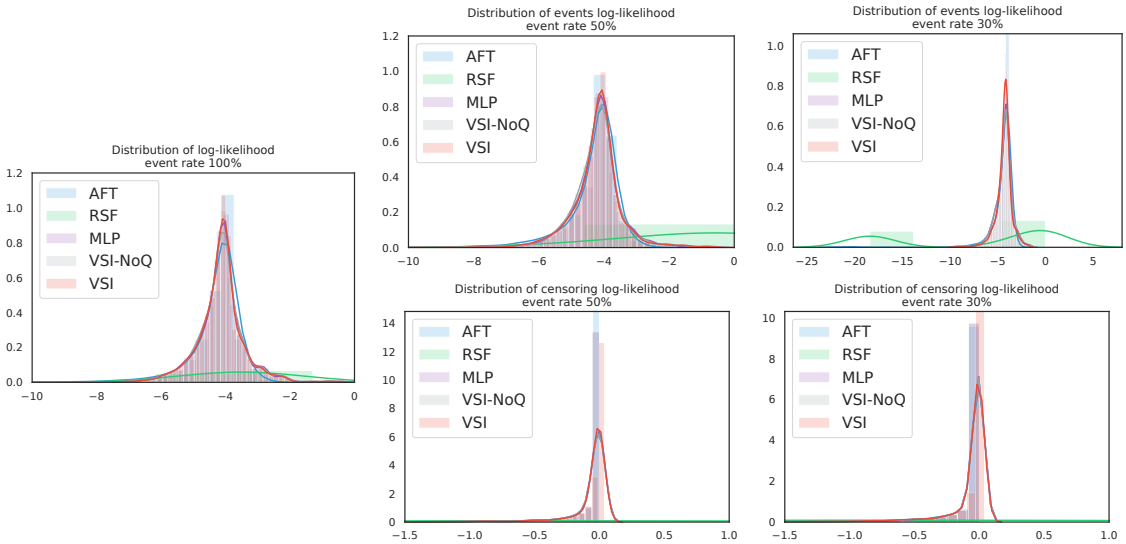

FIGURE C.1: Testing datasets log-likelihood distributions for simulation dataset.

studies are shown in Table C.1. When we have low event rates, VSI stands out from the baseline methods.

Confidence Intervals for Raw C-Index To better capture the differences in raw C-Index, the confidence intervals for each simulation strategy are calculated (See Table C.2). We have comparable performance with oracle methods with regarding to this metric. More illustrative metrics have been discussed in the text.

Table C.1: Quantile range [0.10, 0.90] of log-likelihood in simulation study

Models	Observed			Censored	
	100%	50%	30%	50%	30%
AFT-Weibull	1.55	1.67	1.87	0.26	0.28
MLP	1.66	1.62	1.98	0.24	0.23
VSI-NoQ	1.70	1.59	1.99	0.25	0.23
VSI	1.67	1.60	1.68	0.23	0.27

Table C.2: Raw C-Index with confidence intervals (in parentheses) for simulation studies. RSF and DeepSurv do not provide intrinsic methods to calculate confidence intervals

Event Rate	100%	50%	30%
CoxPH	0.773 (0.768, 0.777)	0.781 (0.775, 0.788)	0.793 (0.785, 0.802)
Coxnet	0.776 (0.762, 0.789)	0.784 (0.763, 0.805)	0.760 (0.730, 0.791)
AFT-Weibull	0.773 (0.768, 0.777)	0.781 (0.775, 0.788)	0.793 (0.785, 0.802)
RSF	0.701	0.718	0.712
DeepSurv	0.772	0.781	0.793
MLP	0.772 (0.767, 0.777)	0.781 (0.775, 0.788)	0.793 (0.784, 0.801)
VSI-NoQ	0.772 (0.768, 0.777)	0.781 (0.774, 0.788)	0.793 (0.784, 0.801)
VSI	0.773 (0.768, 0.777)	0.781 (0.775, 0.788)	0.793 (0.784, 0.801)

Coverage Rate To visualize the proportion of observed time is covered in the predicted personalized time-to-event distributions, we calculated the coverage rate for each percentile ranges. Our model balanced between events and censoring coverage the best among all methods. We compared the coverage with CoxPH, AFT-Weibull, VSI-NoQ, MLP and RSF. CoxPH could serves as the reference, since the simulation study is based on CoxPH assumptions entirely. The results are shown in Figure C.2.

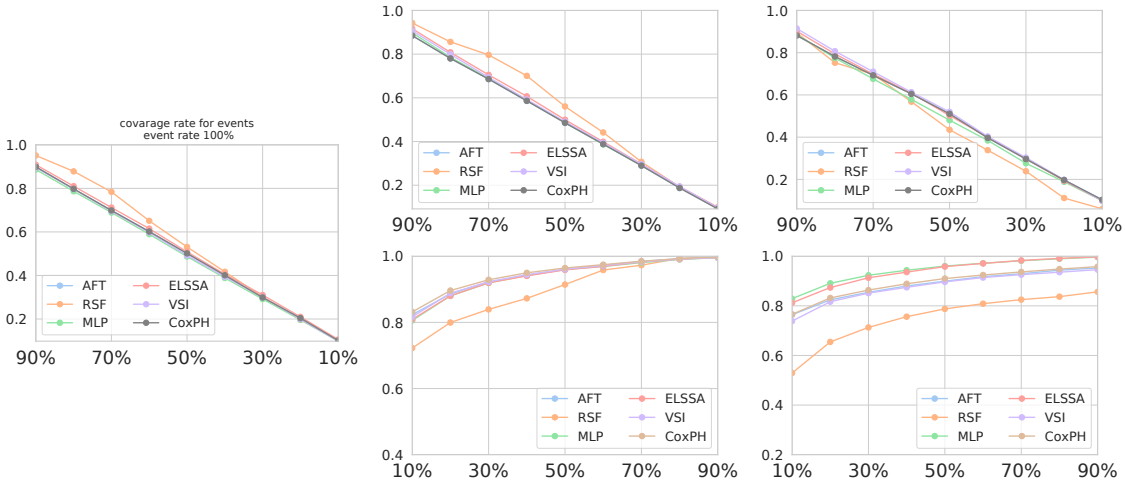

FIGURE C.2: Testing datasets coverage rate for simulation datasets.

C.5 Additional Results for Real Datasets

Comparison of Weighted Average and Median In Table C.3, we compared the C-Index calculated with weighted average and median for real datasets, with weighted average higher in general. In simulation studies, those two numbers are similar to each other.

Table C.3: Comparison median and weighted average for predicted time-to-event distribution with C-index in real-world datasets

C-Index	FLCHAIN	SUPPORT	SEER
Median	0.774	**0.833**	0.814
Weighted Average	**0.792**	0.823	**0.826**

Distribution for Log-likelihood Log-likelihood distribution for real datasets shown in Figure C.3. Similar to the case discussed in the text, VSI performs consistently better in all three realworld datasets.

Confidence Intervals for Raw C-Index As shown in Table C.4, raw C-Index for VSI is significantly better than other method with relatively tight confidence intervals.

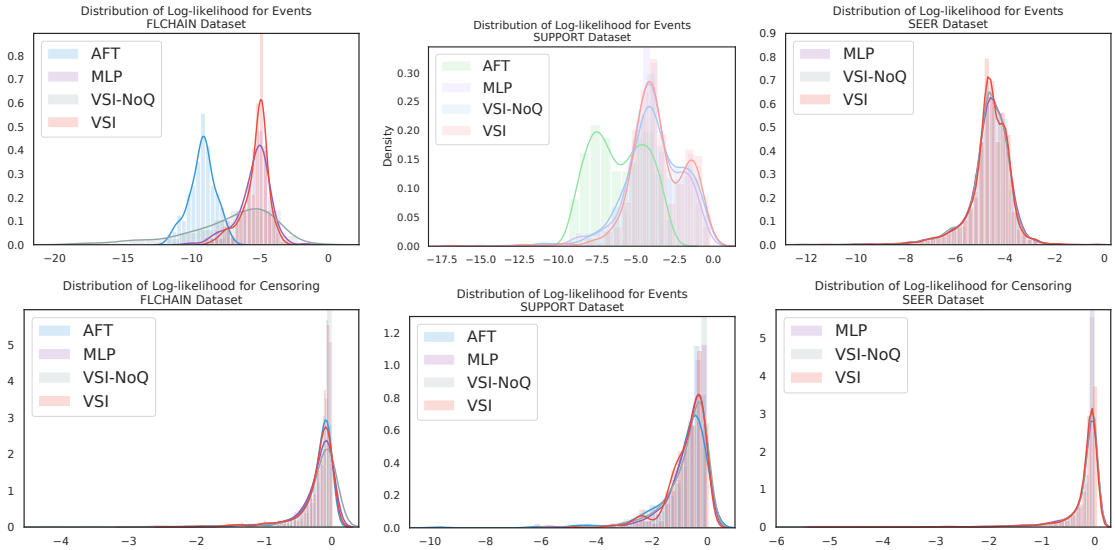

FIGURE C.3: Likelihood distributions for real datasets.

Coverage rate for real datasets Cover rate for real datasets shown in Figure C.4. VSI has relative high coverage for both events and censoring in all three datasets, which similar to the case we have in simulation datasets.

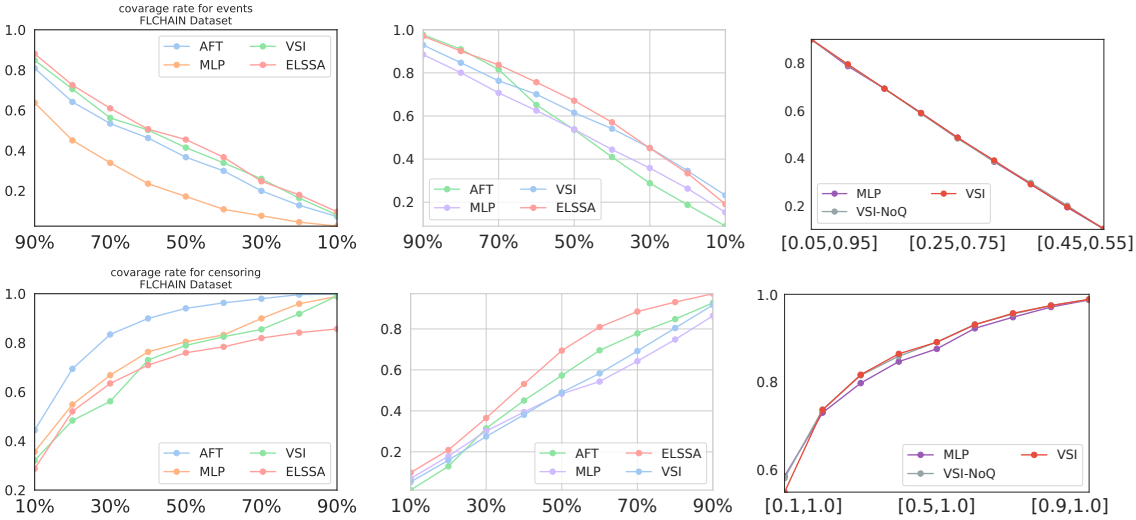

FIGURE C.4: Coverage rates for real-world datasets. FLCHAIN (observed: a, censored: b), SUPPORT (observed: c, censored: d), SEER (observed: e, censored: f)

For the comparison models, VSI, MLP and VSI-NoQ and AFT-Weibull are dis-

Table C.4: Raw C-Index with confidence intervals (in parentheses) for real-world datasets. NA indicates the corresponding evaluation metric cannot be evaluated. RSF and DeepSurv do not provide intrinsic methods to calculate confidence intervals.

Event Rate	FLCHAIN	SUPPORT	SEER
Coxnet	0.790 (0.761, 0.820)	0.784 (0.763, 0.805)	0.819 (0.812, 0.826)
AFT-Weibull	0.792 (0.763, 0.821)	0.797 (0.782, 0.813)	NA
RSF	0.771	0.751	0.796
DeepSurv	0.785	0.678	NA
MLP	0.751 (0.722, 0.781)	0.811 (0.791, 0.831)	0.811 (0.803, 0.818)
VSI-NoQ	0.745 (0.714, 0.777)	0.824 (0.804, 0.843)	0.809 (0.802, 0.817)
VSI	**0.792** (0.762, 0.821)	**0.827** (0.809, 0.846)	**0.826** (0.819, 0.833)

tribution based and could give all statistics in the simulation studies. For RSF and CoxPH, which could calculate the estimated cumulative hazards, the corresponding survival function and time-to-event could be calculated accordingly. For DeepSurv and Coxnet, only raw C-Index can be given.

Bibliography

Arxiv dataset. `https://www.kaggle.com/Cornell-University/arxiv`.

Tiny IMAGENET visual recognition challenge. `https://www.kaggle.com/c/tiny-imagenet`.

Aalen, O., Borgan, O., and Gjessing, H. (2008). *Survival and event history analysis: a process point of view*. Springer Science & Business Media.

Aalen, O. O. (1994). Effects of frailty in survival analysis. *Statistical Methods in Medical Research* **3,** 227–243.

Ahmad, M. A., Eckert, C., and Teredesai, A. (2018). Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pages 559–560.

Alaa, A. M. and van der Schaar, M. (2017). Deep multi-task gaussian processes for survival analysis with competing risks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 2326–2334. Curran Associates Inc.

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. (2016). Deep variational information bottleneck. In *ICLR*.

Antolini, L., Boracchi, P., and Biganzoli, E. (2005). A time-dependent discrimination index for survival data. *Statistics in Medicine* **24,** 3927–3944.

Antoniou, A., Storkey, A., and Edwards, H. (2017). Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340* .

Anzai, Y. (2012). *Pattern recognition and machine learning*. Elsevier.

Aranda-Ordaz, F. J. (1981). On two families of transformations to additivity for binary response data. *Biometrika* **68,** 357–363.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv preprint arXiv:1907.02893* .

Bacchetti, P. (1989). Additive isotonic models. *Journal of the American Statistical Association* **84,** 289–294.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *ICLR*.

Balkema, A. A. and De Haan, L. (1974). Residual life time at great age. *The Annals of probability* pages 792–804.

Bansal, N., Chen, X., and Wang, Z. (2018). Can we gain more from orthogonality regularizations in training deep networks? In *NeurIPS*.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). Statistical inference under order restrictions: The theory and application of isotonic regression. Technical report, Wiley New York.

Bedoya, A. D., Clement, M. E., Phelan, M., Steorts, R. C., O'Brien, C., and Goldstein, B. A. (2019). Minimal impact of implemented early warning score and best practice alert for patient deterioration. *Critical care medicine* **47,** 49–55.

Beltagy, I., Lo, K., and Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In *EMNLP*. ACL.

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate cox proportional hazards models. *Statistics in medicine* **24,** 1713–1723.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35,** 1798–1828.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *JMLR* **3,** 993–1022.

Böhning, D., Mylona, K., and Kimber, A. (2015). Meta-analysis of clinical trials with rare events. *Biometrical Journal* **57,** 633–648.

Botev, Z. I. and Kroese, D. P. (2008). An efficient algorithm for rare-event probability estimation, combinatorial optimization, and counting. *Methodology and Computing in Applied Probability* **10,** 471–505.

Bou-Hamad, I., Larocque, D., Ben-Ameur, H., et al. (2011). A review of survival trees. *Statistics Surveys* **5,** 44–71.

Boyd, K., Costa, V. S., Davis, J., and Page, C. D. (2012). Unachievable region in precision-recall space and its effect on empirical evaluation. In *ICML*.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association* **80,** 580–598.

Bryson, M. C. (1974). Heavy-tailed distributions: properties and tests. *Technometrics* **16,** 61–68.

Bühlmann, P. (2018). Invariance, causality and robustness. *arXiv preprint arXiv:1812.08233* .

Burda, Y., Grosse, R., and Salakhutdinov, R. (2015). Importance weighted autoencoders. In *ICLR*.

Byrd, J. and Lipton, Z. (2019). What is the effect of importance weighting in deep learning? In *ICML*.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. (2019). Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*.

Chapfuwa, P., Li, C., Mehta, N., Carin, L., and Henao, R. (2020). Survival cluster analysis. In *ACM CHIL*.

Chapfuwa, P., Tao, C., Li, C., Page, C., Goldstein, B., Carin, L., and Henao, R. (2018). Adversarial time-to-event modeling. In *NIPS*.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* **16,** 321–357.

Chen, J., Feng, J., and Lu, W. (2018). A wiener causality defined by relative entropy. In *ICONIP*.

Chen, L., Tao, C., Zhang, R., Henao, R., and Carin, L. (2018). Variational inference and model selection with generalized evidence bounds. In *ICML*.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *SIGKDD*.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *ICML*.

Chen, T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. (2018). Neural ordinary differential equations. In *NIPS*.

Cheng, J., Dong, L., and Lapata, M. (2016). Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733* .

Choi, E., Bahadori, M. T., Searles, E., Coffey, C., Thompson, M., Bost, J., Tejedor-Sojo, J., and Sun, J. (2016). Multi-layer representation learning for medical concepts. In *PACM SIGKDD*.

Chopra, S., Hadsell, R., and LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *CVPR*.

Choromanska, A., Agarwal, A., and Langford, J. (2013). Extreme multi class classification. In *NIPS Workshop: eXtreme Classification, submitted*.

Chouldechova, A. and Roth, A. (2018). The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810* .

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., and Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology* **110,** 12–22.

Cogswell, M., Ahmed, F., Girshick, R., Zitnick, L., and Batra, D. (2015). Reducing overfitting in deep networks by decorrelating representations. *arXiv preprint arXiv:1511.06068* .

Coles, S., Bawa, J., Trenner, L., and Dorazio, P. (2001). *An introduction to statistical modeling of extreme values*, volume 208. Springer.

Comon, P. (1994). Independent component analysis, a new concept? *Signal processing* **36,** 287–314.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning* **20,** 273–297.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34,** 187–202.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. (2019). Autoaugment: Learning augmentation strategies from data. In *CVPR*.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. (2019). Class-balanced loss based on effective number of samples. In *CVPR*.

Dai, B., Dai, H., He, N., Liu, W., Liu, Z., Chen, J., Xiao, L., and Song, L. (2018). Coupled variational bayes via optimization embedding. In *NIPS*.

Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., and Bontempi, G. (2017). Credit card fraud detection: a realistic modeling and a novel learning strategy. *IEEE Transactions on Neural Networks and Learning Systems* **29,** 3784–3797.

Davis, P. J. and Rabinowitz, P. (2007). *Methods of numerical integration*. Courier Corporation.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society: Series B (Methodological)* **52,** 393–425.

Desjardins, G., Simonyan, K., Pascanu, R., et al. (2015). Natural neural networks. In *NIPS*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M., and Hankey, J. (2016). Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences* **113,** 2636–2641.

Dispenzieri, A., Katzmann, J. A., Kyle, R. A., Larson, D. R., Therneau, T. M., Colby, C. L., Clark, R. J., Mead, G. P., Kumar, S., Melton III, L. J., et al. (2012). Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population. In *Mayo Clinic Proceedings*.

Donsker, M. D. and Varadhan, S. S. (1983). Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on Pure and Applied Mathematics* **36,** 183–212.

Drummond, C., Holte, R. C., et al. (2003). C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *ICML Workshop*.

Elkan, C. (2001). The foundations of cost-sensitive learning. In *IJCAI*.

Falk, M., Hüsler, J., and Reiss, R.-D. (2010). *Laws of small numbers: extremes and rare events*. Springer Science & Business Media.

Faraggi, D. and Simon, R. (1995). A neural network model for survival data. *Statistics in medicine* **14,** 73–82.

Fard, M. J., Wang, P., Chawla, S., and Reddy, C. K. (2016). A bayesian perspective on early stage event prediction in longitudinal data. *IEEE Transactions on Knowledge and Data Engineering* **28,** 3126–3139.

Fei-Fei, L., Fergus, R., and Perona, P. (2006). One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence* **28,** 594–611.

Fenchel, W. (1949). On conjugate convex functions. *Canadian Journal of Mathematics* **1,** 73–77.

Fernández, T., Rivera, N., and Teh, Y. W. (2016). Gaussian processes for survival analysis. In *Advances in Neural Information Processing Systems*, pages 5021–5029.

Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80,** 27–38.

Fithian, W. and Hastie, T. (2014). Local case-control sampling: Efficient subsampling in imbalanced data sets. *The Annals of Statistics* **42,** 1693.

Fong, R. C. and Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3429–3437.

Fotso, S. (2018). Deep neural networks for survival analysis based on a multi-task framework. *arXiv preprint arXiv:1801.05512* .

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences* **55,** 119–139.

Gan, Z., Chen, C., Henao, R., Carlson, D., and Carin, L. (2015). Scalable deep poisson factor analysis for topic modeling. In *ICML*.

García-Laencina, P. J., Sancho-Gómez, J.-L., and Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: a review. *Neural Computing and Applications* **19,** 263–282.

Goldstein, B. A., Navar, A. M., Pencina, M. J., and Ioannidis, J. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association* **24,** 198–208.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *ICLR*.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research* **13,** 723–773.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. (2009). Covariate shift by kernel mean matching. *Dataset Shift in Machine Learning* **3,** 5.

Guo, Q., Chen, J., Wang, D., Yang, Y., Deng, X., Li, F., Carin, L., and Tao, C. (2021). Tight mutual information estimation with contrastive Fenchel-Legendre optimization. [Available online; accessed 28-May-2021].

Gutmann, M. and Hirayama, J.-i. (2012). Bregman divergence as general framework to estimate unnormalized statistical models. *arXiv preprint arXiv:1202.3727* .

Gutmann, M. and Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304. JMLR Workshop and Conference Proceedings.

Hahn, P. and Jeruchim, M. (1987). Developments in the theory and application of importance sampling. *IEEE Transactions on Communications* **35,** 706–714.

Han, H., Wang, W.-Y., and Mao, B.-H. (2005). Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *ICIC*.

Haneuse, S., Arterburn, D., and Daniels, M. J. (2021). Assessing missing data assumptions in ehr-based studies: A complex and underappreciated task. *JAMA Network Open* **4,** e210184–e210184.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama* **247,** 2543–2546.

Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., and Galstyan, A. (2019). Multitask learning and benchmarking with clinical time series data. *Scientific data* **6,** 1–18.

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *WCCI*.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* **21,** 1263–1284.

He, J., Spokoyny, D., Neubig, G., and Berg-Kirkpatrick, T. (2019). Lagging inference networks and posterior collapse in variational autoencoders. In *ICLR*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*.

Heidelberger, P. (1995). Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* **5,** 43–85.

Heinze-Deml, C., Peters, J., and Meinshausen, N. (2018). Invariant causal prediction for nonlinear models. *Journal of Causal Inference* **6,**.

Henao, R., Gan, Z., Lu, J., and Carin, L. (2015). Deep poisson factor modeling. In *NIPS*.

Hofmann, T., Schölkopf, B., and Smola, A. J. (2008). Kernel methods in machine learning. *The annals of statistics* **36,** 1171–1220.

Holte, R. C., Acker, L., Porter, B. W., et al. (1989). Concept learning and the problem of small disjuncts. In *IJCAI*.

Hougaard, P. (1995). Frailty models for survival data. *Lifetime Data Analysis* **1,** 255–273.

Huang, C., Li, Y., Change Loy, C., and Tang, X. (2016). Learning deep representation for imbalanced classification. In *CVPR*.

Huang, L., Yang, D., Lang, B., and Deng, J. (2018). Decorrelated batch normalization. In *CVPR*.

Huntley, D. A., Cho, D. W., Christman, J., and Csernansky, J. G. (1998). Predicting length of stay in an acute psychiatric hospital. *Psychiatric Services* **49,** 1049–1053.

Hyvarinen, A. and Morioka, H. (2016). Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in Neural Information Processing Systems* **29,** 3765–3773.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks* **13,** 411–430.

Hyvärinen, A. and Pajunen, P. (1999). Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks* **12,** 429–439.

Hyvarinen, A., Sasaki, H., and Turner, R. (2019). Nonlinear ica using auxiliary variables and generalized contrastive learning. In *AISTATS*, pages 859–868.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., Lauer, M. S., et al. (2008). Random survival forests. *The annals of applied statistics* **2,** 841–860.

Jia, K., Li, S., Wen, Y., Liu, T., and Tao, D. (2019). Orthogonal deep neural networks. *IEEE transactions on pattern analysis and machine intelligence* .

Kaggle, I. (2017). Planet: Understanding the amazon from space— kaggle. *Consulté sur https://www. kaggle. com/c/planet-understanding-the-amazon-from-space/data (Consulté le: 2018-01-25)* .

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*, volume 360. John Wiley & Sons.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association* **53,** 457–481.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. (2016). Deep survival: A deep cox proportional hazards network. *stat* **1050,** 2.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. (2018). Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology* **18,** 24.

Kessy, A., Lewin, A., and Strimmer, K. (2018). Optimal whitening and decorrelation. *The American Statistician* **72,** 309–314.

Khan, F. M. and Zubek, V. B. (2008). Support vector regression for censored data (svrc): a novel tool for survival analysis. In *ICDM*.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. (2020). Variational autoencoders and nonlinear ICA: A unifying framework. In *AISTATS*, pages 2207–2217.

King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis* **9,** 137–163.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). Improved variational inference with inverse autoregressive flow. In *NIPS*.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. In *ICLR*.

Klein, J. P. and Moeschberger, M. L. (2006). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

Knaus, W. A., Harrell, F. E., Lynn, J., Goldman, L., Phillips, R. S., Connors, A. F., Dawson, N. V., Fulkerson, W. J., Califf, R. M., Desbiens, N., et al. (1995). The support prognostic model: objective estimates of survival for seriously ill hospitalized adults. *Annals of internal medicine* **122,** 191–203.

Kohane, I. S., Aronow, B. J., Avillach, P., Beaulieu-Jones, B. K., Bellazzi, R., Bradford, R. L., Brat, G. A., Cannataro, M., Cimino, J. J., García-Barrio, N., et al. (2021). What every reader should know about studies using electronic health record data but may be afraid to ask. *Journal of medical Internet research* **23,**.

Kotsiantis, S., Kanellopoulos, D., Pintelas, P., et al. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering* **30,** 25–36.

Lau, B., Cole, S. R., and Gange, S. J. (2009). Competing risk regression models for epidemiologic data. *American journal of epidemiology* **170,** 244–256.

Lee, C., Yoon, J., and Van Der Schaar, M. (2019). Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering* **67,** 122–133.

Lee, C., Zame, W. R., Yoon, J., and van der Schaar, M. (2018). Deephit: A deep learning approach to survival analysis with competing risks. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Li, B., Liu, Y., and Wang, X. (2019). Gradient harmonized single-stage detector. In *AAAI*.

Li, H., Boimel, P., Janopaul-Naylor, J., Zhong, H., Xiao, Y., Ben-Josef, E., and Fan, Y. (2019). Deep convolutional neural networks for imaging data based survival analysis of rectal cancer. *arXiv preprint arXiv:1901.01449* .

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86,** 316–327.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *ICCV*.

Lisboa, P. J., Wong, H., Harris, P., and Swindell, R. (2003). A bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer. *Artificial Intelligence in Medicine* **28,** 1–25.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., and Yu, S. X. (2019). Large-scale long-tailed recognition in an open world. In *CVPR*.

Lu, D., Guo, F., and Li, F. (2020). Evaluating the causal effects of cellphone distraction on crash risk using propensity score methods. *Accident Analysis & Prevention* **143,** 105579.

Lu, D., Tao, C., Chen, J., Li, F., Guo, F., and Carin, L. (2020). Reconsidering generative objectives for counterfactual reasoning. In *NeurIPS*.

Luck, M., Sylvain, T., Cardinal, H., Lodi, A., and Bengio, Y. (2017). Deep learning for patient-specific kidney graft survival analysis. *arXiv preprint arXiv:1705.10245* .

Machado, J. T. and Lopes, A. M. (2020). Rare and extreme events: the case of covid-19 pandemic. *Nonlinear Dynamics* page 1.

Mani, I. and Zhang, I. (2003). knn approach to unbalanced data distributions: a case study involving information extraction. In *ICML Workshop*.

Mariani, G., Scheidegger, F., Istrate, R., Bekas, C., and Malossi, C. (2018). BAGAN: Data augmentation with balancing GAN. *arXiv preprint arXiv:1803.09655* .

Massey Jr, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association* **46,** 68–78.

McFadden, D. (1978). Modeling the choice of residential location. *Transportation Research Record* .

McNeil, A. J. (1997). Estimating the tails of loss severity distributions using extreme value theory. *ASTIN Bulletin: The Journal of the IAA* **27,** 117–137.

Mescheder, L., Nowozin, S., and Geiger, A. (2017). Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *ICML*.

Mitchell, G. F., Hwang, S.-J., Vasan, R. S., Larson, M. G., Pencina, M. J., Hamburg, N. M., Vita, J. A., Levy, D., and Benjamin, E. J. (2010). Arterial stiffness and cardiovascular events: the framingham heart study. *Circulation* **121,** 505.

Mitchell, T. M. (1999). Machine learning and data mining. *Communications of the ACM* **42,** 30–36.

Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. (2018). Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence* **41,** 1979–1993.

Moya, M. M. and Hush, D. R. (1996). Network constraints and multi-objective optimization for one-class classification. *Neural networks* **9,** 463–474.

Mullick, S. S., Datta, S., and Das, S. (2019). Generative adversarial minority oversampling. In *ICCV*.

Munro, D., Ersoy, O., Bell, M., and Sadowsky, J. (1996). Neural network learning of low-probability events. *IEEE Transactions on Aerospace and Electronic Systems* **32,** 898–910.

Murphy, S., Rossini, A., and van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association* **92,** 968–976.

Neal, R. M. (2001). Annealed importance sampling. *Statistics and computing* **11,** 125–139.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)* **135,** 370–384.

Noble, W. S. (2006). What is a support vector machine? *Nature biotechnology* **24,** 1565–1567.

Oh Song, H., Xiang, Y., Jegelka, S., and Savarese, S. (2016). Deep metric learning via lifted structured feature embedding. In *CVPR*.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* .

O'Brien, C., Goldstein, B. A., Shen, Y., Phelan, M., Lambert, C., Bedoya, A. D., and Steorts, R. C. (2020). Development, implementation, and evaluation of an in-hospital optimized early warning score for patient deterioration. *MDM Policy & Practice* **5,** 2381468319899663.

Pal, M. (2005). Random forest classifier for remote sensing classification. *International journal of remote sensing* **26,** 217–222.

Papamakarios, G., Pavlakou, T., and Murray, I. (2017). Masked autoregressive flow for density estimation. In *NIPS*.

Parikh, A. P., Täckström, O., Das, D., and Uszkoreit, J. (2016). A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933* .

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch.

Paulus, R., Xiong, C., and Socher, R. (2017). A deep reinforced model for abstractive summarization. In *ICLR*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830.

Pencina, M. J., Goldstein, B. A., and D'Agostino, R. B. (2020). Prediction models-development, evaluation, and clinical application. *The New England journal of medicine* **382,** 1583–1586.

Peters, J., Bühlmann, P., and Meinshausen, N. (2016). Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B (Statistical Methodology)* .

Phillips, L., Goh, G., and Hodas, N. (2019). Explanatory masks for neural network interpretability. *arXiv preprint arXiv:1911.06876* .

Pickands III, J. et al. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics* **3,** 119–131.

Poole, B., Ozair, S., Van Den Oord, A., Alemi, A., and Tucker, G. (2019). On variational bounds of mutual information. In *ICML*.

Quan, S. F., Howard, B. V., Iber, C., Kiley, J. P., Nieto, F. J., O'Connor, G. T., Rapoport, D. M., Redline, S., Robbins, J., Samet, J. M., et al. (1997). The sleep heart health study: design, rationale, and methods. *Sleep* **20,** 1077–1085.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology* **25,** 111–164.

Raftery, A. E., Madigan, D., and Volinsky, C. T. (1996). Accounting for model uncertainty in survival analysis improves predictive performance. *Bayesian statistics* **5,** 323–349.

Rainforth, T., Kosiorek, A. R., Le, T. A., Maddison, C. J., Igl, M., Wood, F., and Teh, Y. W. (2018). Tighter variational bounds are not necessarily better. In *ICML*.

Ranganath, R., Gerrish, S., and Blei, D. (2014). Black box variational inference. In *AISTATS*.

Ranganath, R., Perotte, A., Elhadad, N., and Blei, D. (2016). Deep survival analysis. In *MLHC*.

Ren, K., Qin, J., Zheng, L., Yang, Z., Zhang, W., Qiu, L., and Yu, Y. (2019). Deep recurrent survival analysis. In *AAAI*.

Resnick, S. (2003). *A probability path*. Birkhauser Verlag AG.

Rezende, D. J. and Mohamed, S. (2015). Variational inference with normalizing flows. In *ICML*.

Ries, L. G., Young, J., Keel, G., Eisner, M., Lin, Y., Horner, M., et al. (2007). Seer survival monograph: cancer survival among adults: Us seer program, 1988-2001, patient and tumor characteristics. *National Cancer Institute, SEER Program, NIH Pub* pages 193–202.

Robert, C. and Casella, G. (2013). *Monte Carlo statistical methods*. Springer Science & Business Media.

Robinson, J., Jegelka, S., and Sra, S. (2020). Strength from weakness: Fast learning using weak supervision. In *ICML*.

Rojas-Carulla, M., Schölkopf, B., Turner, R., and Peters, J. (2018). Invariant models for causal transfer learning. *The Journal of Machine Learning Research* **19,** 1309–1342.

Rothenhäusler, D., Meinshausen, N., Bühlmann, P., and Peters, J. (2018). Anchor regression: heterogeneous data meets causality. *arXiv preprint arXiv:1801.06229* .

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63,** 581–592.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association* **91,** 473–489.

Rubinstein, R. Y. and Kroese, D. P. (2013). *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning*. Springer Science & Business Media.

Rubinstein, R. Y. and Kroese, D. P. (2016). *Simulation and the Monte Carlo method*, volume 10. John Wiley & Sons.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1,** 206–215.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., and Zhong, C. (2021). Interpretable machine learning: Fundamental principles and 10 grand challenges. *arXiv preprint arXiv:2103.11251* .

Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., and Kloft, M. (2018). Deep one-class classification. In *ICML*.

Safavian, S. R. and Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics* **21,** 660–674.

Scharfstein, D. O. and Robins, J. M. (2002). Estimation of the failure time distribution in the presence of informative censoring. *Biometrika* **89,** 617–634.

Schmid, C. H. and Griffith, J. L. (2014). Multivariate classification rules: calibration and discrimination. *Wiley StatsRef: Statistics Reference Online* .

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks* **61,** 85–117.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. (2012). On causal and anticausal learning. *arXiv preprint arXiv:1206.6471* .

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE* **109,** 612–634.

Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., Luigs, H.-G., Mahlein, A.-K., and Kersting, K. (2020). Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence* **2,** 476–486.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *CVPR*.

Schwarzer, G., Vach, W., and Schumacher, M. (2000). On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. *Statistics in medicine* **19,** 541–561.

Shah, P., Kendall, F., Khozin, S., Goosen, R., Hu, J., Laramie, J., Ringel, M., and Schork, N. (2019). Artificial intelligence and machine learning in clinical development: a translational perspective. *NPJ digital medicine* **2,** 1–5.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal* **27,** 379–423.

Shaw, P., Uszkoreit, J., and Vaswani, A. (2018). Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155* .

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference* **90,** 227–244.

Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data* **6,** 60.

Siddharth, N., Paige, B., Van de Meent, J.-W., Desmaison, A., Goodman, N., Kohli, P., Wood, F., and Torr, P. (2017). Learning disentangled representations with semi-supervised deep generative models. In *NIPS*.

Sill, J. (1998). Monotonic networks. In *NIPS*.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software* **39,** 1.

Smith, G. B., Prytherch, D. R., Meredith, P., Schmidt, P. E., and Featherstone, P. I. (2013). The ability of the national early warning score (news) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation* **84,** 465–470.

Sohn, K. (2016). Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* **15,** 1929–1958.

Steinwart, I., Hush, D., and Scovel, C. (2005). A classification framework for anomaly detection. *Journal of Machine Learning Research* **6,**.

Steyerberg, E. W. and Vergouwe, Y. (2014). Towards better clinical prediction models: seven steps for development and an abcd for validation. *European heart journal* **35,** 1925–1931.

Sugiyama, M. (2015). *Introduction to statistical machine learning*. Morgan Kaufmann.

Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences* **116,** 14516–14525.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *CVPR*.

Tao, C., Chen, L., Dai, S., Chen, J., Bai, K., Wang, D., Feng, J., Lu, W., Bobashev, G., and Carin, L. (2019). On fenchel mini-max learning. In *NeurIPS*.

Tao, C., Nichols, T. E., Hua, X., Ching, C. R., Rolls, E. T., Thompson, P. M., Feng, J., Initiative, A. D. N., et al. (2017). Generalized reduced rank latent factor regression for high dimensional tensor fields, and neuroimaging-genetic applications. *NeuroImage* **144,** 35–57.

Tax, D. M. J. (2002). *One-class classification: Concept learning in the absence of counter-examples*. PhD thesis, Technische Universiteit Delft (The Netherlands).

Teshima, T., Sato, I., and Sugiyama, M. (2020). Few-shot domain adaptation by causal mechanism transfer. In *ICML*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58,** 267–288.

Tishby, N., Pereira, F. C., and Bialek, W. (2000). The information bottleneck method. *arXiv preprint physics/0004057* .

Tishby, N. and Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE.

Tokmakov, P., Wang, Y.-X., and Hebert, M. (2019). Learning compositional representations for few-shot recognition. In *ICCV*.

Tomczak, J. M. and Welling, M. (2018). Vae with a vampprior. In *AISTATS*.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. (2018). The iNaturalist species classification and detection dataset. In *CVPR*.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *NIPS*.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. (2016). Matching networks for one shot learning. In *NIPS*.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning* **1,** 1–305.

Walther, D., Rutishauser, U., Koch, C., and Perona, P. (2004). On the usefulness of attention for object recognition. In *Workshop on Attention and Performance in Computational Vision at ECCV*, volume 1. Citeseer.

Wang, P., Li, Y., and Reddy, C. K. (2019). Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)* **51,** 110.

Wang, Y., Yao, Q., Kwok, J. T., and Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)* **53,** 1–34.

Wang, Y.-X., Ramanan, D., and Hebert, M. (2017). Learning to model the tail. In *NIPS*.

Wehenkel, A. and Louppe, G. (2019). Unconstrained monotonic neural networks. In *NeurIPS*.

Weinberger, K. Q. and Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research* **10,**.

Wu, R., Zhang, A., Ilyas, I., and Rekatsinas, T. (2020). Attention-based learning for missing data imputation in holoclean. *Proceedings of Machine Learning and Systems* **2,** 307–325.

Xiu, Z., Chen, J., Goldstein, B. A., Henao, R., Carin, L., and Tao, C. (2020). Supercharging imbalanced data learning with causal representation transfer. *arXiv preprint arXiv:2011.12454* .

Xiu, Z., Tao, C., Gao, M., Davis, C., Goldstein, B., and Henao, R. (2021). Variational disentanglement for rare event modeling. In *AAAI*.

Xiu, Z., Tao, C., Goldstein, B. A., and Henao, R. (2020). Variational learning of individual survival distributions. In *ACM CHIL*.

Yen, S.-J. and Lee, Y.-S. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications* **36,** 5718–5727.

Yu, C.-N., Greiner, R., Lin, H.-C., and Baracos, V. (2011). Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In *NIPS*.

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., and Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine* **15,** e1002683.

Zhang, H. H. and Lu, W. (2007). Adaptive lasso for cox's proportional hazards model. *Biometrika* **94,** 691–703.

Zhang, L., Zhang, C., Quan, S., Xiao, H., Kuang, G., and Liu, L. (2020). A class imbalance loss for imbalanced object recognition. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **13,** 2778–2792.

Zhang, Q. and Zhou, M. (2018). Nonparametric bayesian lomax delegate racing for survival analysis with competing risks. In *NeurIPS*.

Zhou, Z.-H. and Liu, X.-Y. (2005). Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering* **18,** 63–77.

Zong, W., Huang, G.-B., and Chen, Y. (2013). Weighted extreme learning machine for imbalance learning. *Neurocomputing* **101,** 229–242.

Zupan, B., Demšar, J., Kattan, M. W., Beck, J. R., and Bratko, I. (1999). Machine learning for survival analysis: a case study on recurrence of prostate cancer. In *AIMDM*.