Data-driven Decisions — An Anomaly Detection Perspective

Shubhranshu Shekhar

Heinz College & Machine Learning Department
Carnegie Mellon University (CMU)
Pittsburgh, PA

Thesis Committee

Leman Akoglu	CMU (Co-chair)
Christos Faloutsos	CMU (Co-chair)
Daniel Nagin	CMU
David Choi	CMU
Jetson Leder-Luis	Boston University

Doctoral Dissertation
Submitted in partial fulfillment of the requirements
for the Degree of Doctor of Philosophy in Machine Learning and Public Policy

May 2023

To my love, Esha

Abstract

Anomaly detection (AD) algorithms are widely used for data-driven decision support in domains where quantifying risk is critical, such as identifying fraudulent healthcare providers in public health insurance, consumer lending, and detecting aberrant patterns in human electroencephalography (EEG) records. However, AD in decision support is challenging due to the multitude of data modalities (e.g. time-series, or structural data) and data scale, unavailability of ground truth labels for learning and evaluation, and difficulty in yielding human interpretable results for domain-specific problems. This thesis proposes to address the challenges and build intelligent detection systems with the following desirable properties: unsupervised, explainable, scalable, and equitable. Throughout, we propose novel AD algorithms that enable better decision support by addressing domain-specific key challenges such as including domain or expert knowledge, mitigating bias that may adversely affect minority groups, and handling aberrant behavior involving a group of actors. We present applications in public healthcare fraud, and health monitoring.

Acknowledgements

The journey towards achieving a PhD degree has been immensely rewarding, both in terms of knowledge acquisition and the people I have encountered along the way. I am deeply grateful to the invaluable support provided by my advisors, collaborators, colleagues, and loved ones throughout this journey.

First and foremost, I would like to express my deepest gratitude to my advisors, Leman Akoglu and Christos Faloutsos, for their unwavering support, guidance, and encouragement throughout my PhD journey. Their invaluable insights and feedback have been instrumental in shaping this thesis and my overall academic career. I feel fortunate to have had the opportunity to work with them, who are the best example of what it means to be a nurturing researcher.

To Michelle Wirtz and Diane Stidle, for their prompt assistance and support throughout the duration of the program.

To my collaborators, Jetson Leder-Luis, Jonathan Elmer, Leon Iasemidis, Neil Shah, Meng-Chieh Lee, Jaemin Yoo, Bryan Hooi, Dhivya Eswaran, Noah Hutson for their input, suggestions and constant support, and without whom this thesis wouldn't be possible.

To the members of my thesis committee, Jetson Leder-Luis, David Choi, and Daniel Nagin, for their valuable input and suggestions.

To my internship mentors, Deepak Pai, Sriram Ravindran, Moein Saleh for their encouragement and support.

To Meng-Chieh (Jeremy), Dimitris, Meghanath, Emaad, Siddhartha, Jaemin, Lingxiao, Yue, Andre, Max, Saharsh, Tuan, Hung for their camaraderie and intellectual stimulation, which made my journey a truly enriching one.

To my cohort for their companionship and shared experiences in courses, and outside the university, which have made this journey memorable.

To my parents for their unconditional love, and endless encouragement throughout my life. Their sacrifices, hard work, and dedication have been a constant source of inspiration for me.

To Esha, for your unwavering support, love, and understanding. Your love and encouragement have been my anchor and steered me through the challenges.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this thesis, I study the challenges and opportunities in **data-driven decision** support for high-stakes domains (e.g. public healthcare and well-being, clinical decision support, finance). In particular, the thesis focuses on development of Unsupervised and Explainable **Anomaly Detection (AD)** techniques to empower human decision making. Unsupervised AD techniques identify *rare* events, and observations that deviate from underlying data distribution characterizing normal behavior. AD finds applications in domains where quantifying risk is critical, such as identifying fraudulent healthcare providers in public health insurance (Shekhar, Leder-Luis, and Akoglu, 2023), and detecting aberrant patterns in human electroencephalography (EEG) records (Lee, Shekhar, Faloutsos, Hutson, and Iasemidis, 2021). However, AD in decision support is challenging due to multitude of data modalities (e.g. time-series, or structural data) and data scale, unavailability of ground truth labels for learning and evaluation, and difficulty in yielding human interpretable results for domain specific problems. As such, the thesis objective is to build intelligent detection systems with the following desirable properties that aid in decision support.

1. *Unsupervised* detection waives the need for laborious labeling by human experts.
2. *Explainable* tools are user-friendly and assist a human expert in investigation, verification and decision making.
3. *Equitable* detection avoids unjust impact on marginalized groups, since AD as-is can cause unjust flagging of societal minorities (w.r.t. race, sex, etc.) because of their standing as statistical minorities, when minority status does not indicate riskiness.

To this overarching goal, thesis work is broken down to *algorithms* with contributions mostly in anomaly detection, explainable ML, and real-world data mining *applications in decision support*.

Thesis Outline

(A) Algorithms

Chapter 2, based on (Shekhar and Akoglu, 2018), proposes a novel approach for anomaly detection that leverages privileged information to improve the accuracy of unsupervised learning methods. Suppose that our goal is to estimate riskiness of a surgery in three weeks after it is performed based on the information x available before the surgery. Classical detectors use x to learn rules to flag risky patients. However, for patients who had surgery before, there is information about procedures and complications during surgery, or in one or two weeks after surgery, and so on. Availability of such case specific knowledge is fairly common, which are ignored by traditional detectors. Since this kind of domain knowledge is available only for learning, and is not available for the new data points (patients prior to surgery), it is called Privileged Information (PI). The work analyzes how domain knowledge augmentation can benefit anomaly detection, not only when PI is unavailable at test time (as in traditional setup) but also when PI is strategically and willingly avoided at test time. Information that incurs overhead on resources (cost/storage/battery/etc.), run-time computation, or vulnerability could be designated as PI enabling resource-frugal, early, and preventive detection. We show how to incorporate PI into ensemble based detectors and propose SPI, which constructs frames/fragments of knowledge (specifically, density estimates) in the privileged space and transfers them to the anomaly scoring space through "imitation" functions that use only the partial information available for test examples.

Chapter 3, based on (Shekhar, Shah, and Akoglu, 2021), proposes a framework for detecting anomalies in a dataset while simultaneously ensuring fairness towards different subgroups within the dataset. Anomaly detectors are designed exactly to spot rare, statistical minority samples in the data with the hope that outlierness reflects riskiness. For a minority (as defined by race/ethnicity/sex/age/etc.) group, sample size is by definition small, which puts them at odds with AD algorithms. However, when minority status (e.g. Asian) does not reflect positive-class membership (e.g. fraud), AD produces unjust outcomes, by overly flagging the instances from the minority groups as outliers[1]. This conflation of statistical and societal minorities can further become an ethical matter. We discuss sources of bias in AD and its implications for minority groups, and which notions of fairness are suitable for AD that could mitigate the bias in traditional AD. One of the key challeges in fair AD is the absence of ground truth labels for evaluation. We address the challenges of fair AD and design FairOD targeting fairness criteria for AD including statistical parity, treatment parity and equality of opportunity.

Chapter 4, based on (Lee, Shekhar, Faloutsos, Hutson, and Iasemidis, 2021), proposes a novel, generalized framework GEN^{2}OUT to spot and rank *generalized* anomalies to assist domain experts in decision making e.g. to draw attention of a clinician to strange brain activities in multivariate EEG recordings of an epileptic patient. We characterize generalized (point and group) anomalies that may arise in multivariate time series data, for example, in EEG recording during pre-ictal, ictal, and post-ictal phases as seizures arrive as bursts of spatio-temporal activities. The chapter designed an algorithm to assign and compare scores to isolated spikes and groups of spikes, allowing to detect for suspicious events of potential interest to domain expert.

(B) Applications

Chapter 5, based on (Shekhar, Leder-Luis, and Akoglu, 2023), develops new tools to detect health care overbilling or fraud. The US federal government spends more than a trillion dollars per year on health care, largely provided by private third parties and reimbursed by the government. A major concern in this system is overbilling, waste and fraud by providers, who face incentives to misreport on their claims in order to receive higher payments. We develop an ensemble based unsupervised multi-view detector that uses massive Medicare claims data with different modalities – including patient medical history, provider coding patterns, and provider spending – to detect anomalous behavior consistent with fraud. We combine evidence from multiple unsupervised outlier detection algorithms that use different types of global and local analysis – estimating a hospital's impact on patient

[1] Throughout the thesis words *anomaly* and *outlier* are used interchangeably.

expenditure, identifying few ICD codes that a hospital uses differently than the norm, and comparing a hospital's distribution over DRGs to its peers – using which we create a final ranking of suspiciousness.

Chapter 6, based on (Shekhar, Eswaran, Hooi, Elmer, Faloutsos, and Akoglu, 2023), proposed a framework that can assist in early prediction of health outcomes. In the healthcare domain, characterizing the state of a patient in ICU can assist in prediction of health outcomes for the patient, and allow the hospital to redistribute their resources to in-need patients, and potentially achieve better health outcomes overall within the same amount of time. A critical factor in play is the accuracy of such predictions, since incorrectly predicting unfavorable health outcome (e.g withdrawal of life-sustaining therapies) could hinder equitable decision making in the ICU, and may also expose hospitals to very costly lawsuits. We collaborated with a clinician to understand the problem setting better, and to design a solution that is useful to experts for decision making. To that end, the chapter introduces BENEFITTER that unifies earliness and accuracy–competing goals since observing more data can achieve better predictive accuracy– through a cost/benefit framework, and jointly optimizes for the prediction accuracy and earliness. Though the event detection task is supervised due to the nature of the underlying application data, the focus is on effectiveness, and interpretability. Ultimately, we do not propose an autonomous algorithm, rather we provide experts with information that is both more accurate and more timely than currently possible, assisting them in decision making.

Part I

Algorithms

Chapter 2

Knowledge-Augmented Anomaly Detection

Chapter based on: Shubhranshu Shekhar and Leman Akoglu (2018). “Incorporating privileged information to unsupervised anomaly detection”. In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018*. Springer, pp. 87–104.

2.1 Introduction

Outlier detection in point-cloud data has been studied extensively (Aggarwal, 2013). In this work we consider a unique setting with a much sparser literature: the problem of augmenting privileged information into unsupervised anomaly detection. Simply put, privileged information (PI) is *additional* data/knowledge/information that is available only at the learning/model building phase for (subset of) training examples, which however is unavailable for (future) test examples.

The LUPI framework. Learning Using Privileged Information (LUPI) has been pioneered by Vapnik et al. first in the context of SVMs (Vapnik and Vashist, 2009; Vapnik and Izmailov, 2015) (PI-incorporated SVM is named SVM+), later generalized to neural networks (Vapnik and Izmailov, 2017). The setup involves an Intelligent (or non-trivial) Teacher at learning phase, who provides the Student with *privileged* information (like explanations, metaphors, etc.), denoted $\boldsymbol{x}_i^*$, about each training example $\boldsymbol{x}_i$, $i = 1 \ldots n$. The key point in this paradigm is that *privileged information is not available at the test phase* (when Student operates without guidance of Teacher). Therefore, the goal is to build models (in our case, detectors) that can leverage/incorporate such additional information but *yet, not depend on the availability of PI at test time*.

Example: The additional information $\boldsymbol{x}_i^*$'s belong to space X^* which is, generally speaking, different from space X. In other words, the feature spaces of vectors $\boldsymbol{x}_i^*$'s and $\boldsymbol{x}_i$'s do not overlap. As an example, consider the task of identifying cancerous biopsy images. Here the images are in pixel space X. Suppose that there is an Intelligent Teacher that can recognize patterns in such images relevant to cancer. Looking at a biopsy image, Teacher can provide a description like “Aggressive proliferation of A-cells into B-cells” or “Absence of any dynamic”. Note that such descriptions are in a specialized language space X^*, different from pixel space X. Further, they would be available only for a set of examples and not when the model is to operate autonomously in the future.

LUPI's advantages: LUPI has been shown to (*i*) improve rate of convergence for learning, i.e., require asymptotically fewer examples to learn (Vapnik and Vashist, 2009), as well as (*ii*) improve accuracy, when one can learn a model in space X^* that is not much worse than the best model in space X (i.e., PI is intelligent/non-trivial) (Vapnik and Izmailov, 2017). Motivated by these advantages, LUPI has been applied to a number of problems from action recognition (Niu, Li, and Xu, 2016) to risk modeling (Ribeiro, Silva, Chen, Vieira, and Neves, 2012) (expanded in §2.5). However, the focus of all such work has mainly been on *supervised* learning.

LUPI for anomaly detection. The only (perhaps straightforward) extension of LUPI to unsupervised anomaly detection has been introduced recently, generalizing SVM+ to the One-Class SVM (namely OC-SVM+) (Burnaev and Smolyakov, 2016) for malware and bot detection. The issue is that OC-SVM is not a reliable detector since it assumes that normal points can be separated from origin in a single hyperball—experiments on numerous benchmark datasets with ground truth by Emmott et al. that compared popular anomaly detection algorithms find that OC-SVM ranks at the bottom (Table 1, pg. 4 (Emmott, Das, Dietterich, Fern, and Wong, 2013a); also see our results in §2.4). We note that the top performer in (Emmott, Das, Dietterich, Fern, and Wong, 2013a) is the Isolation Forest (iForest) algorithm (Liu, Ting, and Zhou, 2008a), an ensemble of randomized trees.

Our contributions: Motivated by LUPI's potential value to learning and the scarcity in the literature of its generalization to anomaly detection, we propose a new technique called SPI (pronounced 'spy'), for Spotting anomalies with Privileged Information. Our work *bridges the gap (for the first time) between LUPI and unsupervised ensemble based anomaly detection* that is considered state-of-the-art (Emmott, Das, Dietterich, Fern, and Wong, 2013a). We summarize our main contributions as follows.

- **Study of LUPI for anomaly detection:** We analyze how LUPI can benefit anomaly detection, not only when PI is truly unavailable at test time (as in traditional setup) but also when PI is strategically and willingly avoided at test time. We argue that data/information that incurs overhead on resources ($$$/storage/battery/etc.), timeliness, or vulnerability, if designated as PI, can enable resource-frugal, early, and preventive detection (expanded in §3.2).

- **PI-incorporated detection algorithm:** We show how to incorporate PI into ensemble based detectors and propose SPI, which constructs frames/fragments of knowledge (specifically, density estimates) in the privileged space (X^*) and *transfers* them to the anomaly scoring space (X) through "imitation" functions that use only the partial information available for test examples. To the best of our knowledge, ours is the first attempt to leveraging PI for improving the state-of-the-art *ensemble methods* for anomaly detection within an *unsupervised* LUPI framework. Moreover, while SPI augments PI within the tree-ensemble detector iForest (Liu, Ting, and Zhou, 2008a), our solution can easily be applied to any other ensemble based detector (§2.3).

- **Applications:** Besides extensive simulation experiments, we employ SPI on three real-world case studies where PI respectively captures (*i*) expert knowledge, (*ii*) computationally-expensive features, and (*iii*) "historical future" data, which demonstrate the benefits that PI can unlock for anomaly detection in terms of accuracy, speed, and detection latency (§2.4).

2.2 Motivation: How can LUPI benefit anomaly detection?

The implications of the LUPI paradigm for anomaly detection is particularly exciting. Here, we discuss a number of detection scenarios and demonstrate that LUPI unlocks advantages for anomaly detection problems in multiple aspects.

In the original LUPI framework (Vapnik and Vashist, 2009), privileged information (hereafter PI) is defined as data that is available *only* at training stage for training examples but *unavailable at test time* for test examples. Several anomaly detection scenarios admit this definition directly. Interestingly, PI can also be specified as *strategically "unavailable"* for anomaly detection. That is, one can willingly avoid using certain data at test time (while incorporating such data into detection models at train phase[1]) in order to achieve resource efficiency, speed, and robustness. We organize detection scenarios into two with PI as (truly) Unavailable vs. Strategic, and elaborate with examples below. Table 2.1 gives a summary.

TABLE 2.1: Types of data used in anomaly detection with various overhead on resources ($$$, storage, battery, etc.), timeliness, and/or risk, *if used as privileged information can enable resource-frugal, early, as well as preventive detection*.

Properties vs. **Type of Privileged Info**	**Unavailable** **vs. Strategic**	**need** **Resources**	**cause** **Delay**	**incur** **Risk**
1. "historical future" data	U	n/a	n/a	n/a
2. after-the-fact data	U	n/a	n/a	n/a
3. advanced technical data	U	n/a	n/a	n/a
4. restricted-access data	U, S	✓		
5. expert knowledge	U, S	✓	✓	
6. compute-heavy data	S	✓	✓	
7. unsafe-to-collect data	S		✓	✓
8. easy-target-to-tamper data	S			✓

2.2.1 Unavailable PI

This setting includes typical scenarios, where PI is (truly) unknown for test examples.

1. *"historical future" data*: When training an anomaly detection model with offline/historical data that is over time (e.g., temporal features), one may use values both before *and after* time t while creating an example for each t. Such data is PI; not available when the model is deployed to operate in real-time.

2. *after-the-fact data*: In malware detection, the goal is to detect before it gets hold of and harms the system. One may have historical data for some (training) examples from past exposures, including measurements of system variables (number of disk/port read/writes, CPU usage, etc.). Such after-the-exposure measurements can be incorporated as PI.

3. *advanced technical data*: This includes scenarios where some (training) examples are well-understood but those to be detected are simply unknown. For example, the expected behavior of various types of apps on a system may be common domain knowledge that can be converted to PI, but such knowledge may not (yet) be available for new-coming apps.

2.2.2 Strategic PI

Strategic scenarios involve PI that can in principle be acquired but is willingly avoided at test time to achieve gains in resources, time, or risk.

[1] Note that training phase in anomaly detection does not involve the use of any labels.

4. *restricted-access data*: One may want to build models that do not assume access to private data or intellectual property at test time, such as source code (for apps or executables), *even if* they could be acquired through resources. Such information can also be truly unavailable, e.g. encrypted within the software.

5. *expert knowledge*: Annotations about some training examples may be available from experts, which are truly unavailable at test time. One could also strategically choose to avoid expert involvement at test time, which (a) may be costly to obtain and/or (b) cause significant delay, especially for real-time detection.

6. *compute-heavy data*: One may strategically choose not to rely on features that are computationally expensive to obtain, especially in real-time detection, but rather use such data as PI (which can be extracted offline at training phase). Such features not only cause delay but also require compute resources (which e.g., may drain batteries in detecting malware apps on cellphones).

7. *unsafe-to-collect data*: This involves cases where collecting PI at test time is unsafe/dangerous. For example, the slower a drone moves to capture high-resolution (privileged) images for surveillance, not only it causes delay but more importantly, the more susceptible it becomes to be taken down.

8. *easy-target-to-tamper data*: Finally, one may want to avoid relying on features that are easy for adversaries to tamper with. Examples to those features include self-reported data (like age, location, etc.). Such data may be available reliably for some training examples and can be used as PI.

In short, by strategically designating PI one can achieve resource, timeliness, and robustness gains for various anomaly detection tasks. Designating features that need resources as PI → allow resource-frugal ("lazy") detection; features that cause delay as PI → allow early/speedy detection; and designating features that incur vulnerability as PI → allow preventive and more robust detection.

In this subsection, we laid out a long list of scenarios that make LUPI-based learning particularly attractive for anomaly detection. In our experiments (§2.4) we demonstrate its premise for scenarios 1., 5. and 6. above using three real world datasets, while leaving others as what we believe interesting future investigations.

2.3 Privileged Info-Augmented Anomaly Detection

2.3.1 The Learning Setting

Formally, the input for the anomaly detection model at learning phase are tuples of the form

$$\mathcal{D} = \{(\boldsymbol{x}_1, \boldsymbol{x}_1^*), (\boldsymbol{x}_2, \boldsymbol{x}_2^*), \ldots, (\boldsymbol{x}_n, \boldsymbol{x}_n^*)\}\ ,$$

where $\boldsymbol{x}_i = (x_i^1, \ldots, x_i^d) \in X$ and $\boldsymbol{x}_i^* = (x_i^{*1}, \ldots, x_i^{*p}) \in X^*$. Note that this is an unsupervised learning setting where label information, i.e., y_i's are not available. The privileged information is represented as a feature vector $\boldsymbol{x}^* \in \mathbb{R}^p$ that is in space X^*, which is *additional to and different from* the feature space X in which the primary information is represented as a feature vector $\boldsymbol{x} \in \mathbb{R}^d$.

The important distinction from the traditional anomaly detection setting is that the input to the (trained) detector at testing phase are feature vectors

$$\{\boldsymbol{x}_{n+1}, \boldsymbol{x}_{n+2}, \ldots, \boldsymbol{x}_{n+m}\}\ .$$

That is, the (future) test examples do not carry any privileged information. The anomaly detection model is to score the incoming/test examples and make decisions solely based on the primary features $\boldsymbol{x} \in X$.

In this text, we refer to space X^* as the *privileged space* and to X as the *decision space*. Here, a key assumption is that the information in the privileged space is intelligent/nontrivial, that is, it allows to create models $f^*(\boldsymbol{x}^*)$ that detect anomalies with vectors $\boldsymbol{x}^*$ corresponding to vectors $\boldsymbol{x}$ with higher accuracy than models $f(\boldsymbol{x})$. As a result, the main question that arises which we address in this work is: "how can one use the knowledge of the information in space X^* to improve the performance of the desired model $f(\boldsymbol{x})$ in space X?"

In what follows, we present a first-cut attempt to the problem that is a natural knowledge transfer between the two feature spaces (called FT for feature transfer). We then lay out the shortcomings of such an attempt, and present our proposed solution SPI. We compare to FT (and other baselines) in experiments.

2.3.2 First Attempt: Incorporating PI by Transfer of Features

A natural attempt to learning under privileged information that is unavailable for test examples is to treat the task as a *missing data problem*. Then, typical techniques for data imputation can be employed where missing (privileged) features are replaced with their predictions from the available (primary) features.

In this scheme, one simply maps vectors $\boldsymbol{x} \in X$ into vectors $\boldsymbol{x}^* \in X^*$ and then builds a detector model in the transformed space. The goal is to find the transformation of vectors $\boldsymbol{x} = (x^1, \dots, x^d)$ into vectors $\boldsymbol{\phi}(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), \dots, \phi_p(\boldsymbol{x}))$ that minimizes the expected risk given as

$$R(\boldsymbol{\phi}) = \sum_{j=1}^{p} \min_{\phi_j} \int (x^{*j} - \phi_j(\boldsymbol{x}))^2 p(x^{*j}, \boldsymbol{x}) dx^{*j} d\boldsymbol{x} \,, \tag{2.1}$$

where $p(x^{*j}, \boldsymbol{x})$ is the joint probability of coordinate x^{*j} and vector $\boldsymbol{x}$, and functions $\phi_j(\boldsymbol{x})$ are defined by p regressors.

Here, one could construct approximations to functions $\phi_j(\boldsymbol{x})$, $j = \{1, \dots, p\}$ by solving p regression estimation problems based on the training examples

$$(\boldsymbol{x}_1, x_1^{*j}), \dots, (\boldsymbol{x}_n, x_n^{*j}), \;\; j = 1, \dots, p \,,$$

where $\boldsymbol{x}_i$'s are input to each regression ϕ_j and the jth coordinate of the corresponding vector $\boldsymbol{x}_i^*$, i.e. x_i^{*j}'s are treated as the output, by minimizing the regularized empirical loss functional

$$R(\phi_j) = \min_{\phi_j} \sum_{i=1}^{n} (x_i^{*j} - \phi_j(\boldsymbol{x}_i))^2 + \lambda_j \text{penalty}(\phi_j), \;\; j = 1, \dots, p \,. \tag{2.2}$$

Having estimated the transfer functions $\hat{\phi}_j$'s (using linear or non-linear regression techniques), one can then learn any desired anomaly detector $f(\hat{\boldsymbol{\phi}}(\boldsymbol{x}))$ using the training examples, which concludes the learning phase. Note that the detector does not require access to privileged features $\boldsymbol{x}^*$ and can be employed solely on primary features $\boldsymbol{x}$ of the test examples $i = n+1, \dots, m$.

2.3.3 Proposed SPI: Incorporating PI by Transfer of Decisions

Treating PI as missing data and predicting $\boldsymbol{x}^*$ from $\boldsymbol{x}$ could be a difficult task, when privileged features are complex and high dimensional (i.e., p is large). Provided $f^*(\boldsymbol{x}^*)$ is an accurate detection model, a more direct goal would be to *mimic its decisions*—the scores that f^* assigns to the training examples. Mapping *data* between

TABLE 2.2: Three building blocks of knowledge representation in artificial intelligence, in context of SVM-LUPI for classification (Vapnik and Izmailov, 2015) and SPI for anomaly detection [this chapter].

	SVM-LUPI	SPI (**Proposed**)
1. Fundamental elements of knowledge	support vectors	isolation trees
2. Frames (fragments) of the knowledge	kernel functions	tree anomaly scores
3. Structural connections of the frames	weighted sum	weighted sum (by L2R)

two spaces, as compared to *decisions*, would be attempting to solve a more general problem, that is likely harder and unnecessarily wasteful.

The general idea behind transferring decisions/knowledge (instead of data) is to identify a small number of elements in the privileged space X^* that well-approximate the function $f^*(\boldsymbol{x}^*)$, and then try to transfer them to the decision space—through the approximation of those elements in space X. This is the knowledge transfer mechanism in LUPI by Vapnik and Izmailov, 2015. They illustrated this mechanism for the (supervised) SVM classifier. We generalize this concept to unsupervised anomaly detection.

The knowledge transfer mechanism uses three building blocks of knowledge representation in AI, as listed in Table 2.2. We first review this concept for SVMs, followed by our proposed SPI. While SPI is clearly different in terms of the task it is addressing as well as in its approach, as we will show, it is inspired by and builds on the same fundamental mechanism.

Knowledge transfer for SVM:

The *fundamental elements* of knowledge in the SVM classifier are the support vectors. In this scheme, one constructs two SVMs; one in X space and another in X^* space. Without loss of generality, let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t$ be the support vectors of SVM solution in space X and $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_{t^*}^*$ be the support vectors of SVM solution in space X^*, where t and t^* are the respective number of support vectors.

The decision rule f^* in space X^* (which one aims to mimic) has the form

$$f^*(\boldsymbol{x}^*) = \sum_{k=1}^{t^*} y_k \alpha_k^* K^*(\boldsymbol{x}_k^*, \boldsymbol{x}^*) + b^* \;, \tag{2.3}$$

where $K^*(\boldsymbol{x}_k^*, \boldsymbol{x}^*)$ is the kernel function of similarity between support vector $\boldsymbol{x}_k^*$ and vector $\boldsymbol{x}^* \in X^*$, also referred as the *frames* (or *fragments*) of knowledge. Eq. (2.3) depicts the *structural connection* of these fragments, which is a weighted sum with learned weights α_k^*'s.

The goal is to approximate each fragment of knowledge $K^*(\boldsymbol{x}_k^*, \boldsymbol{x}^*)$, $k = 1, \ldots, t^*$ in X^* using the fragments of knowledge in X; i.e., the t kernel functions $K(\boldsymbol{x}_1, \boldsymbol{x}), \ldots, K(\boldsymbol{x}_t, \boldsymbol{x})$ of the SVM trained in X. To this end, one maps t-dimensional vectors

$$\boldsymbol{z} = (K(\boldsymbol{x}_1, \boldsymbol{x}), \ldots, K(\boldsymbol{x}_t, \boldsymbol{x})) \in Z$$

into t^*-dimensional vectors

$$\boldsymbol{z}^* = (K^*(\boldsymbol{x}_1^*, \boldsymbol{x}^*), \ldots, K^*(\boldsymbol{x}_{t^*}^*, \boldsymbol{x}^*)) \in Z^*$$

through t^* regression estimation problems. That is, the goal is to find regressors $\phi_1(\boldsymbol{z}), \ldots, \phi_{t^*}(\boldsymbol{z})$ in X such that

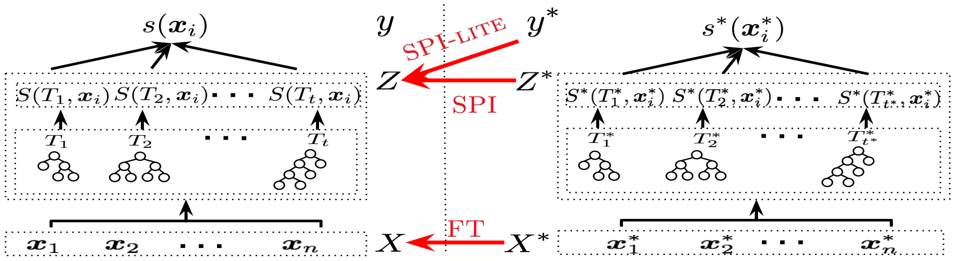

$s(\boldsymbol{x}_i)$
y
SPI-LITE
y^*
$s^*(\boldsymbol{x}_i^*)$
Z
Z^*
SPI
$S(T_1, \boldsymbol{x}_i)$ $S(T_2, \boldsymbol{x}_i)$ ... $S(T_t, \boldsymbol{x}_i)$
T_1 T_2 ... T_t
$S^*(T_1^*, \boldsymbol{x}_i^*)$ $S^*(T_2^*, \boldsymbol{x}_i^*)$ ... $S^*(T_{t^*}^*, \boldsymbol{x}_i^*)$
T_1^* T_2^* ... $T_{t^*}^*$
$\boldsymbol{x}_1$ $\boldsymbol{x}_2$ ... $\boldsymbol{x}_n$
X
FT
X^*
$\boldsymbol{x}_1^*$ $\boldsymbol{x}_2^*$ ... $\boldsymbol{x}_n^*$

FIGURE 2.1: Anomaly detection with PI illustrated. FT maps *data* between spaces (§2.3.2) whereas SPI (and "light" version SPI-LITE) mimic *decisions* (§2.3.3).

$$\phi_k(z_i) \approx K^*(x_k^*, x_i^*), \quad k = 1, \ldots, t^* \tag{2.4}$$

for all training examples $i = 1, \ldots, n$. For each $k = 1, \ldots, t^*$, one can construct the approximation to function ϕ_k by training a regression on the data

$$\{(z_1, K^*(x_k^*, x_1^*)), \; \ldots, \; (z_n, K^*(x_k^*, x_n^*))\}, \; k = 1, \ldots, t^* \; ,$$

where we regress vectors z_i's onto scalar output $K^*(x_k^*, x_i^*)$'s to obtain $\hat{\phi}_k$.

For the prediction of a test example x, one can then replace each $K^*(x_k^*, x^*)$ in Eq. (2.3) (which requires privileged features x^*) with $\hat{\phi}_k(z)$ (which mimics it, using only the primary features x—to be exact, by first transforming x into z through the frames $K(x_j, x), j = 1, \ldots, t$ in the X space).

Knowledge transfer for SPI**:**

In contrast to mapping of features from space X to space X^*, knowledge transfer of decisions maps space Z to Z^* in which fragments of knowledge are represented. Next, we show how to generalize these ideas to anomaly detection with no label supervision. Figure 2.1 shows an overview.

To this end, we utilize a state-of-the-art ensemble technique for anomaly detection, called Isolation Forest (Liu, Ting, and Zhou, 2008a) (hereafters IF, for short), which builds a set of extremely randomized trees. In essence, each tree approximates density in a random feature subspace and anomalousness of a point is quantified by the sum of such partial estimates across all trees.

In this setting, one can think of the individual trees in the ensemble to constitute the *fundamental elements* and the partial density estimates (i.e., individual anomaly scores from trees) to constitute the *fragments* of knowledge, where the structural connection of the fragments is achieved by an unweighted sum.

Similar to the scheme with SVMs, we construct two IFs; one in X space and another in X^* space. Let $\mathcal{T} = T_1, \ldots, T_t$ denote the trees in the ensemble in X and $\mathcal{T}^* = T_1^*, \ldots, T_{t^*}^*$ the trees in the ensemble in X^*, where t and t^* are the respective number of trees (prespecified by the user, typically a few 100s). Further, let $S^*(T_k^*, x^*)$ denote the anomaly score estimated by tree T_k^* for a given x^* (the lower the more anomalous; refer to (Liu, Ting, and Zhou, 2008a) for details of the scoring). $S(T_k, x)$ is defined similarly. Then, the anomaly score s^* for a point x^* in space X^* (which we aim to mimic) is written as

$$s^*(x^*) = \sum_{k=1}^{t^*} S^*(T_k^*, x^*) \; , \tag{2.5}$$

Algorithm 1 SPI-TRAIN: Incorporating PI to Unsupervised Anomaly Detector

```
Input: training examples $\{(\boldsymbol{x}_1, \boldsymbol{x}_1^*), \ldots, (\boldsymbol{x}_n, \boldsymbol{x}_n^*)\}$
Output: detection model (ensemble-of-trees) $\mathcal{T}$ in $X$ space; regressors $\hat{\phi}_k$'s, $k = 1, \ldots, t^*$; $\boldsymbol{\beta}$ (or $\gamma$ for kernelized L2R)
Learn $t^*$ isolation trees $\mathcal{T}^* = \{T_1^*, \ldots, T_{t^*}^*\}$ on $\boldsymbol{x}_i^*$'s $i = 1, \ldots, n$
Learn $t$ isolation trees $\mathcal{T} = \{T_1, \ldots, T_t\}$ on $\boldsymbol{x}_i$'s $i = 1, \ldots, n$
Construct leaf score vectors $\boldsymbol{z}_i$'s, $i = 1, \ldots, n$, based on $\mathcal{T}$
for each $k = 1, \ldots, t^*$ do
    Learn regressor $\hat{\phi}_k$ of $\boldsymbol{z}_i$'s onto $S^*(T_k^*, \boldsymbol{x}_i^*)$'s
    Obtain $\boldsymbol{\beta}$ by optimizing $C$ in (2.9) (or $\gamma$ for kernelized $C_\psi$)
end for
```

which is analogous to Eq. (2.3). To mimic/approximate each fragment of knowledge $S^*(T_k^*, \boldsymbol{x}^*)$, $k = 1, \ldots, t^*$ in X^* using the fragments of knowledge in X; i.e., the t scores for $\boldsymbol{x}$: $S(T_1, \boldsymbol{x}), \ldots, S(T_t, \boldsymbol{x})$ of the IF trained in X, we estimate t^* regressors $\phi_1(\boldsymbol{z}), \ldots, \phi_{t^*}(\boldsymbol{z})$ in X such that

$$\phi_k(\boldsymbol{z}_i) \approx S^*(T_k^*, \boldsymbol{x}_i^*), \quad k = 1, \ldots, t^* \tag{2.6}$$

for all training examples $i = 1, \ldots, n$, where $\boldsymbol{z}_i = (S(T_1, \boldsymbol{x}_i), \ldots, S(T_t, \boldsymbol{x}_i))$. Simply put, each $\hat{\phi}_k$ is an approximate mapping of all the t scores from the ensemble $\mathcal{T}$ in X to an individual score (fragment of knowledge) by tree T_k^* of the ensemble $\mathcal{T}^*$ in X^*. In practice, we learn a mapping from the leaves rather than the trees of $\mathcal{T}$ for a more granular mapping. Specifically, we construct vectors $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}', \ldots, \boldsymbol{z}_{it}')$ where each $\boldsymbol{z}_{ik}'$ is a size ℓ_k vector in which the value at index leaf$(T_k, \boldsymbol{x}_i)$ is set to $S(T_k, \boldsymbol{x}_i)$ and other entries to zero. Here, ℓ_k denotes the number of leaves in tree T_k and leaf$(\cdot)$ returns the index of the leaf that $\boldsymbol{x}_i$ falls into in the corresponding tree (note that $\boldsymbol{x}_i$ belongs to exactly one leaf of any tree, since the trees partition the feature space).

SPI-LITE: A "light" version. We note that instead of mimicking each individual fragment of knowledge $S^*(T_k^*, \boldsymbol{x}^*)$'s, one could also directly mimic the "final decision" $s^*(\boldsymbol{x}^*)$. To this end, we also introduce SPI-LITE, which estimates a *single* regressor $\phi(\boldsymbol{z}_i) \approx s^*(\boldsymbol{x}_i^*)$ for $i = 1, \ldots, n$ (also see Figure 2.1). We compare SPI and SPI-LITE empirically in §2.4.

Learning to Rank (L2R) like in X^* : An important challenge in learning to accurately mimic the scores s^*'s in Eq. (2.5) is to make sure that the regressors ϕ_k's are very accurate in their approximations in Eq. (2.6). Even then, it is hard to guarantee that the final ranking of points by $\sum_{k=1}^{t^*} \hat{\phi}_k(\boldsymbol{z}_i)$ would reflect their ranking by $s^*(\boldsymbol{x}_i^*)$. Our ultimate goal, after all, is to *mimic the ranking* of the ensemble in X^* space since anomaly detection is a ranking problem at its heart.

Algorithm 2 SPI-TEST: PI-Augmented Unsupervised Anomaly Detection

```
Input: test examples $\{\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_{n+m}\}$; $\mathcal{T}$, $\hat{\phi}_k$'s $k = 1, \ldots, t^*$, $\boldsymbol{\beta}$ (or $\gamma$ if kernelized)
Output: estimated anomaly scores $\{s_{n+1}, \ldots, s_{n+m}\}$ for all test examples
for each test example $\boldsymbol{x}_e$, $e = n+1, \ldots, n+m$ do
    Construct leaf score vector $\boldsymbol{z}_e = (\boldsymbol{z}_{e1}', \ldots, \boldsymbol{z}_{et}')$ where entry in each $\boldsymbol{z}_{ek}'$ for index leaf$(T_k, \boldsymbol{x}_e)$ is set to $S(T_k, \boldsymbol{x}_e)$ and to 0 o.w., for $k = 1, \ldots, t$
    Construct $\boldsymbol{\phi}_e = (\hat{\phi}_1(\boldsymbol{z}_e), \ldots, \hat{\phi}_{t^*}(\boldsymbol{z}_e))$
    Estimate anomaly score as $s_e = \boldsymbol{\beta}\boldsymbol{\phi}_e^T$ (or $s_e = \sum_{l=1}^{n} \gamma_l K(\boldsymbol{\phi}_l, \boldsymbol{\phi}_e)$ if kernelized)
end for
```

To this end, we set up an additional pairwise learning to rank objective as follows. Let us denote by $\boldsymbol{\phi}_i = (\hat{\phi}_1(z_i), \ldots, \hat{\phi}_{t^*}(z_i))$ the t^*-dimensional vector of estimated knowledge fragments for each training example i. For each pair of training examples, we create a tuple of the form $((\boldsymbol{\phi}_i, \boldsymbol{\phi}_j), p^*_{ij})$ where

$$p^*_{ij} = P(s^*_i < s^*_j) = \sigma(-(s^*_i - s^*_j)) \,, \tag{2.7}$$

which is the probability that i is ranked ahead of j by anomalousness in X^* space (recall that lower s^* is more anomalous), where $\sigma(v) = 1/(1+e^{-v})$ is the sigmoid function. Notice that the larger the gap between the anomaly scores of i and j, the larger this probability gets (i.e., more surely i ranks above j).

Given the training pair tuples above, our goal of learning-to-rank is to estimate $\boldsymbol{\beta} \in \mathbb{R}^{t^*}$, such that

$$p_{ij} = \sigma(\Delta_{ij}) = \sigma(\boldsymbol{\beta}\boldsymbol{\phi}_i^T - \boldsymbol{\beta}\boldsymbol{\phi}_j^T) = \sigma(-\hat{s^*_i} + \hat{s^*_j})) \approx p^*_{ij}, \;\; \forall i,j \in \{1,\ldots,n\} \,. \tag{2.8}$$

We then utilize the cross entropy as our cost function over all (i,j) pairs, as

$$\min_{\boldsymbol{\beta}} \; C = \sum_{(i,j)} -p^*_{ij}\log(p_{ij}) - (1-p^*_{ij})\log(1-p_{ij}) = \sum_{(i,j)} -p^*_{ij}\Delta_{ij} + \log(1+e^{\Delta_{ij}}) \tag{2.9}$$

where p^*_{ij}'s are given as input to the learning as specified in Eq. (2.7) and p_{ij} is denoted in Eq. (2.8) and is parameterized by $\boldsymbol{\beta}$ that is to be estimated.

The objective function in (2.9) is convex and can be solved via a gradient-based optimization, where $\frac{\mathrm{dC}}{\mathrm{dfi}} = \sum_{(i,j)}(p_{ij} - p^*_{ij})(\boldsymbol{\phi}_i - \boldsymbol{\phi}_j)$ (details omitted for brevity). More importantly, in case the linear mapping $s^*_i \approx \boldsymbol{\beta}\boldsymbol{\phi}_i^T$ is not sufficiently accurate to capture the desired pairwise rankings, the objective can be *kernelized* to learn a *non-linear* mapping that is likely more accurate. The idea is to write $\boldsymbol{\beta}_\psi = \sum_{l=1}^n \gamma_l \psi(\boldsymbol{\phi}_l)$ (in the transformed space) as a weighted linear combination of (transformed) training examples, for feature transformation function $\psi(\cdot)$ and parameter vector $\boldsymbol{\gamma} \in \mathbb{R}^n$ to be estimated. Then, Δ_{ij} in objective (2.9) in the transformed space can be written as

$$\Delta_{ij} = \sum_{l=1}^{n} \gamma_l[\psi(\boldsymbol{\phi}_l)\psi(\boldsymbol{\phi}_i)^T - \psi(\boldsymbol{\phi}_l)\psi(\boldsymbol{\phi}_j)^T] = \sum_{l=1}^{n} \gamma_l[K(\boldsymbol{\phi}_l, \boldsymbol{\phi}_i) - K(\boldsymbol{\phi}_l, \boldsymbol{\phi}_j)]. \tag{2.10}$$

The kernelized objective, denoted C_ψ, can also be solved through gradient-based optimization where we can show partial derivatives (w.r.t. each γ_l) to be equal to $\frac{\partial C_\psi}{\partial \gamma_l} = \sum_{(i,j)}(p_{ij} - p^*_{ij})[K(\boldsymbol{\phi}_l, \boldsymbol{\phi}_i) - K(\boldsymbol{\phi}_l, \boldsymbol{\phi}_j)]$. Given the estimated γ_l's, prediction of score is done by $\sum_{l=1}^n \gamma_l K(\boldsymbol{\phi}_l, \boldsymbol{\phi}_e)$ for any (test) example e.

The SPI **Algorithm:**

We outline the steps of SPI for both training and testing (i.e., detection) in Algo. 1 and Algo. 2, respectively. Note that the test-time detection no longer relies on the availability of privileged features for the test examples, but yet be able to leverage/incorporate them through its training.

2.4 Experiments

We design experiments to evaluate our methods in two different settings:

1. **Benchmark Evaluation**: We show the effectiveness of augmenting PI (see Table 2.3) on 17 publicly available benchmark datasets.[2]

2. **Real-world Use Cases**: We conduct experiments on LingSpam[3] and BotOrNot[4] datasets to show that (*i*) domain-expert knowledge as PI improves spam detection, (*ii*) compute-expensive PI enables fast detection at test time, and (*iii*) "historical future" PI allows early detection of bots.

Baselines

We compare both SPI and SPI-LITE to the following baselines:

1. IF(X-only): Isolation Forest (Liu, Ting, and Zhou, 2008a) serves as a simple baseline that operates solely in decision space X. PI is not used neither for modeling nor detection.

2. OC-SVM+ (PI-incorporated): OC+ for short, is an extension of (unsupervised) One-Class SVM that incorporates PI as introduced in (Burnaev and Smolyakov, 2016).

3. FT(PI-incorporated): This is the direct feature transfer method that incorporates PI by learning a mapping $X \rightarrow X^*$ as we introduced in §2.3.2.

* IF* (X^*-only): IF that operates in X^* space. We report performance by IF* only for reference, since PI is unavailable at test time.

2.4.1 Benchmark Evaluation

The benchmark datasets do not have an explicit PI representation. Therefore, in our experiments we introduce PI as explained below.

Generating privileged representation.

For each dataset, we introduce PI by perturbing normal observations. We designate a small random fraction ($= 0.1$) of n normal data points as anomalies. Then, we randomly select a subset of p attributes and add zero-mean Gaussian noise to the designated anomalies along the selected subset of attributes with matching variances of the selected features. The p selected features represent PI since anomalies stand-out in this subspace due to added noise, while the rest of the d attributes represent X space. Using normal observations allows us to control for features that could be used as PI. Thus we discard the actual anomalies from these datasets where PI is unknown.

We construct 4 versions per dataset with varying fraction γ of perturbed features (PI) retained in X^* space. In particular, each set has γp features in X^*, and $(1-\gamma)p + d$ features in X for $\gamma \in \{0.9, 0.7, 0.5, 0.3\}$.

[2] http://agents.fel.cvut.cz/stegodata/Loda.zip

[3] http://csmining.org/index.php/ling-spam-datasets.html

[4] https://botometer.iuni.iu.edu/bot-repository/datasets/caverlee-2011/caverlee-2011.zip

TABLE 2.3: Mean Average Precision (MAP) on benchmark datasets (avg'ed over 5 runs) for $\gamma = 0.7$. Numbers in parentheses indicate rank of each algorithm on each dataset. iF* (for reference only) reports MAP in the X^* space.

Datasets	p+d	n	iF	OC+	FT	SPI-lite	SPI	iF*
breast-cancer	30	357	0.1279 (4)	0.0935 (6)	0.0974 (5)	0.4574 (3)	**0.5746** (2)	0.6773 (1)
ionosphere	33	225	0.0519 (4)	**0.2914** (1)	0.0590 (3)	0.0512 (5)	0.0470 (6)	0.0905 (2)
letter-recognition	617	4197	0.0889 (6)	0.1473 (4)	0.0908 (5)	0.3799 (3)	**0.6413** (2)	0.9662 (1)
multiple-features	649	1200	0.1609 (5)	0.1271 (6)	0.2044 (4)	0.6589 (3)	**0.8548** (2)	1.0000 (1)
wall-following-robot	24	2923	0.1946 (5)	0.2172 (4)	0.1848 (6)	0.4331 (3)	**0.5987** (2)	0.7538 (1)
cardiotocography	27	1831	0.2669 (5)	0.6107 (4)	0.2552 (6)	0.6609 (3)	**0.6946** (2)	0.8081 (1)
isolet	617	4497	0.1533 (5)	0.1561 (4)	0.1303 (6)	0.5084 (3)	**0.7124** (2)	0.9691 (1)
libras	90	216	0.1368 (5)	0.4479 (4)	0.0585 (6)	0.5175 (3)	**0.6806** (2)	1.0000 (1)
parkinsons	22	147	0.0701 (6)	0.0964 (4)	0.0714 (5)	0.1556 (3)	**0.1976** (1)	0.1778 (2)
statlog-satimage	36	3594	0.2108 (6)	0.5347 (5)	0.5804 (4)	0.9167 (3)	**0.9480** (2)	0.9942 (1)
gisette	4971	3500	0.1231 (4)	0.0814 (6)	0.0977 (5)	0.5593 (3)	**0.8769** (2)	0.9997 (1)
waveform-1	21	3304	0.1322 (4)	0.1481 (3)	0.0841 (6)	0.1234 (5)	**0.1556** (2)	0.4877 (1)
madelon	500	1300	0.7562 (5)	0.1167 (6)	**0.9973** (2)	0.9233 (4)	0.9925 (3)	1.0000 (1)
synthetic-control	60	400	0.3207 (6)	0.7889 (4)	0.6870 (5)	0.8103 (3)	**0.8539** (2)	0.9889 (1)
waveform-2	21	3304	0.1271 (5)	**0.2828** (2)	0.1014 (6)	0.1778 (3)	0.1772 (4)	0.2944 (1)
statlog-vehicle	18	629	0.1137 (6)	0.3146 (5)	0.6326 (4)	0.6561 (3)	**0.7336** (2)	1.0000 (1)
statlog-segment	18	1320	0.1250 (6)	0.2323 (4)	0.1868 (5)	0.3304 (3)	**0.3875** (2)	0.7399 (1)
(Average Rank)			(5.11)	(4.23)	(4.88)	(3.29)	(2.35)	(1.11)

Results

We report the results on perturbed datasets with $\gamma = 0.7$[5] as fraction of features retained in space X^*. Table 2.3 reports mean Average Precision (area under the precision-recall curve) against 17 datasets for different methods. The results are averaged across 5 independent runs on stratified train-test splits.

Our SPI outperforms competition in detection performance in most of the datasets. To compare the methods statistically, we use the non-parametric Friedman test (Demšar, 2006) based on the average ranks. Table 2.3 reports the ranks (in parentheses) on each dataset as well as the average ranks. With p-value $= 2.16 \times 10^{-11}$, we reject the null hypothesis that all the methods are equivalent using Friedman test. We proceed with Nemenyi post-hoc test to compare the algorithms pairwise and to find out the ones that differ significantly. The test identifies performance of two algorithms to be significantly different if their average ranks differ by at least the "critical difference" (CD). In our case, comparing 6 methods on 17 datasets at significance level $\alpha = 0.05$, CD $= 1.82$.

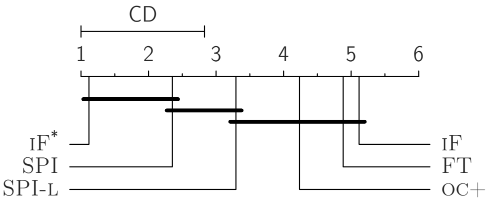


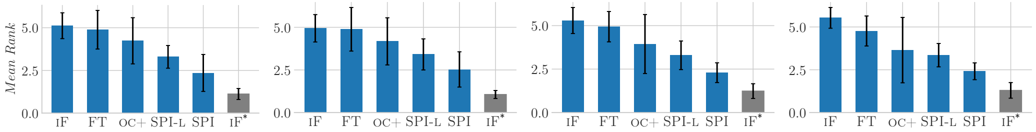
FIGURE 2.2: Average rank of algorithms (w.r.t. MAP) and comparison by the Nemenyi test. Groups of methods not significantly different (at p-val $= 0.05$) are connected with horizontal lines. CD depicts critical distance required to reject equivalence. Note that SPI is significantly better than the baselines.

Results of the post-hoc test are summarized through a graphical representation in Figure 2.2. We find that SPI is significantly better than all the baselines. We also notice that SPI has no significant difference from iF* which uses PI at test time,

[5] The results with $\gamma \in \{0.9, 0.5, 0.3\}$ are similar and reported in the supplementary material available at http://www.andrew.cmu.edu/user/shubhras/SPI

demonstrating its effectiveness in augmenting PI. While all the baselines are comparable to SPI-LITE, its average rank is better (also see last row in Table 2.3), followed by other PI-incorporated detectors, and lastly IF with no PI.

Average Precision (AP) is a widely-accepted metric to quantify overall performance of ranking methods like anomaly detectors. We also report average rank of the algorithms against other popular metrics including AUC of ROC curve, NDCG@10 and PRECISION@10 in Figure 2.3. Notice that the results are consistent across measures, SPI and SPI-LITE performing among the best.



(A) MAP (B) AUC (C) NDCG@10 (D) PRECISION@10

FIGURE 2.3: SPI and SPI-LITE outperform competition w.r.t. different evaluation metrics. Average rank (bars) across benchmark datasets. IF* shown for reference.

2.4.2 Real-world Use Cases

Data description. LingSpam dataset[3] consists of 2412 non-spam and 481 spam email messages from a linguistics mailing-list. We evaluate two use cases (1) domain-expert knowledge as PI and (2) compute-expensive PI on LingSpam.

BotOrNot dataset[4] is collected from Twitter during December 30, 2009 to August 2, 2010. It contains 22,223 content polluters (bots) and 19,276 legitimate users, along with their number of followings over time and tweets. For our experiments, we select accounts with age less than 10 days (for early detection task) at the beginning of dataset collection. The subset contains 901 legitimate (human) accounts and 4535 bots. We create 10 sets containing all the legitimate and a random 10% sample of the bots. We evaluate use case (3) "historical future" as PI and report the results averaged over these sets.

Case 1: Domain-expert Knowledge as PI for Email Spam Detection.

***X^** space:** The Linguistic Inquiry and Word Count (LIWC) software[6] is a widely used text analysis tool in social sciences. It uses a manually-curated keyword dictionary to categorize text into 90 psycholinguistic classes. Construction of LIWC dictionary relies exclusively on human experts which is a slow and evolving process. For the LingSpam dataset, we use the percentage of word counts in each class (assigned by LIWC software) as the privileged features. ***X* space:** The bag-of-word model is widely used as feature representation in text analysis. As such, we use the term frequencies for our email corpus as the primary features.

Figure 2.4 shows the detection performance[7] of algorithms in ROC curves (averaged over 15 independent runs on stratified train-test splits). We find that IF, which does not leverage PI but operates solely in X space, is significantly worse than most PI-incorporated methods. OC-SVM+ is nearly as poor as IF despite using PI—this is potentially due to OC-SVM being a poor anomaly detector in the first place, as shown in (Emmott, Das, Dietterich, Fern, and Wong, 2013a) and as we argued in

[6] https://liwc.wpengine.com/

[7] See supplementary material quantifying the performance of methods against other ranking metrics

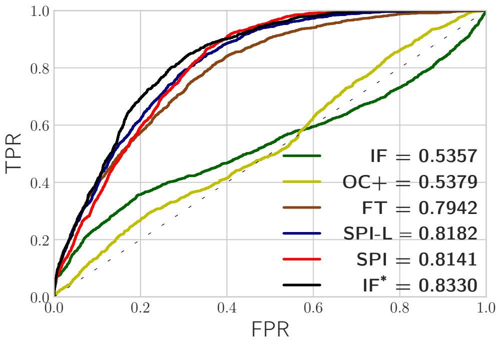


FIGURE 2.4: Detection performance on Case 1: using expert knowledge as PI. Legend depicts the AUC values. PI-incorporated detectors (except OC-SVM+) outperform non-PI IF and achieve similar performance to IF*.

§6.1. All knowledge transfer methods, SPI, SPI-LITE, and FT, perform similarly on this case study, and are as good as IF*, directly using X^*.

Case 2: Compute-Expensive Features as PI for Email Spam Detection.

X^* space: Beyond bag-of-words, one can use syntactic features to capture *stylistic* differences between spam and non-spam emails. To this end, we extract features from the parse trees of emails using the StanfordParser[8]. The parser provides the taxonomy (tree) of Part-of-Speech (PoS) tags for each sentence, based on which we construct (*i*) PoS bi-gram frequencies, and (*ii*) quantitative features (width, height, and horizontal/vertical imbalance) of the parse tree.

On average, StanfordParser requires 66 seconds[9] to parse and extract features from a single raw email in LingSpam. Since the features are computationally demanding, we incorporate those as PI to facilitate faster detection at test time.

X space: We use the term frequencies as the primary features as in Case 1.

Figure 2.5 (a) shows the detection performance[7] of methods in terms of AUC under ROC. We find that IF*using (privileged) syntactic features achieves lower AUC of ∼0.65 as compared to ∼0.83 using (privileged) LIWC features in Case 1. Accordingly, all methods perform relatively lower, suggesting that the syntactic features are less informative of spam than psycholinguistic ones. Nonetheless, we observe that the performance ordering remains consistent, where IF ranks at the bottom and SPI and SPI-LITE get closest to IF*.

Figure 2.5 (b) shows the comparison of wall-clock time required by each detector to compute the anomaly scores at test time for varying fraction of test data. On average, SPI achieves 5500× speed-up over IF* that employs the parser at test time. This is a considerable improvement of response time for comparable accuracy. Also notice the inset plot showing the AUC vs. total test time, where our proposed SPI and SPI-LITE are closest to the ideal point at the top left.

Case 3: "Historical Future" as PI for Twitter Bot Detection.

We use temporal data from the activity and network evolution of an account to capture behavioral differences between a human and a bot. We construct temporal features including volume, rate-of-change, and lag-autocorrelations of the number of

[8] https://nlp.stanford.edu/software/lex-parser.shtml

[9] Using a single thread on 2.2 GHz Intel Core i7 CPU with 8 cores and 16GB RAM

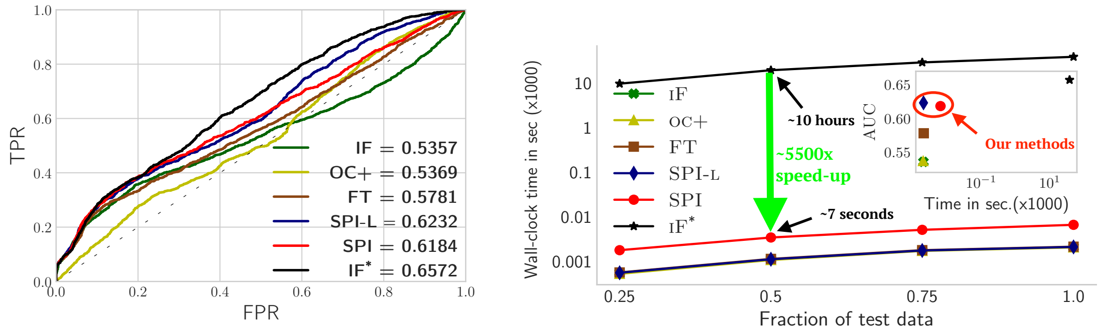


FIGURE 2.5: Comparison of detectors on Case 2: using computationally-expensive features as PI. (a) detection performance, legend depicts AUC values; and (b) wall-clock time required (in seconds, note the logarithmic scale) vs. test data size [inset plot on top right: AUC vs. time (methods depicted with symbols)].

followings. We also extract temporal features from text such as count of tweets, links, hash-tags and mentions.

X^* **space**: All the temporal features within f_t days in the future (relative to detection at time t) constitute privileged features. Such future values would not be available at any test time point but can be found in historical data.

X **space**: Temporal features within h_t days in the past as well as static user features (from screen name and profile description) constitute primary features.

Figure 2.6 (a) reports the detection performance of algorithms in terms of ROC curves (averaged over 10 sets) at time $t = 2$ days after the data collection started; for $h_t = 2$, $f_t = 7$.[10] The findings are similar to other cases: SPI and SPI-LITE outperform the competing methods in terms of AUC and OC-SVM+ performs similar to non-PI IF; demonstrating that knowledge transfer based methods are more suitable for real-world use cases.

Figure 2.6 (b) compares the detection performance of SPI and IF over time; for detection at $t = \{0, 1, 2, 3, 4\}$. As time passes, historical data grows as $h_t = \{0, 1, 2, 3, 4\}$ where "historical future" data is fixed at $f_t = 7$ for PI-incorporated methods. Notice that at time $t = 1$, SPI achieves similar detection performance to IF's performance at $t = 2$ that uses more historical data of 2 days. As such, SPI enables 24 hours early detection as compared to non-PI IF for the same accuracy. Notice that with the increase in historical data, the performances of both methods improve, as expected. At the same time, that of SPI improves faster, ultimately reaching a higher saturation level, specifically ∼7% higher relative to IF. Moreover, SPI gets close to IF*'s level in just around 3 days.

2.5 Related Work

We review the history of LUPI, follow up and related work on learning with side/hidden information, as well as LUPI-based anomaly detection.

Learning Under Privileged Information: The LUPI paradigm is introduced by Vapnik and Vashist, 2009 as the SVM+ method, where, Teacher provides Student not only with (training) examples but also explanations, comparisons, metaphors, etc. which accelerate the learning process. Roughly speaking, PI adjusts Student's concept of similarity between training examples and reduces the amount of data required for learning. Lapin et al. Lapin, Hein, and Schiele, 2014 showed that

[10] Same conclusions can be drawn for $f_t \in \{1, 3, 5, 7\}$ (see supplementary material).

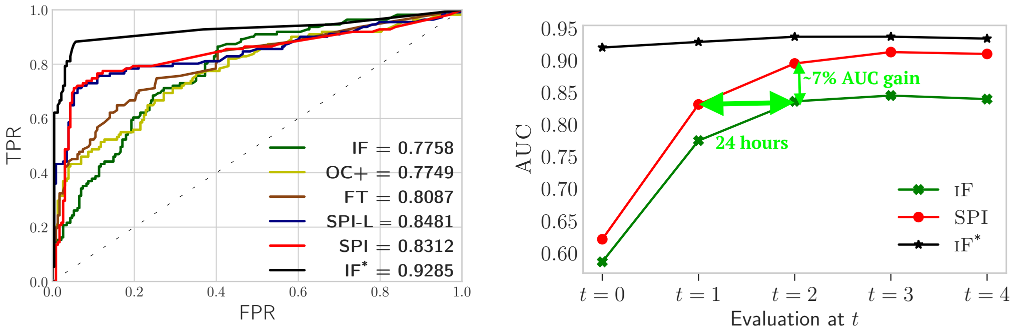


FIGURE 2.6: Comparison of detectors on Case 3: using "historical future" data as PI. (a) SPI outperforms competition in performance and is closest to IF*'s; (b) SPI achieves same detection performance as IF 24 hours earlier, and gets close to IF*in 3 days of history.

learning with PI is a particular instance of importance weighting in SVMs. Another such mechanism was introduced more recently by Vapnik and Izmailov, 2015, where knowledge is transferred from the space of PI to the space where the decision function is built. The general idea is to specify a small number of fundamental concepts of knowledge in the privileged space and then try to transfer them; i.e., construct additional features in decision space via e.g., regression techniques in decision space. Importantly, the knowledge transfer mechanism is not restricted to SVMs, but generalizes, e.g. to neural networks (Vapnik and Izmailov, 2017).

LUPI has been applied to a number of different settings including clustering (Feyereisl and Aickelin, 2012; Marcacini, Domingues, Hruschka, and Rezende, 2014), metric learning (Fouad, Tino, Raychaudhury, and Schneider, 2013), learning to rank (Sharmanska, Quadrianto, and Lampert, 2013), malware and bot detection (Burnaev and Smolyakov, 2016; Celik, McDaniel, Izmailov, Papernot, and Swami, 2016), risk modeling (Ribeiro, Silva, Chen, Vieira, and Neves, 2012), as well as recognizing objects (Sharmanska, Quadrianto, and Lampert, 2014), actions and events (Niu, Li, and Xu, 2016).

Learning with Side/Hidden Information: Several other work, particularly in computer vision (Chen, Liu, and Lyu, 2012; Wang and Ji, 2015), propose methods to learn with data that is unavailable at test time referred as side and hidden information (e.g., text descriptions or tags for general images, facial expression annotations for face images, etc.). In addition, Jonschkowski, Höfer, and Brock, 2015 describe various patterns of learning with side information. All of these work focus on supervised learning problems.

LUPI-based Anomaly Detection: With the exception of One-Class SVM (OC-SVM+) (Burnaev and Smolyakov, 2016), which is a direct extension of Vapnik's (supervised) SVM+, the LUPI framework has been utilized only for supervised learning problems. While anomaly detection has been studied extensively (Aggarwal, 2013), we are unaware of any work other than (Burnaev and Smolyakov, 2016) leveraging privileged information for unsupervised anomaly detection. As we argued in §3.2 and empirically demonstrated through benchmark experiments (Emmott, Das, Dietterich, Fern, and Wong, 2013a), OC-SVM+ is not an effective solution to anomaly detection. Motivated by this along with the premises of the LUPI paradigm, we are the first to design a new technique that ties LUPI with unsupervised tree-based ensemble methods, which are considered state-of-the-art for anomaly detection.

Chapter 3

Fairness-aware Outlier Detection

Chapter based on: Shubhranshu Shekhar, Neil Shah, and Leman Akoglu (2021). "Fairod: Fairness-aware outlier detection". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 210–220.

3.1 Introduction

Fairness in machine learning (ML) has received a surge of attention in the recent years. The community has largely focused on designing different notions of fairness (Barocas, Hardt, and Narayanan, 2017; Corbett-Davies and Goel, 2018a; Verma and Rubin, 2018) mainly tailored towards supervised ML problems (Hardt, Price, and Srebro, 2016; Zafar, Valera, Gomez Rodriguez, and Gummadi, 2017; Goel, Yaghini, and Faltings, 2018). However, perhaps surprisingly, fairness in the context of outlier detection (OD) is vastly understudied. OD is critical for numerous applications in security (Gogoi, Bhattacharyya, Borah, and Kalita, 2011; Zavrak and İskefiyeli, 2020; Zhang and Zulkernine, 2006), finance (Van Vlasselaer et al., 2015; Lee et al., 2020; Johnson and Khoshgoftaar, 2019), healthcare (Luo and Gallagher, 2010; Bosc, Heitz, Armspach, Namer, Gounot, and Rumbach, 2003) etc. and is widely used for detection of rare positive-class instances.

Outlier detection for "policing": In such critical systems, OD is often used to flag instances that reflect *riskiness*, which are then "policed" (or audited) by human experts. For example, law enforcement agencies might employ automated surveillance systems in public spaces to spot suspicious individuals based on visual characteristics, who are subsequently stopped and frisked. Alternatively, in the financial domain, analysts can police fraudulent-looking claims, and corporate trust and safety employees can police bad actors on social networks.

Group sample size disparity yields unfair OD: Importantly, outlier detectors are designed exactly to spot rare, *statistical minority* samples[1] with the hope that outlierness reflects riskiness, which prompts their bias against *societal minorities* (as defined by race/ethnicity/sex/age/etc.) as well, since minority group sample size is by definition small.

However, when minority status (e.g. Hispanic) does not reflect positive-class membership (e.g. fraud), OD produces ***unjust outcomes, by overly flagging the instances from the minority groups as outliers.*** This conflation of statistical and societal minorities can become an ethical matter.

Unfair OD leads to disparate impact: What would happen downstream if we did not strive for *fairness-aware* OD given the existence of societal minorities? OD

[1] In this work, the words sample, instance, and observation are used interchangeably throughout text.

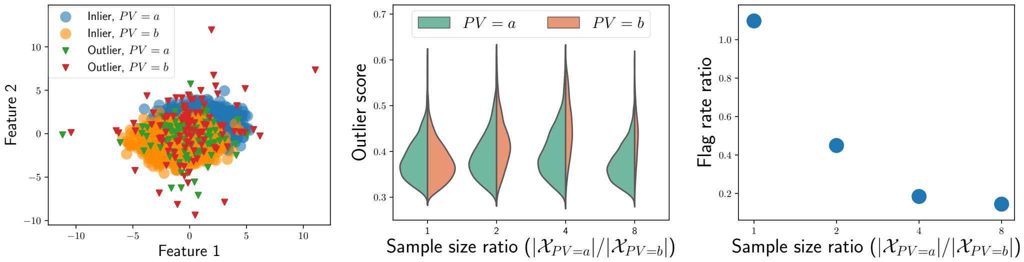


FIGURE 3.1: (left) Simulated 2-dim. data with equal sized groups i.e. $|\mathcal{X}_{PV=a}|=|\mathcal{X}_{PV=b}|$. (middle) Group score distributions induced by $PV = a$ and $PV = b$ are plotted by varying the simulated $|\mathcal{X}_{PV=a}|/|\mathcal{X}_{PV=b}|$ ratio. Notice that minority group ($PV = b$) receives larger outlier scores as the size ratio increases. (right) Flag rate ratio of the groups for the varying sample size ratio $|\mathcal{X}_{PV=a}|/|\mathcal{X}_{PV=b}|$. As we increase size disparity, the minority group is "policed" (i.e. flagged) comparatively more.

models' inability to distinguish societal minorities (as induced by so-called *protected* variables (PVs)), from statistical minorities, contributes to the likelihood of minority group members being flagged as outliers (see Fig. 3.1). This is further exacerbated by proxy variables which partially-redundantly encode (i.e. correlate with) the PV(s), by increasing the number of subspaces in which minorities stand out. The result is *overpolicing* due to over-representation of minorities in OD outcomes. Note that overpolicing the minority group also implies underpolicing the majority group given limited policing capacity and constraints.

Overpolicing can also feed *back* into a system when the policed outliers are used as labels in downstream supervised tasks. Alarmingly, this initially skewed sample (due to unfair OD), may be amplified through a feedback loop via predicting policing where more outliers are identified in more heavily policed groups. Given that OD's use in societal applications has direct bearing on social well-being, ensuring that OD-based outcomes are non-discriminatory is pivotal. This demands the design of fairness-aware OD models, which our work aims to address.

Prior research and challenges: Abundant work on algorithm fairness has focused on supervised ML tasks (Beutel et al., 2019; Hardt, Price, and Srebro, 2016; Zafar, Valera, Gomez Rodriguez, and Gummadi, 2017). Numerous notions of fairness (Barocas, Hardt, and Narayanan, 2017; Verma and Rubin, 2018) have been explored in such contexts, each with their own challenges in achieving equitable decisions (Corbett-Davies and Goel, 2018a). In contrast, there is little to no work on addressing fairness in *unsupervised* OD. Incorporating fairness into OD is challenging, in the face of (1) many possibly-incompatible notions of fairness and, (2) the absence of ground-truth outlier labels.

The two works tackling[2] unfairness in the OD literature are by P and Abraham, 2020 which proposes an ad-hoc procedure to introduce fairness specifically to the LOF algorithm (Breunig, Kriegel, Ng, and Sander, 2000), and Zhang and Davidson, 2020 (concurrent to our work) which proposes an adversarial training based deep SVDD detector. Amongst other issues (see Sec. 3.5), the approach proposed in (P and Abraham, 2020) invites disparate treatment, necessitating explicit use of *PV at decision time*, leading to taste-based discrimination (Corbett-Davies and Goel, 2018b) that is unlawful in several critical applications. On the other hand, the approach in (Zhang and Davidson, 2020) has several drawbacks (see Sec. 3.5), and in light

[2] Davidson and Ravi, 2020 aims to *quantify* fairness of OD model outcomes *post hoc*, which thus has a different scope.

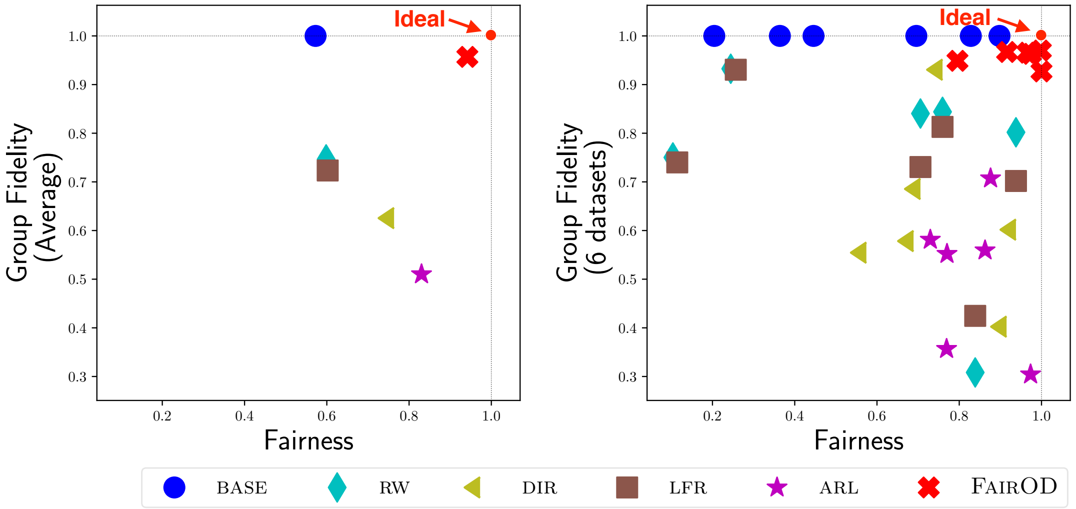


FIGURE 3.2: Fairness (statistical parity) vs. GroupFidelity (group-level rank preservation) of baselines and our proposed FAIROD (red cross), (left) averaged across 6 datasets, and (right) on individual datasets. FAIROD outperforms existing solutions (tending towards ideal), achieving fairness while preserving group fidelity from the BASE detector. See Sec. 6.5 for more details.

of unavailable implementation, we include a similar baseline called ARL that we compare against our proposed method.

Alternatively, one could re-purpose existing fair representation learning techniques (Zemel, Wu, Swersky, Pitassi, and Dwork, 2013; Edwards and Storkey, 2015; Beutel, Chen, Zhao, and Chi, 2017) as well as data preprocessing strategies (Kamiran and Calders, 2012; Feldman, Friedler, Moeller, Scheidegger, and Venkatasubramanian, 2015) for subsequent fair OD. However, as we show in Sec. 6.5 and discuss in Sec. 3.5, isolating representation learning from the detection task is suboptimal, largely (needlessly) sacrificing detection performance for fairness.

Our contributions: Our work strives to design a fairness-aware OD model to achieve equitable policing across groups and avoid an unjust conflation of statistical and societal minorities. We summarize our main contributions as follows:

1. **Desiderata & Problem Definition for Fair Outlier Detection:** We identify 5 properties characterizing detection quality and fairness in OD as desiderata for fairness-aware detectors. We discuss their justifiability and achievability, based on which we formally define the (unsupervised) fairness-aware OD problem (Sec. 3.2).
2. **Fairness Criteria & New, Fairness-Aware OD Model:** We introduce well-motivated fairness criteria and give mathematical objectives that can be optimized to obey the desiderata. These criteria are universal, in that they can be embedded into the objective of any end-to-end OD model. We propose FAIROD, a new detector which directly incorporates the prescribed criteria into its training. Notably, FAIROD (1) aims to equalize flag rates across groups, achieving group fairness via statistical parity, while (2) striving to flag truly high-risk samples within each group, and (3) avoiding disparate treatment. (Sec. 3.3)
3. **Effectiveness on Real-world Data:** We apply FAIROD on several real-world and synthetic datasets with diverse applications such as credit risk assessment and hate speech detection. Experiments demonstrate FAIROD's effectiveness in achieving both fairness goals (Fig. 3.2) as well as accurate detection (Fig. 3.6, Sec. 6.5), significantly outperforming alternative solutions.

3.2 Desiderata for Fair Outlier Detection

Notation

We are given N samples (also, observations or instances) $\mathcal{X} = \{X_i\}_{i=1}^N \subseteq \mathbb{R}^d$ as the input for OD where $X_i \in \mathbb{R}^d$ denotes the feature representation for observation i. Each observation is additionally associated with a binary[3] protected (also, sensitive) variable, $\mathcal{PV} = \{PV_i\}_{i=1}^N$, where $PV_i \in \{a, b\}$ identifies two groups – the majority ($PV_i = a$) group and the minority ($PV_i = b$) group. We use $\mathcal{Y} = \{Y_i\}_{i=1}^N$, $Y_i \in \{0, 1\}$, to denote the *unobserved* ground-truth binary labels for the observations where, for exposition, $Y_i = 1$ denotes an outlier (positive outcome) and $Y_i = 0$ denotes an inlier (negative outcome). We use $O : X \mapsto \{0, 1\}$ to denote the predicted outcome of an outlier detector, and $s : X \mapsto \mathbb{R}$ to capture the corresponding numerical outlier score as the estimate of the outlierness. Thus, $O(X_i), s(X_i)$ respectively indicate predicted outlier label and outlier score for sample X_i. We use $\mathcal{O} = \{O(X_i)\}_{i=1}^N$ and $\mathcal{S} = \{s(X_i)\}_{i=1}^N$ to denote the set of all predicted labels and scores from a given model without loss of generality. Note that we can derive $O(X_i)$ from a simple thresholding of $s(X_i)$. We routinely drop i-subscripts to refer to properties of a single sample without loss of generality. We denote the group *base rate* (or prevalence) of outlierness as $br_a = P(Y = 1|PV = a)$ for the majority group. Finally, we let $fr_a = P(O = 1|PV = a)$ depict the *flag rate* of the detector for the majority group. Similar definitions extend to the minority group with $PV = b$.

Having presented the problem setup and notation, we state our fair OD problem (informally) as follows.

Informal Problem 1 (Fair Outlier Detection)**.** *Given samples $\mathcal{X}$ and protected variable values $\mathcal{PV}$, estimate outlier scores $\mathcal{S}$ and assign outlier labels $\mathcal{O}$, such that*

(i) assigned labels and scores are "fair" w.r.t. the PV, and

(ii) higher scores correspond to higher riskiness encoded by the underlying (unobserved) $\mathcal{Y}$.

How can we design a fairness-aware OD model that is *not biased* against minority groups? What constitutes a "fair" outcome in OD, that is, what would characterize fairness-aware OD? What specific notions of fairness are most applicable to OD?

To approach the problem and address these motivating questions, we first propose a list of desired properties that an ideal fairness-aware detector should satisfy, followed by our proposed solution, FairOD.

3.2.1 Proposed Desiderata

D1. **Detection effectiveness:** We require an OD model to be accurate at detection, such that the scores assigned to the instances by OD are well-correlated with the ground-truth outlier labels. Specifically, OD benefits the policing effort only when the detection rate (also, precision) is strictly larger than the *base rate* (also, prevalence), that is,

$$P(Y = 1 \mid O = 1) > P(Y = 1) \,. \tag{3.1}$$

This condition ensures that any policing effort concerted through the employment of an OD model is able to achieve a *strictly larger precision* (LHS) *as compared to random sampling*, where policing via the latter would simply yield a precision that is equal to

[3] For simplicity of presentation, we consider a single, binary protected variable (PV). We discuss extensions to multi-valued PV and multi-attribute PVs in Appendix A.2.

the prevalence of outliers in the population (RHS) in expectation. Note that our first condition in (3.1) is related to detection performance, and specifically, the usefulness of OD itself for policing applications.

How-to: We can indirectly control for detection effectiveness via careful feature engineering. Assuming domain experts assist in feature design, it would be reasonable to expect a better-than-random detector that satisfies Eq. (3.1).

Next, we present *fairness-related* conditions for OD.

D2. **Treatment parity:** OD should exhibit non-disparate treatment that explicitly avoid the use of PV for producing a decision. In particular, OD decisions should obey

$$P(O=1 \mid X) \;=\; P(O=1 \mid X, PV=v),\ \forall v\ . \tag{3.2}$$

In words, the probability that the detector outputs an outlier label O for a given feature vector X remains unchanged even upon observing the value of the PV. In many settings (e.g. employment), explicit PV use is unlawful at inference.

How-to: We can build an OD model using a disparate learning process (Lipton, McAuley, and Chouldechova, 2018) that uses PV only during the model training phase, but does not require access to PV for producing a decision, hence satisfying treatment parity.

Treatment parity ensures that OD decisions are effectively "blindfolded" to the PV. However, this notion of fairness alone is not sufficient to ensure equitable policing across groups; namely, removing the PV from scope may still allow discriminatory OD results for the minority group (e.g., African American) due to the presence of several other features (e.g., zipcode) that (partially-)redundantly encode the PV. Consequently, by default, OD will use the PV *indirectly*, through access to those correlated proxy features. Therefore, additional conditions follow.

D3. **Statistical parity (SP):** One would expect the OD outcomes to be independent of group membership, i.e. $O \perp\!\!\!\perp PV$. In the context of OD, this notion of fairness (also, demographic parity, group fairness, or independence) aims to enforce that the outlier flag rates are independent of PV and equal across the groups as induced by PV.

Formally, an OD model satisfies statistical parity under a distribution over (X, PV) where $PV \in \{a, b\}$ if

$$\begin{gathered} fr_a = fr_b \text{ or equivalently,} \\ P(O=1|PV=a) \;=\; P(O=1|PV=b)\ . \end{gathered} \tag{3.3}$$

SP implies that the fraction of minority (majority) members in the flagged set is the same as the fraction of minority (majority) in the overall population. Equivalently, one can show

$$\begin{aligned} fr_a = fr_b \text{ (SP)} \iff & P(PV=a|O=1) = P(PV=a) \\ & \text{and } P(PV=b|O=1) = P(PV=b)\ . \end{aligned} \tag{3.4}$$

The motivation for SP derives from luck egalitarianism (Knight, 2009) – a family of egalitarian theories of distributive justice that aim to counteract the distributive effects of "brute luck". By redistributing equality to those who suffer through no fault of their own choosing, mediated via race, gender, etc., it aims to counterbalance the manifestations of such "luck". Correspondingly, SP ensures equal flag rates across PV groups, eliminating such group-membership bias. Therefore, it merits incorporation in OD since OD results are used for policing or auditing by human experts in downstream applications.

How-to: We could enforce SP during OD model learning by comparing the distributions of the predicted outlier labels O amongst groups, and update the model to ensure that these output distributions match across groups.

SP, however, is not sufficient to ensure both equitable *and* accurate outcomes as it permits so-called "laziness" (Barocas, Hardt, and Narayanan, 2017). Being an unsupervised quantity that is agnostic to the ground-truth labels $\mathcal{Y}$, SP could be satisfied while producing decisions that are arbitrarily inaccurate for any or all of the groups. In fact, an extreme scenario would be random sampling; where we select a certain fraction of the given population uniformly at random and flag all the sampled instances as outliers. As evident via Eq. (3.4), this entirely random procedure would achieve SP (!). The outcomes could be worse – that is, not only inaccurate (put differently, as accurate as random) but also unfair for only *some* group(s) – when OD flags mostly the true outliers from one group while flagging randomly selected instances from the other group(s), leading to discrimination *despite* SP. Therefore, additional criteria is required to explicitly penalize "laziness," aiming to not only flag *equal fractions* of instances across groups but also those *true outlier* instances from both groups.

D4. **Group fidelity (also, Equality of Opportunity):** It is desirable that the *true* outliers are equally likely to be assigned higher scores, and in turn flagged, regardless of their membership to any group as induced by PV. We refer to this notion of fairness as group fidelity, which steers OD outcomes toward being faithful to the ground-truth outlier labels equally across groups, obeying the condition

$$P(O=1|Y=1, PV=a) \;=\; P(O=1|Y=1, PV=b)\,. \tag{3.5}$$

Mathematically, this condition is equivalent to the so-called Equality of Opportunity[4] in the supervised fair ML literature, and is a special case of Separation (Verma and Rubin, 2018; Hardt, Price, and Srebro, 2016). In either case, it requires that all PV-induced groups experience the same true positive rate. Consequently, it penalizes "laziness" by ensuring that the true-outlier instances are ranked above (i.e., receive higher outlier scores than) the inliers within each group.

The key caveat here is that (3.5) is a supervised quantity that requires access to the ground-truth labels $\mathcal{Y}$, which are explicitly unavailable for the *unsupervised* OD task. What is more, various impossibility results have shown that certain fairness criteria, including SP and Separation, are mutually exclusive or incompatible (Barocas, Hardt, and Narayanan, 2017), implying that simultaneously satisfying both of these conditions (exactly) is not possible.

How-to: The unsupervised OD task does not have access to $\mathcal{Y}$, therefore, group fidelity cannot be enforced directly. Instead, we propose to enforce group-level rank preservation that maintains fidelity to within-group ranking from the BASE model, where BASE is a fairness-agnostic OD model. Our intuition is that rank preservation acts as a proxy for group fidelity, or more broadly Separation, via our assumption that within-group ranking in the BASE model is accurate and top-ranked instances within each group encode the highest risk samples within each group.

Specifically, let π^{BASE} represent the ranking of instances based on BASE OD scores, and let $\pi^{\text{BASE}}_{PV=a}$ and $\pi^{\text{BASE}}_{PV=b}$ denote the group-level ranked lists for majority and minority groups, respectively. Then, the rank preservation is satisfied when $\pi^{\text{BASE}}_{PV=v} =$

[4] Opportunity, because positive-class assignment by a supervised model in many fair ML problems is often associated with a positive outcome, such as being hired or approved a loan.

$\pi_{PV=v}; \forall v \in \{a,b\}$ where $\pi_{PV=v}$ is the ranking of group-v instances based on outlier scores from our proposed OD model. Group rank preservation aims to address the "laziness" issue that can manifest while ensuring SP; we aim to not lose the within-group detection prowess of the original detector while maintaining fairness. Moreover, since we are using only a proxy for Separation, the mutual exclusiveness of SP and Separation may no longer hold, though we have not established this mathematically.

D5. **Base rate preservation:** The flagged outliers from OD results are often audited and then used as human-labeled data for supervised detection (as discussed in previous section) which can introduce bias through a feedback loop. Therefore, it is desirable that group-level base rates within the flagged population is reflective of the group-level base rates in the overall population, so as to not introduce group bias of outlier incidence downstream. In particular, we expect OD outcomes to ideally obey

$$P(Y=1|O=1, PV=a) = br_a \text{ , and} \tag{3.6}$$
$$P(Y=1|O=1, PV=b) = br_b \text{ .} \tag{3.7}$$

Note that group-level base rate within the flagged population (LHS) is mathematically equivalent to group-level precision in OD outcomes, and as such, is also a supervised quantity which suffers the same caveat as in D4, regarding unavailability of $\mathcal{Y}$.

How-to: As noted, $\mathcal{Y}$ is not available to an unsupervised OD task. Importantly, provided an OD model satisfies D1 and D3, we show that it cannot simultaneously also satisfy D5, i.e. per-group equal base rate in OD results (flagged observations) and in the overall population.

Claim 1. *Detection effectiveness:* $P(Y=1|O=1) > P(Y=1)$ *and SP:* $P(O=1|PV=a) = P(O=1|PV=b)$ *jointly imply that* $P(Y=1|O=1, PV=v) > P(Y=1|PV=v), \exists v$.

Proof. We prove the claim in Appendix A.1.1. □

Claim 1 shows an incompatibility and states that, provided D1 and D3 are satisfied, the base rate in the flagged population cannot be equal to (but rather, is an overestimate of) that in the overall population for *at least one of the groups*. As such, base rates in OD outcomes cannot be reflective of their true values. Instead, one may hope for the preservation of the *ratio* of the base rates (i.e. it is not impossible). As such, a relaxed notion of D5 is to preserve proportional base rates across groups in the OD results, that is,

$$\frac{P(Y=1|O=1, PV=a)}{P(Y=1|O=1, PV=b)} = \frac{P(Y=1|PV=a)}{P(Y=1|PV=b)} \text{ .} \tag{3.8}$$

Note that ratio preservation still cannot be explicitly enforced as (3.8) is also label-dependent. Finally we show in Claim 2 that, provided D1, D3 and Eq. (3.8) are all satisfied, then it entails that the base rate in OD outcomes is an overestimation of the true group-level base rate *for every group*.

Claim 2. *Detection effectiveness:* $P(Y=1|O=1) > P(Y=1)$*, SP:* $P(O=1|PV=a) = P(O=1|PV=b)$*, and Eq.* (3.8)*:* $\frac{P(Y=1|O=1,PV=a)}{P(Y=1|O=1,PV=b)} = \frac{P(Y=1|PV=a)}{P(Y=1|PV=b)}$ *jointly imply* $P(Y=1|PV=v, O=1) > P(Y=1|PV=v), \forall v$.

Proof. We prove the claim in Appendix A.1.2. □

Claim 1 and Claim 2 indicate that if we have both (*i*) better-than-random precision (D1) and (*ii*) SP (D3), interpreting the base rates in OD outcomes for downstream learning tasks would not be meaningful, as they would not be reflective of true population base rates. Due to both these incompatibility results, and also feasibility issues given the lack of $\mathcal{Y}$, we leave base rate preservation – despite it being a desirable property – out of consideration.

3.2.2 Problem Definition

Based on the definitions and enforceable desiderata, our fairness-aware OD problem is formally defined as follows:

Problem 1 (Fairness-Aware Outlier Detection)**.** *Given samples $\mathcal{X}$ and protected variable values $\mathcal{PV}$, estimate outlier scores $\mathcal{S}$ and assign outlier labels $\mathcal{O}$, to achieve*

(i) $P(Y = 1|O = 1) > P(Y = 1)$,

[Detection effectiveness]

(ii) $P(O \mid X, PV = v) = P(O \mid X),\ \forall v \in \{a, b\}$,

[Treatment parity]

(iii) $P(O = 1|PV = a) = P(O = 1|PV = b)$,

[Statistical parity]

(iv) $\pi^{\text{BASE}}_{PV=v} = \pi_{PV=v}$ *,* $\forall v \in \{a, b\}$*, where* BASE *is a fairness-agnostic detector. [Group fidelity proxy]*

Given a dataset along with PV values, the goal is to design an OD model that builds on an existing BASE OD model and satisfies the criteria (i)–(iv), following the proposed desiderata D1 – D4.

3.2.3 Caveats of a Simple Approach

A simple yet naïve fairness-aware OD approach to address Problem 1 can be designed as follows:

1. Obtain ranked lists $\pi^{\text{BASE}}_{PV=a}$ and $\pi^{\text{BASE}}_{PV=b}$ from BASE, and
2. Flag top instances as outliers from each ranked list at equal fraction such that $P(O = 1|PV = a) = P(O = 1|PV = b), PV \in \{a, b\}$

This approach fully satisfies (iii) and (iv) in Problem 1 by design, as well as (i) given suitable features. However, it explicitly suffers from *disparate treatment*.

3.3 Fairness-aware Outlier Detection

In this section, we describe our proposed FAIROD – an unsupervised, fairness-aware, end-to-end OD model that embeds our proposed learnable (i.e. optimizable) fairness constraints into an existing BASE OD model. The key features of our model are that FAIROD aims for equal flag rates across groups (statistical parity), and encourages correct top group ranking (group fidelity), while not requiring PV for decision-making on new samples (non-disparate treatment). As such, it aims to target the proposed desiderata D1 – D4 as described in Sec. 3.2.

3.3.1 Base Framework

Our proposed OD model instantiates a deep-autoencoder (AE) framework for the base outlier detection task. However, we remark that the fairness regularization criteria introduced by FAIROD can be plugged into any end-to-end *optimizable* anomaly detector, such as one-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, and Williamson, 2001), deep anomaly detector (Chalapathy, Menon, and Chawla, 2018), variational AE for OD (An and Cho, 2015), and deep one-class classifiers (Ruff et al., 2018). Our choice of AE as the BASE OD model stems from the fact that AE-inspired methods have been shown to be state-of-the-art outlier detectors (Chen, Sathe, Aggarwal, and Turaga, 2017; Ma, Zhang, Cao, and Guo, 2013; Zhou and Paffenroth, 2017) and that our fairness-aware loss criteria can be optimized in conjunction with the objectives of such models. The main goal of FAIROD is to incorporate our proposed notions of fairness into an end-to-end OD model, irrespective of the choice of the BASE model family.

AE consists of two main components: an encoder $G_E : X \in \mathbb{R}^d \mapsto Z \in \mathbb{R}^m$ and a decoder $G_D : Z \in \mathbb{R}^m \mapsto X \in \mathbb{R}^d$. $G_E(X)$ encodes the input X to a hidden vector (also, code) Z that preserves the important aspects of the input. Then, $G_D(Z)$ aims to generate X', a reconstruction of the input from the hidden vector Z. Overall, the AE can be written as $G = G_D \circ G_E$, such that $G(X) = G_D(G_E(X))$. For a given AE based framework, the outlier score for X is computed using the reconstruction error as

$$s(X) = \|X - G(X)\|_2^2 \,. \tag{3.9}$$

Outliers tend to exhibit large reconstruction errors because they do not conform to to the patterns in the data as coded by an auto-encoder, hence the use of reconstruction errors as outlier scores (Aggarwal, 2015; Pang, Shen, Cao, and Hengel, 2020; Shah, Beutel, Gallagher, and Faloutsos, 2014). This scoring function is general in that it applies to many reconstruction-based OD models, which have different parameterizations of the reconstruction function G. We show in the following how FAIROD regularizes the reconstruction loss from BASE through fairness constraints that are conjointly optimized during the training process. The BASE OD model optimizes the following

$$\mathcal{L}_{\text{BASE}} = \sum_{i=1}^{N} \|X_i - G(X_i)\|_2^2 \tag{3.10}$$

and we denote its outlier scoring function as $s^{\text{BASE}}(\cdot)$.

3.3.2 Fairness-aware Loss Function

We begin with designing a loss function for our OD model that optimizes for achieving SP and group fidelity by introducing regularization to the BASE objective criterion. Specifically, FAIROD minimizes the following loss:

$$\boxed{\mathcal{L} = \alpha \underbrace{\mathcal{L}_{\text{BASE}}}_{\text{Reconstruction}} + (1-\alpha) \underbrace{\mathcal{L}_{SP}}_{\text{Statistical Parity}} + \gamma \underbrace{\mathcal{L}_{GF}}_{\text{Group Fidelity}}} \tag{3.11}$$

where $\alpha \in (0,1)$ and $\gamma > 0$ are hyperparameters which govern the balance between different fairness criteria and reconstruction quality in the loss function.

The first term in Eq. (3.11) is the objective for learning the reconstruction (based on BASE model family) as given in Eq. (3.10), which quantifies the goodness of the encoding Z via the squared error between the original input and its reconstruction generated from Z. The second component in Eq. (3.11) corresponds to regularization introduced to enforce the fairness notion of independence, or statistical parity (SP) as given in Eq. (3.4). Specifically, the term seeks to minimize the absolute correlation between the outlier scores $\mathcal{S}$ (used for producing predicted labels $\mathcal{O}$) and protected variable values $\mathcal{PV}$. $\mathcal{L}_{SP}$ is given as

$$\mathcal{L}_{SP} = \left| \frac{\left(\sum_{i=1}^{N} s(X_i) - \mu_s\right)\,\left(\sum_{i=1}^{N} PV_i - \mu_{PV}\right)}{\sigma_s\;\sigma_{PV}} \right| \tag{3.12}$$

where $\mu_s = \frac{1}{N}\sum_{i=1}^{N} s(X_i)$, $\sigma_s = \frac{1}{N}\sum_{i=1}^{N}(s(X_i) - \mu_s)^2$, $\mu_{PV} = \frac{1}{N}\sum_{i=1}^{N} PV_i$, and $\sigma_{PV} = \frac{1}{N}\sum_{i=1}^{N}(PV_i - \mu_{PV})^2$.

We adapt this absolute correlation loss from (Beutel et al., 2019), which proposed its use in a supervised setting with the goal of enforcing statistical parity. As Beutel et al., 2019 mentions, while minimizing this loss does not guarantee independence, it performs empirically quite well and offers stable training. We observe the same in practice; it leads to minimal associations between OD outcomes and the protected variable (see details in Sec. 6.5).

Finally, the third component of Eq. (3.11) emphasizes that FairOD should maintain fidelity to within-group rankings from the BASE model (penalizing "laziness"). We set up a listwise learning-to-rank objective in order to enforce group fidelity. Our goal is to train FairOD such that it reflects the within-group rankings based on $s^{\text{BASE}}(\cdot)$ from BASE. To that end, we employ a listwise ranking loss criterion that is based on the well-known Discounted Cumulative Gain (DCG) (Järvelin and Kekäläinen, 2002) measure, often used to assess ranking quality in information retrieval tasks such as search. For a given ranked list, DCG is defined as

$$\text{DCG} = \sum_{r} \frac{2^{rel_r} - 1}{\log_2(1 + r)}$$

where rel_r depicts the relevance of the item ranked at the r^{th} position. In our setting, we use the outlier score $s^{\text{BASE}}(X)$ of an instance X to reflect its relevance since we aim to mimic the group-level ranking by BASE. As such, DCG per group can be re-written as

$$\text{DCG}_{PV=v} = \sum_{X_i \in \mathcal{X}_{PV=v}} \frac{2^{s^{\text{BASE}}(X_i)} - 1}{\log_2\left(1 + \sum_{X_k \in \mathcal{X}_{PV=v}} \mathbb{1}[s(X_i) \leq s(X_k)]\right)}$$

where $\mathcal{X}_{PV=a}$ and $\mathcal{X}_{PV=b}$ would respectively denote the set of observations from majority and minority groups, and $s(X)$ is the estimated outlier score from our FairOD model under training.

A key challenge with DCG is that it is not differentiable, as it involves ranking (sorting). Specifically, the sum term in the denominator uses the (non-smooth) indicator function $\mathbb{1}(\cdot)$ to obtain the position of instance i as ranked by the estimated outlier scores. We circumvent this challenge by replacing the indicator function by the (smooth) sigmoid approximation, following (Qin, Liu, and Li, 2010). Then, the group fidelity loss component $\mathcal{L}_{GF}$ is given as

$$\mathcal{L}_{GF} = \sum_{v \in \{a,b\}} \left(1 - \sum_{X_i \in \mathcal{X}_{PV=v}} \frac{2^{s^{\text{BASE}}(X_i)} - 1}{\text{DNM}} \right) \tag{3.13}$$

$$\text{DNM} = \log_2 \Big(1 + \sum_{X_k \in \mathcal{X}_{PV=v}} \text{sigm}(s(X_k) - s(X_i))\Big) \cdot IDCG_{PV=v} \quad ,$$

$\text{sigm}(x) = \frac{\exp(-cx)}{1+\exp(-cx)}$ is the sigmoid function where $c > 0$ is the scaling constant, and, $IDCG_{PV=v} = \sum_{j=1}^{|\mathcal{X}_{PV=v}|} ((2^{s^{\text{BASE}}(X_j)} - 1)/\log_2(1+j))$ is the ideal (hence I), i.e. largest DCG value attainable for the respective group. Note that IDCG can be computed per group apriori to model training via BASE outlier scores alone, and serves as a normalizing constant in Eq. (3.13).

Note that having trained our model, scoring instances does not require access to the value of their PV, as PV is only used in Eq. (3.12) and (3.13) for training purposes. At test time, the anomaly score of a given instance X is computed simply via Eq. (3.9). Thus, FAIROD also fulfills the desiderata on treatment parity.

TABLE 3.1: Summary statistics of real-world and synthetic datasets used for evaluation.

Dataset	N	d	PV	PV = b	$\lvert\mathcal{X}_{\mathbf{PV=a}}\rvert/\lvert\mathcal{X}_{\mathbf{PV=b}}\rvert$	% outliers	Labels
Adult	25262	11	gender	*female*	4	5	{income $\leq 50K$, income $> 50K$}
Credit	24593	1549	age	*age* ≤ 25	4	5	{paid, delinquent}
Tweets	3982	10000	racial dialect	*African-American*	4	5	{normal, abusive}
Ads	1682	1558	simulated	1	4	5	{non-ad, ad}
Synth1	2400	2	simulated	1	4	5	$\{0,1\}$
Synth2	2400	2	simulated	1	4	5	$\{0,1\}$

Optimization and Hyperparameter Tuning

We optimize the parameters of FAIROD by minimizing the loss function given in Eq. (3.11) by using the built-in Adam optimizer (Kingma and Ba, 2014) implemented in PyTorch.

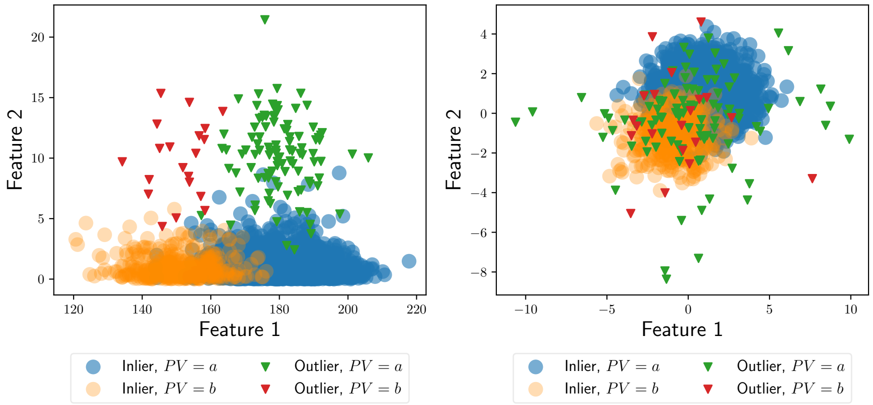
FAIROD comes with two tunable hyperparameters, α and γ. We define a grid for these and pick the configuration that achieves the best balance between SP and our proxy quantity for group fidelity (based on group-level ranking preservation). Note that both of these quantities are unsupervised (i.e., do not require access to ground-truth labels), therefore, FAIROD model selection can be done in a completely unsupervised fashion. We provide further details about hyperparameter selection in Sec. 6.5.

Generalizing to Multi-valued and Multiple Protected Attributes

FAIROD generalizes beyond binary PV, and easily applies to settings with multi-valued and multiple-protected attributes. We provide the details in Appendix A.2.

3.4 Experiments

Our proposed FAIROD is evaluated through extensive experiments on a set of synthetic datasets as well as diverse real-world datasets. In this section, we present dataset description and the experimental setup, followed by key evaluation questions and results.



(A) `Synth1` (B) `Synth2`

FIGURE 3.3: Synthetic datasets. See Sec. 3.4.1 for the details of the data generating process.

3.4.1 Dataset Description

Table 3.1 gives an overview of the datasets used in evaluation. A brief summary follows, with details on generative process of synthetic data and detailed descriptions.

Synthetic

We illustrate the efficacy of FAIROD on two synthetic datasets, `Synth1` and `Synth2`. These datasets present scenarios that mimic real-world settings, where we may have features that are uncorrelated with the outcome labels but partially correlated with the PV, or features which are correlated both to outcome labels and PV.

- `Synth1`: In `Synth1`, we simulate a 2-dimensional dataset comprised of samples $X = [x_1, x_2]$ where x_1 is correlated with the protected variable PV, but does not offer any predictive value with respect to ground-truth outlier labels $\mathcal{Y}$, while x_2 is correlated with these labels $\mathcal{Y}$ (see Fig. A.1a). We draw 2400 samples, of which $PV = a$ (majority) for 2000 points, and $PV = b$ (minority) for 400 points. 120 (5%) of these points are outliers. x_1 differs in terms of shifted means, but equal variances, for both majority and minority groups. x_2 is distributed similarly for both majority and minority groups, drawn from a normal distribution for outliers, and an exponential for inliers. The detailed generative process for the data is below (left), and Fig. A.1a shows a visual.

- `Synth2`: In `Synth2`, we again simulate a 2-dimensional dataset comprised of samples $X = [x_1, x_2]$ where x_1, x_2 are partially correlated with both the protected variable PV as well as ground-truth outlier labels $\mathcal{Y}$ (see Fig. A.1b). We draw 2400 samples, of which $PV = a$ (majority) for 2000 points, and $PV = b$ (minority) for 400 points. 120 (5%) of these points are outliers. For inliers, both x_1, x_2 are normally distributed, and differ across majority and minority groups only in terms of shifted means, but equal variances. Outliers are drawn from a product distribution of an exponential and linearly transformed Bernoulli distribution (product taken for symmetry). The detailed generative process for the data is below, and Fig. A.1b shows a visual.

`Synth1`

Simulate samples $X = [x_1, x_2]$ by...

$$PV \sim \text{Bernoulli}(4/5)$$

$$Y \sim \text{Bernoulli}(1/20)$$

$$x_1 \sim \begin{cases} \text{Normal}(-1, 1.44) & \text{if } \quad Y = 0,\ PV = 1 \quad \text{[a, majority; inlier]} \\ \text{Normal}(1, 1.44) & \text{if } \quad Y = 0,\ PV = 0 \quad \text{[b, minority; inlier]} \\ 2 \times \text{Exponential}(1)(1 - 2 \times \text{Bernoulli}(1/2)) & \text{if } \quad Y = 1 \quad \text{[outlier]} \end{cases}$$

$$x_2 \sim \begin{cases} \text{Normal}(-1, 1) & \text{if } \quad Y = 0,\ PV = 1 \quad \text{[a, majority; inlier]} \\ \text{Normal}(1, 1) & \text{if } \quad Y = 0,\ PV = 0 \quad \text{[b, minority; inlier]} \\ 2 \times \text{Exponential}(1)(1 - 2 \times \text{Bernoulli}(1/2)) & \text{if } \quad Y = 1 \quad \text{[outlier]} \end{cases}$$

`Synth2`

Simulate samples $X = [x_1, x_2]$ by...

$$PV \sim \text{Bernoulli}(4/5)$$

$$Y \sim \text{Bernoulli}(1/20)$$

$$x_1 \sim \begin{cases} \text{Normal}(180, 10) & \text{if } \quad PV = 1 \quad \text{[a, majority]} \\ \text{Normal}(150, 10) & \text{if } \quad PV = 0 \quad \text{[b, minority]} \end{cases}$$

$$x_2 \sim \begin{cases} \text{Normal}(10, 3) & \text{if } \quad Y = 1 \quad \text{[outlier]} \\ \text{Exponential}(1) & \text{if } \quad Y = 0 \quad \text{[inlier]} \end{cases}$$

Real-world

We experiment on 4 real-world datasets from diverse domains that have various types of PV: specifically gender, age, and race (see Table 3.1). Detailed descriptions are as follows.

- **Adult** (Lichman et al., 2013) (`Adult`). The dataset is extracted from the 1994 Census database where each data point represents a person. The dataset records income level of an individual along with features encoding personal information on education, profession, investment and family. In our experiments, *gender* $\in$ {*male, female*} is used as the protected variable where *female* represents minority group and high earning individuals who exceed an annual income of 50,000 i.e. annual *income* $> 50,000$ are assigned as outliers ($Y = 1$). We further downsample *female* to achieve a *male* to *female* sample size ratio of 4:1 and ensure that percentage of outliers remains the same (at 5%) across groups induced by the protected variable.

- **Credit-defaults** (Lichman et al., 2013) (`Credit`). This is a risk management dataset from the financial domain that is based on Taiwan's credit card clients' default cases. The data records information of credit card customers including their payment status, demographic factors, credit data, historical bill and payments. Customer *age* is used as the protected variable where *age* > 25 indicates the majority group and *age* ≤ 25 indicates the minority group. We assign individuals with delinquent *payment status* as outliers ($Y = 1$). The *age* > 25 to *age* ≤ 25 imbalance ratio is 4:1 and contains 5% outliers across groups induced by the protected variable.

- **Abusive Tweets** (Blodgett, Green, and O'Connor, 2016) (`Tweets`). The dataset is a collection of Tweets along with annotations indicating whether a tweet is abusive or not. The data are not annotated with any protected variable by default; therefore, to assign protected variable to each Tweet, we employ the following process: We predict the racial dialect — *African-American* or *Mainstream* — of the tweets in the corpus using the language model proposed by (Blodgett, Green, and O'Connor, 2016). The dialect is assigned to a Tweet only when the prediction probability is greater than 0.7, and then the predicted *racial dialect* is used as protected variable where *African-American dialect* represents the minority group. In this setting, abusive tweets are labeled as outliers ($Y = 1$) for the task of flagging abusive content on Twitter. The group sample size ratio of *racial dialect = African-American* to *racial dialect = Mainstream* is set to 4:1. We further sample data points to ensure equal percentage (5%) of outliers across dialect groups.

- **Internet ads** (Lichman et al., 2013) (`Ads`). This is a collection of possible advertisements on web-pages. The features characterize each ad by encoding phrases occurring in the ad URL, anchor text, alt text, and encoding geometry of the ad image. We assign observations with class label *ad* as outliers ($Y = 1$) and downsample the data to get an outlier rate of 5%. There exists no demographic information available, therefore we simulate a binary protected variable by randomly assigning each observation to one of two values (i.e. groups) $\in \{0, 1\}$ such that the group sample size ratio is 4:1.

3.4.2 Baselines

We compare FAIROD to two classes of baselines: (i) a fairness-agnostic base detector that aims to solely optimize for detection performance, and (ii) preprocessing methods that aim to correct for bias in the underlying distribution and generate a dataset obfuscating the PV.

Base detector model:

- BASE: A deep anomaly detector that employs an autoencoder neural network. The reconstruction error of the autoencoder is used as the anomaly score. BASE omits the protected variable from model training.

Preprocessing based methods:

- RW (Kamiran and Calders, 2012): A preprocessing approach that assigns weights to observations in each group differently to counterbalance the under-representation of minority samples.
- DIR (Feldman, Friedler, Moeller, Scheidegger, and Venkatasubramanian, 2015) A preprocessing approach that edits feature values such that protected variables can not be predicted based on other features in order to increase group fairness. It uses *repair_level* as a hyperparameter, where 0 indicates no repair, and the larger the value gets, the more obfuscation is enforced.
- LFR: This baseline is based on Zemel, Wu, Swersky, Pitassi, and Dwork, 2013 that aims to find a latent representation of the data while obfuscating information about protected variables. In our implementation, we omit the classification loss component during representation learning. It uses two hyperparameters – A_z to control for SP, and A_x to control for the quality of representation.
- ARL: This is based on Beutel, Chen, Zhao, and Chi, 2017 that finds new latent representations by employing an adversarial training process to remove information about the protected variables. In our implementation, we use reconstruction error in place of the classification loss. ARL uses λ to control for the trade-off between accuracy (in our implementation, reconstruction quality)

and obfuscating protected variable. This baseline optimizes an objective similar to that proposed in (Zhang and Davidson, 2020) which substitutes SVDD loss for reconstruction loss.

The OD task proceeds the preprocessing, where we employ the BASE detector on the modified data transformed or learned by each of the preprocessing based baselines. We do not compare to the LOF-based fair detector in (P and Abraham, 2020) as it exhibits disparate treatment and is inapplicable in settings that we consider.

Hyperparameters

We choose the hyperparameters of FAIROD from $\alpha \in \{0.01, 0.5, 0.9\} \times \gamma \in \{0.01, 0.1, 1.0\}$ by evaluating the Pareto curve for fairness and group fidelity criteria. The BASE and FAIROD methods both use an auto-encoder with two hidden layers. We fix the number of hidden nodes in each layer to 2 if $d \leq 100$, and 8 otherwise. The representation learning methods LFR and ARL use the model configurations as proposed by their authors. The hyperparameter grid for the preprocessing baselines are set as follows: $repair\_level \in \{0.0001, 0.001, 0.01, 0.1, 1.0\}$ for DIR, $A_z \in \{0.0001, 0.001, 0.01, 0.1, 0.9\}$ and $A_x = 1 - A_z$ for LFR, and $\lambda \in \{0.0001, 0.001, 0.01, 0.1, 0.9\}$ for ARL. We pick the best model for the preprocessing baselines using Fairness as they only optimize for statistical parity. The best BASE model is selected based on reconstruction error through cross validation upon multiple runs with random seeds.

3.4.3 Evaluation

We design experiments to answer the following questions:

- **[Q1] Fairness:** How well does FAIROD (a) achieve fairness as compared to the baselines, and (b) retain the within-group ranking from BASE?
- **[Q2] Fairness-accuracy trade-off:** How accurately are the outliers detected by FAIROD as compared to fairness-agnostic BASE detector?
- **[Q3] Ablation study:** How do different elements of FAIROD influence group fidelity and detector fairness?

Evaluation Measures

Fairness

Fairness is measured in terms of statistical parity. We use flag-rate ratio $r = \frac{P(O=1|PV=a)}{P(O=1|PV=b)}$ which measures the statistical fairness of a detector based on the predicted outcome where $P(O = 1|PV = a)$ is the flag-rate of the *majority* group and $P(O = 1|PV = b)$ is the flag-rate of the *minority* group. We define Fairness $= \min(r, 1/r) \in [0, 1]$. For a maximally fair detector, Fairness $= 1$ as $r = 1$.

GroupFidelity

We use the Harmonic Mean (HM) of per-group NDCG to measure how well the group ranking of BASE detector is preserved in the fairness-aware detectors. HM between two scalars p and q is defined as $1/(\frac{1}{p} + \frac{1}{q})$. We use HM to report GroupFidelity since it is (more) sensitive to lower values (than e.g. arithmetic mean); as such, it takes large values when *both* of its arguments have large values. We define GroupFidelity $= \text{HM}(NDCG_{PV=a}, NDCG_{PV=b})$, where

$$NDCG_{PV=a} = \sum_{i=1}^{|\mathcal{X}_{PV=a}|} \frac{2^{s^{\text{BASE}}(X_i)} - 1}{\log_2(1 + \sum_{k=1}^{|\mathcal{X}_{PV=a}|} \mathbb{1}(s(X_i) \leq s(X_k))) \cdot IDCG},$$

$|\mathcal{X}_{PV=a}|$ is the number of instances in group with $PV = a$, $\mathbb{1}(cond)$ is the indicator function that evaluates to 1 if *cond* is true and 0 otherwise, $s(X_i)$ is the predicted score of the fairness-aware detector, $s^{\text{BASE}}(X_i)$ is the outlier score from BASE detector and $IDCG = \sum_{j=1}^{|\mathcal{X}_{PV=a}|} \frac{2^{s^{\text{BASE}}(X_j)}-1}{\log_2(j+1)}$. GroupFidelity ≈ 1 indicates that group ranking from the BASE detector is well preserved.

Top-k Rank Agreement

We also measure how well the final ranking of the method aligns with the purely performance-driven BASE detector, as BASE optimizes only for reconstruction error. We compute top-k rank agreement as the Jaccard set similarity between the top-k observations as ranked by two methods. Let $\pi^{\text{BASE}}_{[1:k]}$ denote the top-k of the ranked list based on outlier scores $s^{\text{BASE}}(X_i)$'s, and $\pi^{detector}_{[1:k]}$ be the top-k of the ranked list for competing methods where $detector \in \{$RW, DIR, LFR, ARL, FAIROD $\}$. Then the measure is given as Top-k Rank Agreement $= |\pi^{\text{BASE}}_{[1:k]} \cap \pi^{detector}_{[1:k]}| / |\pi^{\text{BASE}}_{[1:k]} \cup \pi^{detector}_{[1:k]}|$

AUC-ratio and AP-ratio

Finally, we consider supervised parity measures based on ground-truth labels, defined as the ratio of ROC AUC and Average Precision (AP) performances across groups; AUC-ratio $= \text{AUC}_{PV=a}/\text{AUC}_{PV=b}$ and AP-ratio $= \text{AP}_{PV=a}/\text{AP}_{PV=b}$.

[Q1] Fairness

In Fig. 3.2 (presented in Introduction), FAIROD is compared against BASE, as well as all the preprocessing baselines across datasets. The methods are evaluated using the best configuration of each method on each dataset. The best hyperparameters for FAIROD are the ones for which GroupFidelity and Fairness[5] are closest to the "ideal" point as indicated in Fig. 3.2.

In Fig. 3.2 (left), the average of Fairness and GroupFidelity for each method over datasets is reported. FAIROD achieves 9× and 5× improvement in Fairness as compared to BASE method and the nearest competitor, respectively. For FAIROD, Fairness is very close to 1, while at the same time the group ranking from the BASE detector is well preserved where GroupFidelity also approaches 1. FAIROD dominates the baselines (see Fig. 3.2 (right)) as it is on the Pareto frontier of GroupFidelity and Fairness. Here, each point on the plot represents an evaluated dataset. Notice that FAIROD preserves the group ranking while achieving SP consistently across datasets.

Fig. 3.4 reports Top-k Rank Agreement (computed at top-5% of ranked lists) of each method evaluated across datasets. The agreement measures the degree of alignment of the ranked results by a method with the fairness-agnostic BASE detector. In Fig. 3.4 (left), as averaged over datasets, FAIROD achieves better rank agreement as compared to the competitors. In Fig. 3.4 (right), FAIROD approaches ideal statistical parity across datasets while achieving better rank agreement with the BASE detector. Note that FAIROD does not strive for a perfect Top-k Rank Agreement (=1) with BASE,

[5] Note that we can do model selection in this manner without access to any labels, since both are unsupervised measures.

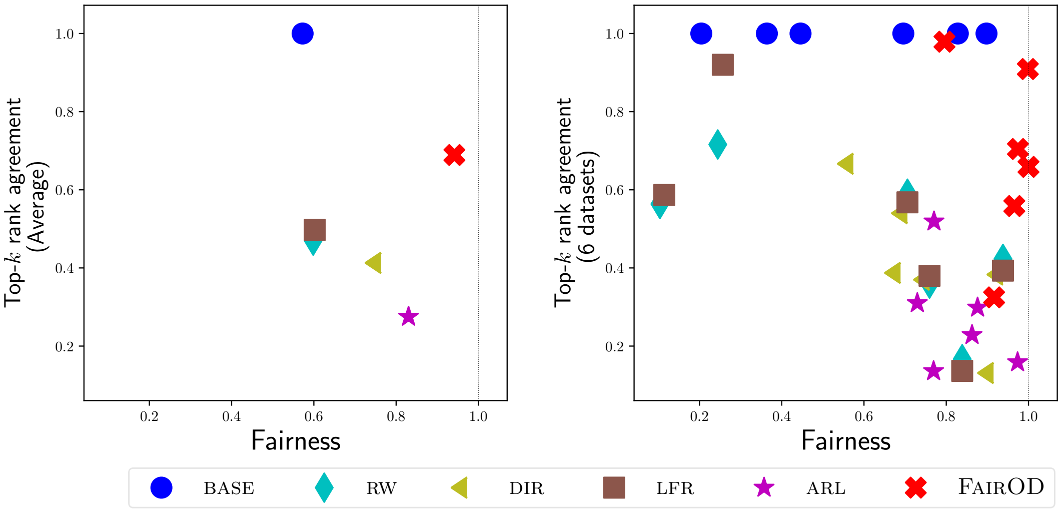


FIGURE 3.4: (left) FAIROD achieves the best Top-k Rank Agreement compared to the competitors (BASE is shown for reference) in addition to the best overall Fairness, across datasets on average, and (right) measures are shown on individual datasets.

since BASE is shown to fall short with respect to our desired fairness criteria. Our purpose in illustrating it is to show that the ranked list by FAIROD is not drastically different from BASE, which simply aims for detection performance.

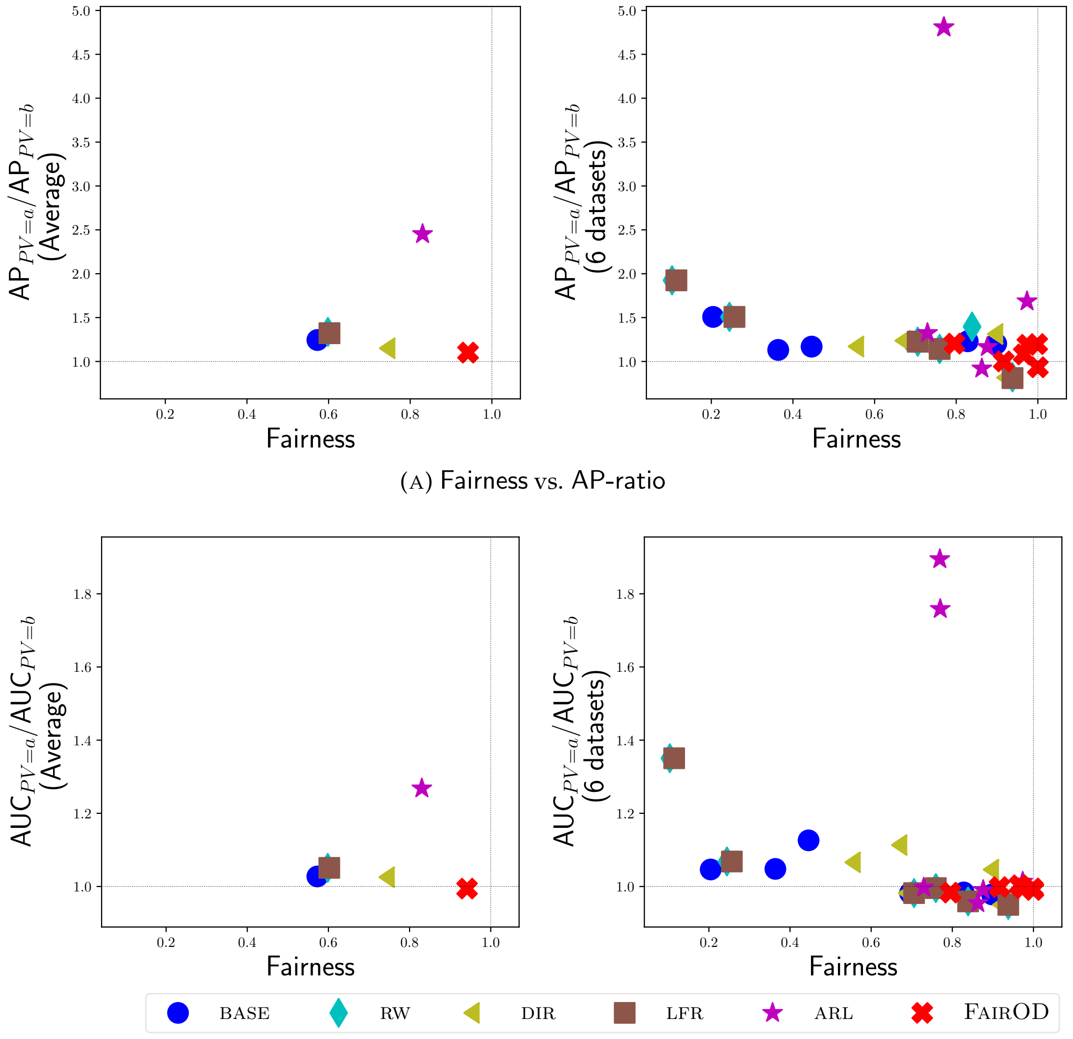
Next we evaluate the competing methods against supervised (label-aware) fairness metrics. Note that FAIROD does not (by design) optimize for these supervised fairness measures. Fig. 3.5a evaluates the methods against Fairness and label-aware parity criterion – specifically, group AP-ratio (ideal AP-ratio is 1). FAIROD approaches ideal Fairness as well as ideal AP-ratio across all datasets. FAIROD outperforms the competitors on the averaged metrics over datasets (Fig. 3.5a (left)) and across individual datasets (Fig. 3.5a (right)). In contrast, the preprocessing baselines are up to $\sim$5$\times$ worse than FAIROD over AP-ratio measure across datasets. Fig. 3.5b reports evaluation of methods against Fairness and another label-aware parity measure – specifically, group AUC-ratio (ideal AUC-ratio = 1). As shown in Fig. 3.5b (left), FAIROD outperforms all the baselines in expectation as averaged over all datasets. Further, in Fig. 3.5b (right), FAIROD consistently approaches ideal AUC-ratio across datasets, while the preprocessing baselines are up to $\sim$1.9$\times$ worse comparatively.

We note that impressively, FAIROD approaches parity across different supervised fairness measures despite not being able to optimize for label-aware criteria explicitly.

[Q2] Fairness-accuracy trade-off

In the presence of ground-truth outlier labels, the performance of a detector could be measured using a ranking accuracy metric such as area under the ROC curve (ROC AUC).

In Fig. 3.6, we compare the AUC performance of FAIROD to that of BASE detector for all datasets. Notice that each of the symbols (i.e. datasets) is slightly below the diagonal line indicating that FAIROD achieves equal or sometimes even better (!) detection performance as compared to BASE. The explanation is that since FAIROD enforces SP and does not allow "laziness", it addresses the issue of falsely or unjustly flagged minority samples by BASE, thereby, improving detection performance.



(A) Fairness vs. AP-ratio

(B) Fairness vs. AUC-ratio

FIGURE 3.5: FAIROD outperforms all competitors on averaged label-aware parity metrics over datasets (left) and for individual datasets (right): we report Fairness against (a) Group AP-ratio and (b) Group AUC-ratio.

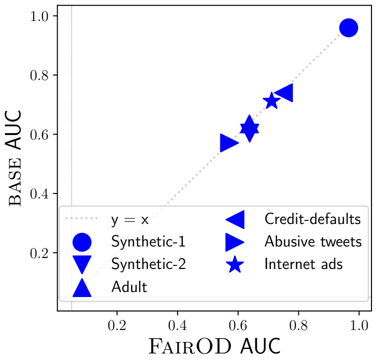


FIGURE 3.6: ROCAUC of FAIROD vs. BASE: FAIROD matches the performance of BASE detector, while enforcing fairness criteria (maintaining good performance *with* fairness).

From Fig. 3.6, we conclude that FAIROD does not trade-off detection performance much, and in some cases it even improves performance by eliminating false

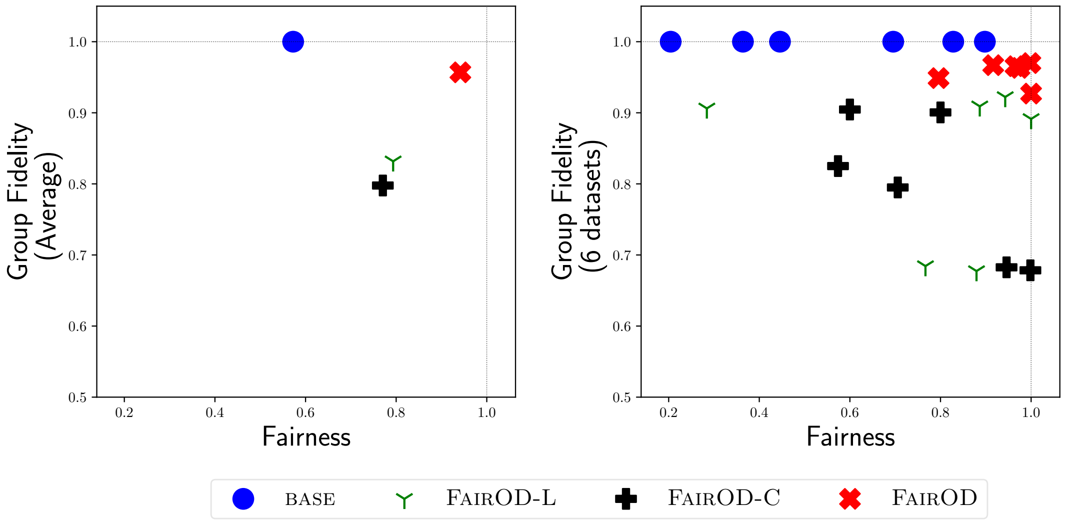


FIGURE 3.7: FAIROD compared to its variants FAIROD-L and FAIROD-C across datasets, to evaluate the effect of different regularization components. FAIROD-L achieves comparable Fairness to FAIROD while compromising GroupFidelity. FAIROD-C improves Fairness as compared to BASE, but is ill-suited to optimizing for GroupFidelity.

positives from the minority group, as compared to the performance-driven, fairness-agnostic BASE detector.

[Q3] Ablation study

Finally, we evaluate the effect of various components in the design of FAIROD's objective. Specifically, we compare to the results of two relaxed variants of FAIROD, namely FAIROD-L and FAIROD-C, described as follows.

- FAIROD-L: We retain only the SP-based regularization term from FAIROD objective along with the reconstruction error. This relaxation of FAIROD is partially based on the method proposed in Beutel et al., 2019, which minimizes the correlation between model prediction and group membership to the PV. In FAIROD-L, the reconstruction error term substitutes the classification loss used in the optimization criteria in Beutel et al., 2019. Note that FAIROD-L concerns itself with only group fairness to attain SP which may suffer from "laziness" (hence, FAIROD-L) (see Sec. 3.2).

- FAIROD-C: Instead of training with NDCG-based group fidelity regularization, FAIROD-C utilizes a simpler regularization, aiming to minimize the correlation (hence, FAIROD-C) of the outlier scores per-group with the corresponding scores from BASE detector. Thus, FAIROD-C attempts to maintain group fidelity over the entire ranking within a group, in contrast to FAIROD's NDCG-based regularization which emphasizes the quality of the ranking at the top. Specifically, FAIROD-C substitutes $\mathcal{L}_{GF}$ in Eq. (3.11) with the following.

$$\mathcal{L}_{GF} = - \sum_{v \in \{a,b\}} \left| \frac{\left(\sum_{X_i \in \mathcal{X}_{PV=v}} s(X_i) - \mu_s\right) \left(\sum_{X_i \in \mathcal{X}_{PV=v}} s^{\text{BASE}}(X_i) - \mu_{s^{\text{BASE}}}\right)}{\sigma_s \, \sigma_{s^{\text{BASE}}}} \right|$$

where $v \in \{a, b\}$, and $\mu_{s^{\text{BASE}}}, \sigma_{s^{\text{BASE}}}$ are defined similar to μ_s, σ_s respectively.

Fig. 3.7 presents the comparison of FAIROD and its variants. In Fig. 3.7 (left), we report the evaluation against GroupFidelity and Fairness averaged over datasets, and in Fig. 3.7 (right), the metrics are reported for each individual dataset. FAIROD-L approaches SP and achieves comparable Fairness to FAIROD except on one dataset as shown in Fig. 3.7 (right). This results in lower Fairness compared to FAIROD when averaged over datasets as shown in Fig. 3.7 (left). However, FAIROD-L suffers with respect to GroupFidelity as compared to FAIROD. This is because FAIROD-L may randomly flag instances to achieve SP since it does not include any group ranking criterion in its objective. On the other hand, FAIROD-C improves Fairness when compared to BASE, while under-performing on the majority of datasets compared to FAIROD across metrics. Since FAIROD-C tries to preserve group-level ranking, it trades-off on Fairness as measured against FAIROD-L. We also observe that FAIROD outperforms FAIROD-C on all datasets, which suggests that preserving the entire group-level rankings may be a harder task than preserving top of the rankings; it is also a needlessly ill-suited one since what matters for outlier detection is the top of the ranking.

3.5 Related Work

A majority of work on algorithmic fairness focuses on supervised learning problems. We refer to Barocas, Hardt, and Narayanan, 2019; Mehrabi, Morstatter, Saxena, Lerman, and Galstyan, 2019 for an excellent overview. We organize related work in three sub-areas related to fairness in outlier detection, fairness-aware representation learning, and data de-biasing strategies.

Outlier Detection and Fairness Outlier detection (OD) is a well-studied problem in the literature (Aggarwal, 2015; Gupta, Gao, Aggarwal, and Han, 2013; Chandola, Banerjee, and Kumar, 2009), and finds numerous applications in high-stakes domains like health-care (Luo and Gallagher, 2010), security (Gogoi, Bhattacharyya, Borah, and Kalita, 2011), and finance (Phua, Lee, Smith, and Gayler, 2010). However, only a few studies focus on OD's fairness aspects. P and Abraham, 2020 propose a detector called FairLOF that applies an ad-hoc procedure to introduce fairness specifically to the LOF algorithm (Breunig, Kriegel, Ng, and Sander, 2000). This approach suffers from several drawbacks: (i) it mandates disparate treatment, which may be at times infeasible/unlawful, e.g. in domains like housing or employment, (ii) only prioritizes SP, which as we discussed in Sec. 3.2, can permit "laziness," (iii) it is heuristic, and cannot be concretely optimized end-to-end. Concurrent to our work, Zhang and Davidson, 2020 introduce a deep SVDD based detector employing adversarial training to obfuscate protected group membership, similar to our ARL baseline. This approach also has issues: (i) it only considers SP, and (ii) it suffers from well-known instability due to adversarial training (Kodali, Abernethy, Hays, and Kira, 2017; Madras, Creager, Pitassi, and Zemel, 2018; Cevora, 2020). A related work by Davidson and Ravi, 2020 focuses on quantifying the fairness of an OD model's outcomes after detection, which thus has a different scope.

Fairness-aware Representation Learning Several works aim to map input samples to an embedding space, where the representations are indistinguishable across groups (Zemel, Wu, Swersky, Pitassi, and Dwork, 2013; Louizos, Swersky, Li, Welling, and Zemel, 2015). Most recently, adversarial training has been used to obfuscate PV association in representations while preserving accurate classification (Edwards and Storkey, 2015; Beutel, Chen, Zhao, and Chi, 2017; Madras, Creager, Pitassi, and Zemel, 2018; Adel, Valera, Ghahramani, and Weller, 2019; Zhang, Lemoine, and Mitchell, 2018). Most of these methods are supervised. Substituting classification or label-aware loss terms with unsupervised reconstruction loss can plausibly extend such methods to OD (by using masked representations as inputs to a detector). However, a common shortcoming is that statistical parity (SP) is employed as the primary fairness criterion in these methods, e.g. in fair principal component analysis (Olfat and Aswani, 2019) and fair variational autoencoder (Louizos, Swersky, Li, Welling, and Zemel, 2015). To summarize, fair representation learning techniques exhibit two key drawbacks for unsupervised OD: (i) they only employ SP, which may be prone to “laziness", and (ii) isolating embedding from detection makes embedding oblivious to the task itself, and therefore can yield poor detection performance (as shown in experiments in Sec. 6.5).

Strategies for Data De-Biasing Some of the popular de-biasing methods (Kamiran and Calders, 2012; Krasanakis, Spyromitros-Xioufis, Papadopoulos, and Kompatsiaris, 2018) draw from topics in learning with imbalanced data (He and Garcia, 2009) that employ under- or over-sampling or point-wise weighting of the instances based on the class label proportions to obtain balanced data. These methods apply preprocessing to the data in a manner that is agnostic to the subsequent or downstream task and consider only the fairness notion of SP, which is prone to “laziness.”

Chapter 4

Detecting and Ranking Generalized Anomalies

Chapter based on: Meng-Chieh Lee, Shubhranshu Shekhar, Christos Faloutsos, T Noah Hutson, and Leon Iasemidis (2021). "Gen2Out: Detecting and ranking generalized anomalies". In: *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, pp. 801–811

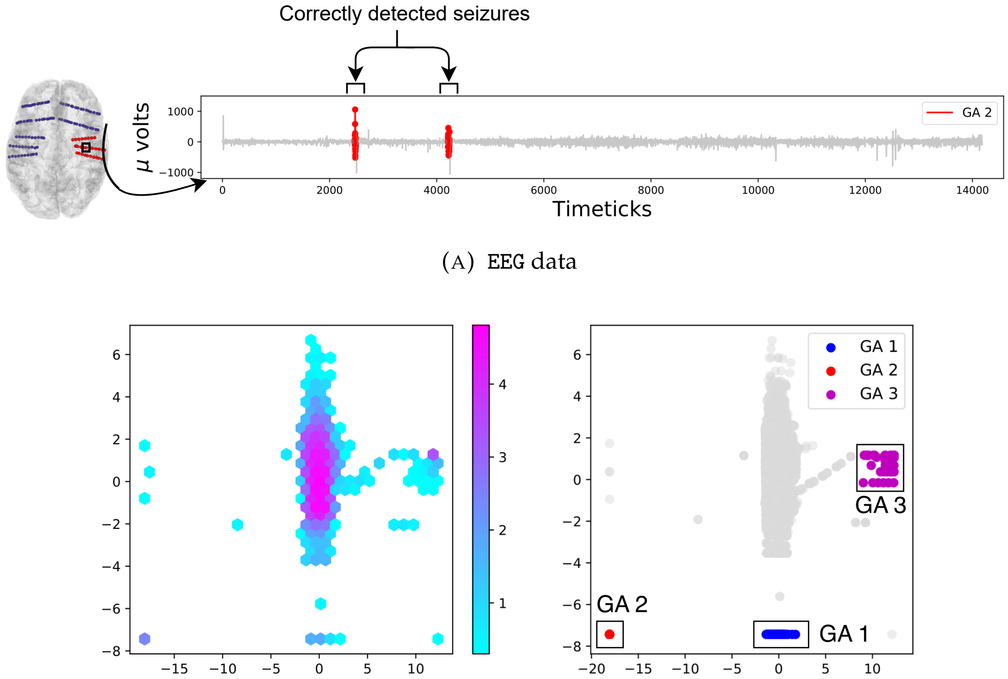
4.1 Introduction

How would we spot and rank single-point- as well as group-anomalies? How can we draw attention of the clinician to strange brain activities in multivariate EEG recordings of an epileptic patient? How could we design an anomaly score function, so that it assigns intuitive scores to both point-, as well as group-anomalies? Our goal is to design a principled and fast anomaly detection algorithm for a given cloud of m-dimensional point-cloud data that provides a unified view as well as a scoring function for each generalized anomaly. This has numerous applications (intrusion detection in computer networks, automobile traffic analysis, outlier[1] detection in a collection of feature vectors from, say, medical images, or twitter users, or DNA strings, and more).

Our motivating application is seizure detection in EEG recordings. Specifically, we want to spot those parts of the brain, and those time-ticks, that a seizure happened. Epilepsy is a neurodegenerative disease that affects $1-2\%$ of the world's population and is characterized by recurrent seizures that intermittently disrupt the normal function of the brain through paroxysmal electrical discharges (Shorvon, 2009). At least 30% of patients with medically refractory epilepsy are resistant to the mainstay treatment by anti-epileptic drugs (AEDs). These patients may benefit from surgical therapy. A significant challenge of this therapy is identification of the region of the brain where seizures are originating, that is, the epileptogenic focus (Krishnan et al., 2015; Vlachos, Krishnan, Treiman, Tsakalis, Kugiumtzis, and Iasemidis, 2016). This region is then surgically either resected or electrically stimulated over time to control upcoming seizures long prior to their occurrence (Tsakalis and Iasemidis, 2006; Chakravarthy, Sabesan, Tsakalis, and Iasemidis, 2009; Hutson, Pizarro, Pati, and Iasemidis, 2018). Accurate identification of the epileptogenic focus is therefore of high significance for the treatment of epilepsy.

As suggested by the application domain, to achieve better outcomes for patients, it is critical to direct attention of the clinician to the anomalous time periods in the

[1] We use outlier and anomaly interchangeably in this work.



(A) EEG data

(B) Heatmap of `http` data

(C) Detected group-anomalies

FIGURE 4.1: (a) GEN^{2}OUT matches ground truth. Brain scan of the patient with electrode positions (*top row*), and detected groups shown in color red, that matches the ground truth seizure locations. (b) Heatmap of *http* intrusion detection dataset (c) GEN^{2}OUT correctly spots group (DDoS) attacks in the intrusion detection dataset, marked GA1, GA2 and GA3.

brain activity in order of their suspiciousness. The problem is two-fold: (a) *detection*, as well as (b) *ranking* of generalized anomalies. We want a scoring function for generalized anomalies, such that in the EEG/epilepsy setting it would score the groups which may correspond to anomalous periods e.g. seizure and draw attention to most anomalous time periods; thus aiding a domain expert in decision making. As we show in Section 4.3.1, we propose some intuitive axioms, that a generalized anomaly detector should obey.

Informal Problem 1 (Doubly-general anomaly problem)**.**

- ***Given*** *a point-cloud dataset from an application setting,*
- ***find*** *the point-anomalies and group-anomalies, and*
- ***rank*** *them in suspiciousness order.*

Generality of approach: In most machine learning (ML) algorithms, we operate on clouds of points (after embedding, after auto-encoding etc). For example, time series is transformed into some form of $m-$dimensional cloud (Blázquez-García, Conde, Mori, and Lozano, 2021) for further analysis; in images, numerical features are generated for learning tasks e.g. Imagenet (Krizhevsky, Sutskever, and Hinton, 2012). Thus, the proposed approach can be applied to diverse real data including point cloud, time-series and image data.

Figure 6.1 illustrates the effectiveness of our method – GEN^{2}OUT detects group-anomalies that correspond to seizure period in the patient; and, detects DoS/DDoS attack as group-anomalies.

We propose GEN^{2}OUT, which has the following properties:

TABLE 4.1: GEN^{2}OUT matches all the specs. Qualitative comparison of GEN^{2}OUT against top competitors showing that every competitor misses one or more features.

Method Property	LODA (Pevný, 2016)	RRCF (Guha, Mishra, Roy, and Schrijvers, 2016)	IF (Liu, Ting, and Zhou, 2008b)	OCSMM(Muandet and Schölkopf, 2013)	AAE-VAE(Chalapathy, Toth, and Chawla, 2018)	MGM(Xiong, Póczos, Schneider, Connolly, and VanderPlas, 2011)	GLAD (Yu, He, and Liu, 2015)	GEN^{2}OUT
Obeys Axioms (see §4.3.1)				?	?	?	?	✓
Discover point anomalies	✓	✓	✓					✓
Rank point anomalies	✓	✓	✓					✓
Discover group anomalies							?	✓
Rank group anomalies				✓	✓	✓	✓	✓
Jointly rank point- and group- anomalies								✓
Scalable	✓	?	✓	?	✓	?	?	✓

- **Principled and Sound:** We identify five axioms (see Section §4.3.1) and show that the proposed GEN^{2}OUT obeys them, in contrast to top competitors.
- **Doubly-general:** GEN^{2}OUT is doubly general. First dimension of generalization is size of anomalies – detecting point- and group-anomalies. Second dimension of generalization is scoring and ranking of generalized anomalies– both point- and group-anomalies.
- **Scalable:** Linear on the input size (see Figure 4.10).
- **Effectiveness**: Applied on real-world data (see Figure 6.1 and 4.7), GEN^{2}OUT wins in most cases over benchmark datasets for point anomaly detection. For group anomaly detection, GEN^{2}OUT has no competitors as they need group structure information, and it agrees with ground truth on seizure detection.

4.2 Background and Related Work

Anomaly detection is a well-studied problem. Recent works (Boukerche, Zheng, and Alfandi, 2020; Chandola, Banerjee, and Kumar, 2009; Aggarwal, 2015; Gupta, Gao, Aggarwal, and Han, 2013; Toth and Chawla, 2018) provide a detailed review of many methods for anomaly and outlier detection. As shown in Table 4.1, GEN^{2}OUT is the only method that matches the specs. Here, we review anomaly detection methods for point- and -group- anomalies.

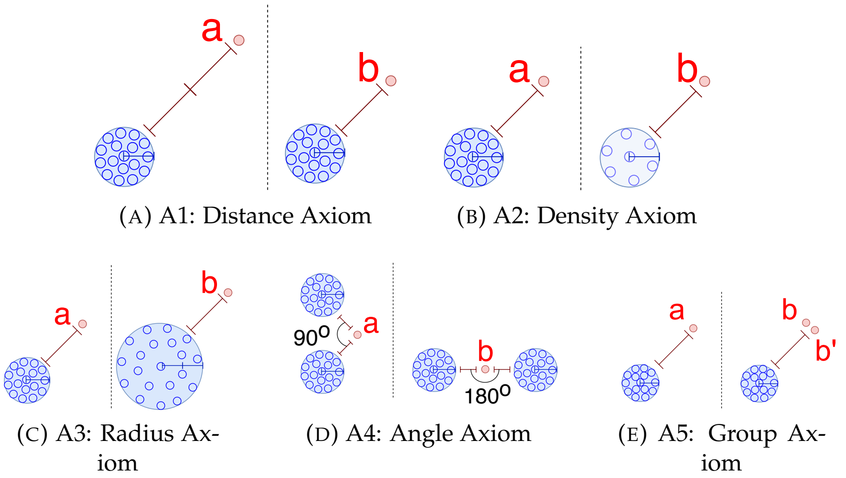
Point Anomaly Detection. Model-based and density-based methods for outlier detection are quite popular for point cloud data. Principal component analysis (PCA) based detectors (Shyu, 2003) assume that the data follows a multi-variate normal distribution. Local outlier factor (LOF) (Breunig, Kriegel, Ng, and Sander, 2000) flags instances that lie in low-density regions. Clustering based methods (He, Xu, and Deng, 2003) score instances or small clusters by their distance to large clusters. However, these methods suffer from too many false positives as they are not optimized for detection (Liu, Ting, and Zhou, 2012). Recently, a surge of focus has been on ensemble-based detectors that have been shown to outperform base detectors and are considered state-of-the-art for outlier detection (Emmott, Das, Dietterich, Fern, and Wong, 2013b). Isolation forest (Liu, Ting, and Zhou, 2008b) (IF), a state-of-the-art ensemble technique, builds a set of randomized trees that allows approximating the density of instances in a random feature subspace. Emmott, Das, Dietterich, Fern, and Wong, 2013b show that IF significantly outperforms other detectors such as LOF. IF (Liu, Ting, and Zhou, 2008b) shows that LOF has a high computation complexity (quadratic) and does not scale for large datasets. After that two more methods LODA (Pevnỳ, 2016) and Random Cut Forests (RRCF) (Guha, Mishra, Roy, and Schrijvers, 2016) have been proposed as state-of-the-art methods. LODA is projection-based histogram ensemble that works well in many real settings. RRCF improve upon IF and use a data sketch that preserves pairwise distances.

Group Anomaly Detection. Numerous methods have been proposed for group anomaly detection (Muandet and Schölkopf, 2013; Chalapathy, Toth, and Chawla, 2018; Xiong, Póczos, Schneider, Connolly, and VanderPlas, 2011; Yu, He, and Liu, 2015). Earlier approaches (Muandet and Schölkopf, 2013; Chalapathy, Toth, and Chawla, 2018; Xiong, Póczos, Schneider, Connolly, and VanderPlas, 2011) require the group memberships of the points known apriori, while Yu, He, and Liu, 2015 requires information on pairwise relations among data points. Moreover, these methods focus only on scoring group-anomalies, and ignore point-anomalies unlike our method. GEN^{2}OUT detects and ranks anomalous points and groups, without requiring additional information on group structure of the dataset. As mentioned above, Table 4.1 summarizes comparison of GEN^{2}OUT against state-of-the-art point- and group- anomaly detection methods. As such none of the methods has all the features of Table 4.1 .

Fractals and multifractals: In order to stress test our method, we use self-similar (fractals) clouds of points. We created the fractal images (Sierpinski triangle, biased line and 'fern' etc.), using the method and the code from Barnsley and Sloan, 1988. We used the 'uniform' version (that is, for the Sierpinski triangle, all the miniature versions have the same weight of 1/3), also generated the 'biased' version of triangle using weights (0.6, 0.3, 0.1), and 'biased line' with *bias* $b = 0.8$ using weights (0.8, 0.2) that is b of the data points go to the first half of the line, and in this half, b of the data points go to first quarter of the line, and so on recursively (this, informally, is the 80-20 law).

4.3 Proposed Axioms and Insights

In this section, we explain our proposed axioms in detail and give the insights. It is worth noting that these axioms are proposed to examine whether an anomaly detector is provided with the ability to compare the scores across datasets. The assumption is that, there are two different datasets with the same application setting. Although some of the axioms seem to be popular in single dataset setting, they are



(A) A1: Distance Axiom (B) A2: Density Axiom

(C) A3: Radius Axiom (D) A4: Angle Axiom (E) A5: Group Axiom

FIGURE 4.2: Illustration of Axioms

not considered and even ever mentioned by other studies when there is more than one dataset. The observed insights are critical and penetrate this research. These greatly inspire us on selecting the core part of our anomaly detector.

4.3.1 Proposed Axioms

We propose five axioms an ideal anomaly detector should follow: producing higher anomaly scores when an instance is farther away from data kernel (*d*istance aware), or lies in low density locality (*density, radius and group aware*), and not aligned with majority of data (*a*ngle aware (Kriegel, Schubert, and Zimek, 2008)). In the following, let $\boldsymbol{a} \in \mathrm{R}^m$ and $\boldsymbol{b} \in \mathrm{R}^m$ be $m-$dimensional anomalies in point cloud datasets S_a and S_b respectively. Additionally, suppose that normal observations are distributed uniformly in a disc in the datasets as shown in Figure 4.2 and $s(.)$ is the generalized anomaly score function.

Axiom 1 (Distance Aware). All else being equal, the farther point from the normal observations is more anomalous.

$$\left.\begin{array}{l} S_a - \{\boldsymbol{a}\} = S_b - \{\boldsymbol{b}\}, \\ dist(\boldsymbol{a}, S_a) > dist(\boldsymbol{b}, S_b) \end{array}\right\} \implies s(\boldsymbol{a}) > s(\boldsymbol{b})$$

Axiom 2 (Density Aware). All else being equal, denser the cluster of points, more anomalous the outlier.

$$\left.\begin{array}{l} dist(\boldsymbol{a}, S_a) = dist(\boldsymbol{b}, S_b), \\ density(S_a) > density(S_b) \end{array}\right\} \implies s(\boldsymbol{a}) > s(\boldsymbol{b})$$

Axiom 3 (Radius Aware). All else being equal, for a given number of observations, smaller the radius of the cluster of points, more anomalous the outlier.

$$\left.\begin{array}{l} |S_a| = |S_b|, \\ dist(\boldsymbol{a}, S_a) = dist(\boldsymbol{b}, S_b), \\ radius(S_a) < radius(S_b) \end{array}\right\} \implies s(\boldsymbol{a}) > s(\boldsymbol{b})$$

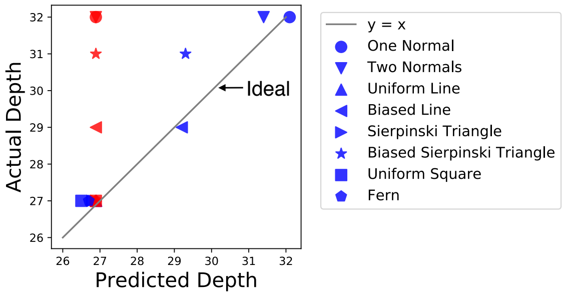


FIGURE 4.3: GEN^{2}OUT wins (in color blue) as the estimated depth is close to 45^o line. IF estimates the same depth for each dataset with #samples=1M.

Axiom 4 (Angle Aware). All else being equal, smaller the angle of a point with respect to cluster of observation, more anomalous the outlier.

$$\left.\begin{array}{l} |S_a| = |S_b|, \\ density(S_a) = density(S_b), \\ radius(S_a) = radius(S_b), \\ angle(\mathbf{a}, S_a) < angle(\mathbf{b}, S_b) \end{array}\right\} \implies s(\mathbf{a}) > s(\mathbf{b})$$

Axiom 5 (Group-size Aware). All else being equal, the least populous group, the more anomalous it is.

$$\textit{Let } g_a \subset S_a, g_b \subset S_b \textit{ are the groups.}$$

$$\left.\begin{array}{l} |g_a| < |g_b| \\ |S_a - g_a| = |S_b - g_b|, \\ density(S_a) = density(S_b), \\ radius(S_a) = radius(S_b), \end{array}\right\} \implies s(g_a) > s(g_b)$$

Justification for Axioms Axiom A1 is self explanatory as shown in Figure 4.2a. The outlier point (shown in color red) in the left dataset (Figure 4.2a) being farther from the normal observations should be more anomalous.

Consider the case of social networks. A node reachable via k hops from a close friends group should be more anomalous compared to reachable via k hops from a colleagues group. Figure 4.2b illustrates Axiom A2 where the outlier in the left dataset should be more anomalous.

As shown in Figure 4.2c, for the same number of observations, the larger radius cluster would have a larger distance among points. Therefore, the outlier in the left dataset with smaller radius should be more anomalous.

The farther points would tend to form a smaller angle with the cluster of observations (see Figure 4.2d) and should be more anomalous in the left dataset.

The group $g_a = \{\mathbf{a}\}$ consisting of one point Figure 4.2e is intuitively more anomalous compared to group $g_b = \{\mathbf{b}, \mathbf{b}'\}$ containing more data points. For example, if g_b has 1000 points, it is not an anomaly anymore.

4.3.2 Insights

In this section, we are given the observations $X = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ where $\mathbf{x_i} \in \mathrm{R}^m$ (see Table 4.2 for symbol definitions) for the anomaly detection. Our goal is to design an anomaly detector that obeys the axioms proposed in §4.3.1. The intuition for the selection of basic model is that, according to the five axioms in Figure 4.2, point 'a' in the first dataset should always have higher probability to be separated out

TABLE 4.2: Table of symbols.

Symbol	Definition
$X = \{\boldsymbol{x_i}\}$	point cloud dataset where $\boldsymbol{x_i} \in \mathrm{R}^m$ for $i \in 1, 2, \ldots n$
$s(.)$	anomaly score function for an outlier detector
$h(\boldsymbol{q})$	path length estimate for instance $\boldsymbol{q}$ as it traverses a depth limited ATOMICTREE
$\mathbb{E}[h(\boldsymbol{q})]$	path length averaged over the ensemble
$H(n)$	depth estimation function for an ATOMICTREE containing n observations
d_{limit}	depth limit of a ATOMICTREE

comparing the point 'b' in the second dataset. ATOMICTREE has the properties which are very close to our demand. Here, we consider a randomized tree ATOMICTREE data structure with the following properties – (i) Each node in the tree is either *leaf* node, or an internal node with two children, (ii) internal nodes store an attribute-value pair and dictate tree traversal. Given $X = \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}\}$, ATOMICTREE is grown through recursive division of X by randomly selecting an attribute and a split value until all the leaf nodes contain exactly one instance (hence the name ATOMICTREE) of observations assuming that observations are distinct. We randomly generate more than one tree to build a forest, to reduce the variance and detect outliers in subspaces.

We make a number of interesting observations while empirically investigating the process of tree growth for a variety of data distributions including multi-fractals. In Figure 4.4, we report depth (height) distribution of randomized trees averaged over 100 trees. We sample a number of points ($|X| \in \{2^{10}, 2^{11}, 2^{12}, 2^{20}\}; m = 2$) from each data distribution (shown in Figure 4.4 (a), (b), (c), (d), (i), (j), (k), (l)) and plot their corresponding depth (height) distribution (shown in Figure 4.4 (e), (f), (g), (h), (m), (n), (o), (p)). Notice that the number of points ($2^{\mathrm{x}}; \mathrm{x} \in \{6, 7, \ldots\}$) in the tree grows linearly with the average depth for any given dataset. In Figure 4.3, we plot the predicted depth for each of the distributions against the actual depth of the tree for those distributions shown in Figure 4.4 by fitting to this linear trend. We present the following lemma based on the observations and draw the following insights.

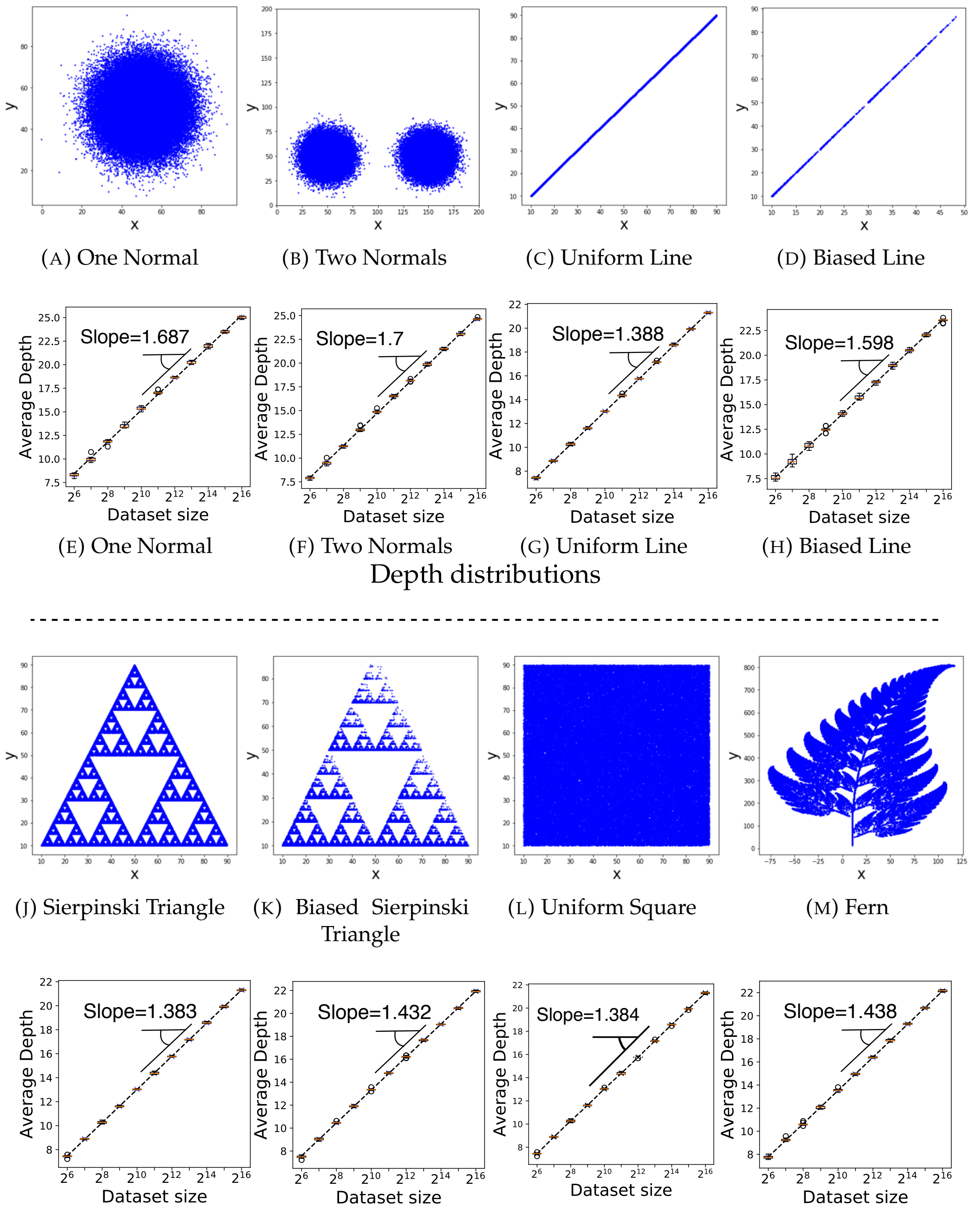
Insight 1 (Power Depth Property (PDP)). *The growth of the tree depth with the logarithm counts of observations is linear irrespective of the data distribution.*

Justification for PDP property: In our attempt to explain PDP property, we study the expected depth computation for datasets with known distributions. However, in general, it is difficult. Let us consider *biased* line dataset with a bias factor b. Here we study a related setting: random points, but with fixed cuts. We refer to this model as 'fixed-cut' tree FIXEDCUTTREE. For this case, we can show that the PDP property holds, and the slope grows as the 'bias' factor b grows. Then, the depth of FIXEDCUTTREE for a *biased* fractal line (data in Fig. 4.4d) obeys the following lemma.

Lemma 1 (Expected Depth of FIXEDCUTTREE). *The expected tree depth $H(n, b)$ for a* biased *line with a bias factor b containing $n \geq 2$ data points is given as:*

$$H(n,b) = \sum_{k=0}^{n} \Big[\binom{n}{k} b^k (1-b)^{n-k} \times$$

$$\big(\frac{k}{n} H(k,b) + \frac{n-k}{n} H(n-k,b) + 1\big)\Big]$$



(A) One Normal (B) Two Normals (C) Uniform Line (D) Biased Line

(E) One Normal (F) Two Normals (G) Uniform Line (H) Biased Line

Depth distributions

(J) Sierpinski Triangle (K) Biased Sierpinski Triangle (L) Uniform Square (M) Fern

(N) Sierpinski Triangle (O) Biased Sierpinski Triangle (P) Uniform Square (Q) Fern

Depth distributions

FIGURE 4.4: Illustrating depth distribution for several diverse datasets (including Gaussian, Uniform, multifractals).

Proof. Let $H(k, b)$ be the depth of FIXEDCUTTREE with k observations constructed using $X_k \subseteq X$. Since FIXEDCUTTREE is grown via recursive partitioning on a randomly chosen attribute-value, therefore, for a biased line, $b =$ probability of a point going to left node i.e. the point less than chosen attribute-value. Let k be the number of points partitioned onto the left node, then $n - k$ points go to right node. Define $\mathrm{B}(n, k, b) = \binom{n}{k} b^n (1 - b)^{n-k}$ the Binomial probability for a fixed k. Let $f(n, k, b)$ be the estimate of the depth when k observations are in left node, then $f(n, k, b) = (\frac{k}{n} H(k, b) +$

$\frac{n-k}{n}H(n-k,b)+1\big)$ as each random partition increases depth by 1. Therefore, the expected depth of the tree is given as $H(n,b) = \sum_{k=0}^{n} f(n,k,b)\mathrm{B}(n,k,b)$. ■

We denote $H(n,b) = H(n)$ for $b = 1/2$. A tree with one data point would have a depth of one i.e. $H(1,b) = 1 = H(1)$; and $H(0,b) = 0 = H(0)$. In Figure 4.5, we show the effect of bias on the the (analytical) depth computed using $H(n,b)$. Notice that increase in bias – indicating deviation from uniformity – increases depth which matches intuition.

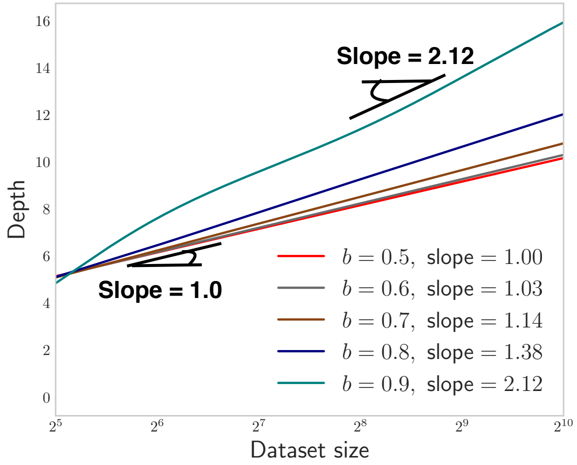


FIGURE 4.5: Depth ($\boldsymbol{H(n,b)}$) vs. Dataset size: slope increases with increase in bias for a *biased* line data

Corollary 1. *For bias* $1-b$, $H(n, 1-b)$ *follow the results for* $H(n,b)$.

Following the PDP property, the depth estimation function is given as

$$H(n) \approx w_0 + w_1 log_2(n) \tag{4.1}$$

where w_0 and w_1 are parameters that we estimate for each data distribution, and n is the number of instances in the dataset.

Insight 2. *The slope of the linear fit varies significantly depending on the dataset distribution.*

For example, the slope for Uniform Line (see Fig. 4.4g) is 1.38, while for a Uniform Square (see Fig. 4.4p) is 1.66. These insights lead to the following lemma.

Lemma 2. GEN^{2}OUT *includes* IF *as a special case.*

Proof. In Eq. 4.1, setting $w_0 = 2 \times 0.57 - (2(n-1)/n)$ and $w1 = 2 \times log_e(2)$ yields the average path length function used in IF. Here, 0.57 is the Euler's constant, and $log_e(2)$ accounts for the difference in log bases. ■

Drawing from these insights, next we present the details of our proposed anomaly detector algorithm.

4.4 Proposed Method

For ease of exposition, we describe the algorithm in two steps – GEN^{2}OUT$_0$ for point anomalies, and then GEN^{2}OUT for generalized anomalies.

TABLE 4.3: GEN^{2}OUT wins as it obeys all the axioms a generalized anomaly detector should follow. We compare the methods statistically, by conducting two-sample t-test based on scores obtained for points $\boldsymbol{a}, \boldsymbol{b}$. A positive difference in score indicates that the detector follows that axiom (see Figure 4.2). 🟩 indicates that the detector follows the axiom, 🟥 indicates that the detector does not obey the axiom.

	LODA		RRCF		IF		GEN^{2}OUT	
	Statistic	p-value	Statistic	p-value	Statistic	p-value	Statistic	p-value
A1: Distance Axiom	0	1	3.6	0.002**	2.1	0.054	11.4	1.2e-9***
A2: Density Axiom	7e15	2e-275***	-0.14	0.89	-10	8.6e-9***	25.2	1.7e-15***
A3: Radius Axiom	0	1	6.4	4.8e-6***	11.9	5.9e-10***	21.3	3.4e-14***
A4: Angle Axiom	6.6	3.2e-6***	17.5	9.6e-13***	-0.2	0.83	53.7	2.5e-21***
A5: Group Axiom	-14.7	1.8e-11***	1.1	0.27	0.95	0.35	28.2	2.6e-16***

Our goal is to design a principled, universal anomaly detector that allows for score comparison across point-cloud datasets, and we present the proposed GEN^{2}OUT. Motivated by our epilepsy application, we aim to apply it to draw attention to anomalous periods in a multivariate time series data, and to point to variables that characterize the flagged anomalous period. We are given $\mathbb{T} = \{\mathcal{T}_i\}, i = 1, 2, \ldots$ a set of multivariate time series. Each time series $\mathcal{T} \in \mathbb{T}$ consists of signal measurements $\boldsymbol{x_t} = (x_{t1}, \ldots, x_{tM})$ across M variables (electrode channels in seizure data) at each time step $t = 1, 2, ..., T$. To find anomalous patterns that occur in multivariate time series $\mathcal{T}$, we divide time-series data $\mathcal{T}$ into subsequence S_j according to the sliding window size K. We characterize each univariate series of length K with several features such that each univariate series becomes a m-dimensional point. In particular, each univariate series in the multivariate time series is summarized by a set of features extracted from its statistical properties such as first four moments, and each subsequence S_j is transformed into a point cloud data matrix $X_j \in \mathrm{R}^{M \times m}$. We then apply GEN^{2}OUT (details to follow next) to each data matrix X_j and record the anomalousness of each (electrode channel) variable $\in 1, \ldots, M$ for each period of time (window size K). The key property of GEN^{2}OUT is that it obeys the axioms that an ideal detector should follow enabling us to reason about the anomalousness of each time period and the anomalousness of each variable within a time period by comparing their respective anomaly scores. Therefore, each time period is now represented by an anomaly score vector through which we can draw attention of the domain experts to anomalous time periods.

4.4.1 Point anomalies – GEN^{2}OUT$_0$

Given the observations $X = \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_M}\}$ where $\boldsymbol{x_i} \in \mathrm{R}^m$, GEN^{2}OUT$_0$'s goal is to detect and assign anomaly score to outlier points. GEN^{2}OUT$_0$ uses an ensemble of depth-limited randomized tree ATOMICTREE (§4.3.2) that recursively partition instances in X.

Definition 1 (Depth Limited ATOMICTREE). *An* ATOMICTREE *that is constructed by recursively partitioning the given set of observations X until a depth limit d_{limit} is reached or the* leaf *nodes contain exactly one instance.*

As evidenced in prior works, the random trees induce shorter path lengths (number of steps from root node to leaf node while traversing the tree) for anomalous observations since the instances that deviate from other observations are likely to be partitioned early. Therefore, a shorter average path length from the ensemble would likely indicate an anomalous observation. Anomaly detection is essentially a ranking task where the rank of an instance indicates its relative degree of anomalousness.

```
Data: A data matrix $X$, number of ATOMICTREE estimators numTrees,
      ATOMICTREE depth limit $d_{limit}$
Result: $w_0, w_1$ of depth estimation function $H(\cdot)$ and ATOMICTREE ensemble
Initialize $Y$ and $Z$;
/* Estimating the function $H(\cdot)$                               */
for $i = n_1, n_1 + 1, \ldots$;                      // a small $n_1$ e.g. 10
 do
    Draw $X_s \subset X$ s.t. $|X_s| = 2^i$ ;
    $F_s \leftarrow$ CONSTRUCT-ATOMICTREE ($X_s$, $\infty$);
    $Z \leftarrow Z \cup$ average depth of $F_s$ containing observations $X_s$;
    $Y \leftarrow Y \cup i$;
end
$H(.) \leftarrow$ Fit linear regression $Y$ and $Z$;
$w_0, w_1 \leftarrow$ coefficients($H(.)$);
/* Construct the ensemble of ATOMICTREE                              */
for $t$ = 1 to numTrees do
    ensemble $\leftarrow$ ensemble $\cup$ CONSTRUCT-ATOMICTREE ($X$, $d_{limit}$);
end
return  $w_0, w_1$, ensemble
```

Algorithm 1: GEN^{2}OUT$_0$-FIT

We next design anomaly score function for our algorithm to facilitate ranking of observations.

```
Data: A data matrix $X$,$d_{limit}$, currDepth:0
Result: ATOMICTREE
Initialize ATOMICTREE;
if $d_{limit} \leq$ currDepth or $|X| \leq 1$ then
    return  a leaf node of size $|X|$
else
    pick an attribute at random from $X$;
    pick an attribute value at random;
    $X_l \leftarrow$ set of points on the left ($<$) of the chosen attribute-value pair;
    $X_r \leftarrow$ set of points on the right ($\geq$) of the chosen attribute-value pair;
    left $\leftarrow$ CONSTRUCT-ATOMICTREE ($X_l$,$d_{limit}$, currDepth + 1);
    right $\leftarrow$ CONSTRUCT-ATOMICTREE ($X_r$,$d_{limit}$, currDepth + 1)
    return  an internal node with {left, right, {chosen attribute-value
      pair}}
end
```

Algorithm 2: CONSTRUCT-ATOMICTREE

Proposed Anomaly Score. We construct anomaly score using the path length $h(\boldsymbol{q})$ for each instance $\boldsymbol{q} \in \mathrm{R}^m$ as it traverses a depth limited ATOMICTREE. The path length for $\boldsymbol{q}$ is $h(\boldsymbol{q}) = h_0 + H(l_{busy})$ if $l_{busy} > 1$; otherwise $h(\boldsymbol{q}) = h_0$ where h_0 is the number of edges $\boldsymbol{q}$ traverses from *root* node to *leaf* node that contains l_{busy} points in a depth limited ATOMICTREE. When $l_{busy} > 1$, we estimate the expected depth from the leaf node using $H(l_{busy})$ (uses Eq. 4.1). We normalize $h(\boldsymbol{q})$ by the average tree height $H(n)$ (height of ATOMICTREE containing n observations) for the depth limited ATOMICTREE ensemble to produce an anomaly score $s(\boldsymbol{q}, n)$ for a given observation $\boldsymbol{q}$. Referring to the PDP insights we presented in Section §4.3.2, we estimate the data dependent $H(\cdot)$ using Eq. 4.1 since the tree depth grows linearly with the number of observations (in log_2) in the tree (see Figure 4.4). The slope of the linearity is characterized by underlying data distribution; each distribution follows a linear growth. The score function is

$$s(\boldsymbol{q}, n) = 2^{-\frac{\mathrm{E}[h(\boldsymbol{q})]}{H(n)}} \tag{4.2}$$

where $\mathrm{E}[h(\boldsymbol{q})]$ is the average path length of observation $\boldsymbol{q}$ in the ATOMICTREE ensemble, n is number of data points used to construct each ATOMICTREE, and $H(n)$ is the function for estimating depth of the tree given in Eq. 4.1.

GEN^{2}OUT$_0$ **Parameter Fitting.** GEN^{2}OUT is a depth limited ATOMICTREE ensemble. The algorithm for fitting GEN^{2}OUT$_0$ parameters is provided in Algorithms 1 and 2.

```
Data: A data matrix $X$, ATOMICTREE ensemble
Result: Anomaly scores scores for observations in $X$
Initialize depths;
Initialize scores;
Initialize $l_{busy}$;
$n \leftarrow$ numSamplesInATOMICTREE;
for $\boldsymbol{x} \in X$ do
    depths $\leftarrow$ depths $\cup$ compute path-lengths for $\boldsymbol{x}$ (see §4.4.1);
    $l_{busy} \leftarrow l_{busy} \cup$ compute number of samples in leaf where traversal of $\boldsymbol{x}$
      terminated ;
end
for $depth \in$ depths, $l \in l_{busy}$ do
    $h$ =depth$+H(l)$;
    $s = 2^{\frac{-h}{H(n)}}$;
    scores $\leftarrow$ scores $\cup$ $s$;
end
return scores;
```

Algorithm 3: GEN^{2}OUT$_0$-Scoring

GEN^{2}OUT$_0$ **Scoring.** To assign anomaly scores to the instances in a data matrix X, the expected path length $\mathrm{E}(h(\boldsymbol{q}))$ for each instance $\boldsymbol{q} \in X$. $\mathrm{E}(h(\boldsymbol{q}))$ is estimated by averaging the path length after tree traversal through each ATOMICTREE in GEN^{2}OUT ensemble. We outline the steps to assign anomaly score to a data point using GEN^{2}OUT$_0$ in Algorithm 3.

4.4.2 Full algorithm – GEN^{2}OUT

GEN^{2}OUT$_0$ can spot point-anomalies. How can design an algorithm that can spot both point- as well as group-anomalies, simultaneously?

The main insight is to exploit the less-appreciated ability of sampling to drop outliers, with high probability. How can we use this property to spot group-anomalies, of size, say n_g (in a population of n data points)? The idea is that, with a sampling rate of n_g/n, a point $\boldsymbol{a}$ of the group will probably be stripped of its cohorts, and thus behave like a point-anomaly, exhibiting a high anomaly score. For dis-ambiguation versus the sampling of GEN^{2}OUT$_0$, we will refer to this sampling process as '*qualification*', and to the corresponding rate as *qr= qualification rate*.

```
Initialize $n \leftarrow |X|$;
/* Step 0: Fit a sequence of GEN$^2$OUT$_0$                                */
for $qr \in \{1, 1/2, 1/4, \cdots\}$ do
    Draw $X_s \subset X$ s.t. $|X_s| = n \times qr$;
    GEN$^2$OUT$_0$-ensembles $\leftarrow$ GEN$^2$OUT$_0$-ensembles $\cup$ GEN$^2$OUT$_0$-FIT ($X_s$, ., .);
end
/* Step 1: create X-RAY plot                                              */
for $e \in$ GEN$^2$OUT$_0$-ensembles do
    /* generate score for specific qualification rate                    */
    scores $\leftarrow$ scores $\cup$ GEN$^2$OUT$_0$-Scoring(X, e);
end
/* Step 2: Apex extraction                                                */
/* max score and qualification rate for each point across
   qualified datasets                                                     */
max_scores, max_qr $\leftarrow \arg\max$(scores) ;
/* select points with max score above threshold                           */
candidate-points $\leftarrow X$[max_scores $\geq$ threshold];
/* Step 3: Outlier grouping                                               */
for $r \in$ unique(max_qr) do
    candidate-points_r ; // candidate points at this qualification
     rate
    /* identify more than one group per qualification rate               */
    clusters $\leftarrow$ cluster candidate-points_r;
end
/* Step 4: Compute iso-curves                                             */
for $cl \in$ clusters do
    /* points similar to outlier at (score=1, qr=1) is more
       anomalous                                                          */
    iso_scores $\leftarrow \frac{2 - ManhattanDistance([\frac{\log_2 max\_qr(\boldsymbol{a})}{10}+1, max\_score(\boldsymbol{a})],[1,1])}{2}\ \forall \boldsymbol{a} \in cl$;
end
/* Step 5: Scoring                                                        */
assign scores $\leftarrow median($iso_scores$(cl))\ \forall cl \in$ clusters
```

Algorithm 4: GEN^{2}OUT

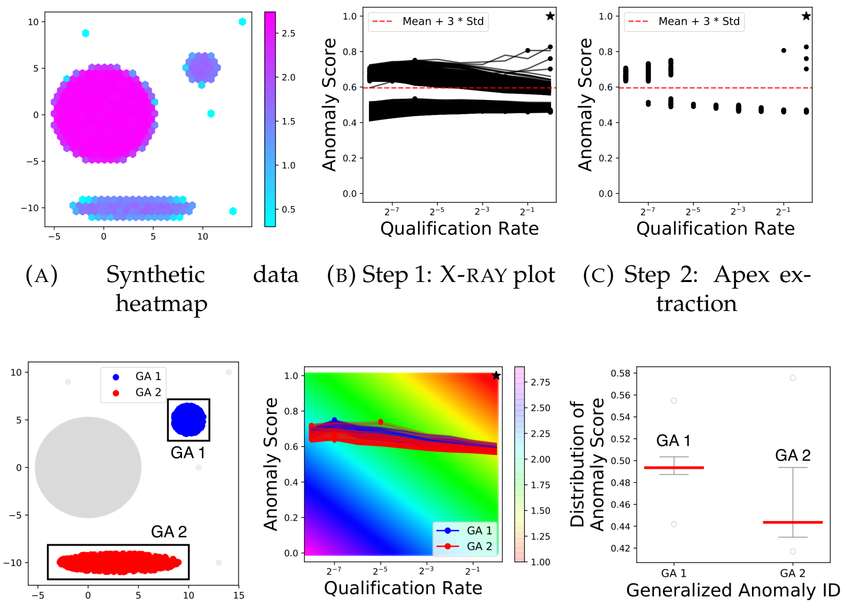
In more detail, to determine whether point $\boldsymbol{a}$ belongs to a group-anomaly, we compute its (GEN^{2}OUT$_0$) score $s(\boldsymbol{a}, qr)$ for several qualification rates qr; when the score peaks (say, at rate n_g/n) then n_g is roughly the size of the group-anomaly (= micro-cluster) that $\boldsymbol{a}$ belongs to. Some definitions:

Definition 2 (X-RAY-line). *For a given data point* $\boldsymbol{a}$*, the* X-RAY *line is the function (score(*$\boldsymbol{a}$*, qr) vs qr).*

Definition 3 (X-RAY plot). *For a cloud of n points, the* X-RAY *plot is the 2-d plot of all the n* X-RAY*-lines (one for each data point)*

See Figure 4.6b for an example.

Definition 4 (APEX). *Apex of point* $\boldsymbol{a}$ *is the point (score, qr) with the highest anomaly score.*



(A) Synthetic data heatmap (B) Step 1: X-RAY plot (C) Step 2: Apex extraction

(D) Step 3: Outlier grouping (E) Step 4: Anomaly iso-curves (F) Step 5: Scoring

FIGURE 4.6: GEN^{2}OUT works. Illustration of GEN^{2}OUT on synthetic dataset

See Figure 4.6c for an example.

Algorithm 4 describes the steps of the proposed GEN^{2}OUT. In summary, we find the X-RAY plot (Step 1) and then find the apex point for every data point $\boldsymbol{a}$ (Step 2); keep the ones with high apex and then cluster the corresponding data points (Step 3); and then assign scores to the each group (Step 4 and Step 5).

Figure 4.6 illustrates the steps in GEN^{2}OUT on a synthetic dataset that has two anomalous groups along with several point anomalies.

Figure 4.6b finds the X-RAY plot and Figure 4.6c shows the apex with the red threshold line. We find two groups after applying clustering (dbscan (Schubert, Sander, Ester, Kriegel, and Xu, 2017) in our implementation) shown in color red, and blue in Figure 4.6d. Then we compute the similarity of points in X-RAY plot representation in each cluster to the theoretically most anomalous point at score= 1, qr= 1 (see iso curves in Figure 4.6e), and then assign generalized anomaly score using the median of the similarity scores as shown in Figure 4.6f. GEN^{2}OUT correctly assigns higher score to GA1 (blue cluster in Figure 4.6f) which contains 1000 points as compared to GA2 (red cluster in Figure 4.6f) containing 2000 points (also see Axiom A5). For ease of visualization, we do not show point-anomalies in this plot.

4.5 Experiments

We evaluate our method through extensive experiments on a set of datasets from real world use-cases. We now provide dataset details and the experimental setup, followed by the experimental results.

4.5.1 Dataset Description

- **Epilepsy Dataset.** We analyzed intracranial electroencephalographic (EEG) signals recorded at the Epilepsy Monitoring Unit of a large public university from one

TABLE 4.4: Benchmark datasets summary.

Datasets	#Samples	Dimension	% Outliers
Size < 3000			
arrhythmia	452	274	14.6%
cardio	1831	21	9.6%
glass	214	9	4.2%
ionosphere	351	33	35.9%
letter	1600	32	6.3%
lympho	148	18	4.1%
pima	768	8	34.9%
vertebral	240	6	12.5%
vowels	1456	12	3.4%
wbc	378	30	5.6%
breastw	683	9	35%
wine	129	13	7.8%
Size ≥ 3000			
mnist	7603	100	9.2%
musk	3062	166	3.2%
optdigits	5216	64	2.9%
pendigits	6870	16	2.3%
satellite	6435	36	31.6%
satimage-2	5803	36	1.2%
shuttle	49097	9	7.2%
annthyroid	7200	6	7.4%
cover	286048	10	0.96%
http	567498	3	0.39%
mammography	11183	6	2.3%
smtp	95156	3	0.032%
speech	3686	400	1.7%
thyroid	3772	6	2.5%

patient with refractory epilepsy. Electrodes were stereotactically placed in the brain and EEG signals were then recorded across 122 electrode contacts at a sampling rate of 2KHz with focal region in the right temporal lobe.

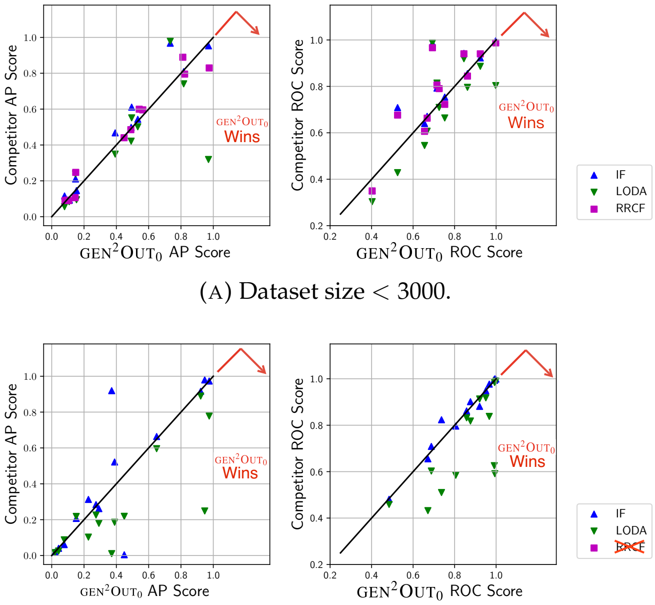
- **Benchmark Datasets.** Our benchmark set consist of 4 real-world outlier detection datasets from ODDS repository (Rayana, 2016). The datasets cover diverse application domains and have diverse range dimensionality and outlier percentage (summarized in Table 4.4). The ODDS datasets provide ground truth outliers that we use for the quantitative evaluation of the methods.

4.5.2 Point Anomalies

We compare GEN^{2}OUT$_0$ to the following state-of-the-art ensemble baselines.

1. IF: Isolation Forest (Liu, Ting, and Zhou, 2008b) uses an ensemble of randomized trees to flag anomalies.
2. LODA: Lightweight on-line detector of anomalies (Pevnỳ, 2016) is a projection based histogram ensemble.
3. RRCF: Robust Random Cut Forest (Guha, Mishra, Roy, and Schrijvers, 2016) are tree ensembles that use sketch based anomaly detector.

To evaluate the effectiveness, we compare GEN^{2}OUT$_0$ to state-of-the-art ensemble baselines on a set of real-world point-cloud benchmark outlier detection datasets. We use average precision (AP) and receiver operating characteristic (ROC) scores as our evaluation metrics. We plot the scores (AP and ROC score) for each competing method on all the benchmark datasets in Figure 4.7.



(A) Dataset size < 3000.

(B) Dataset size ≥ 3000 where RRCF doesn't scale.

FIGURE 4.7: GEN^{2}OUT$_0$ wins. We plot average precision (AP) and area under the ROC curve for GEN^{2}OUT$_0$ against the same metric of the competitors (none of which obey all our axioms). Points representing benchmark datasets are below the line for the majority of datasets.

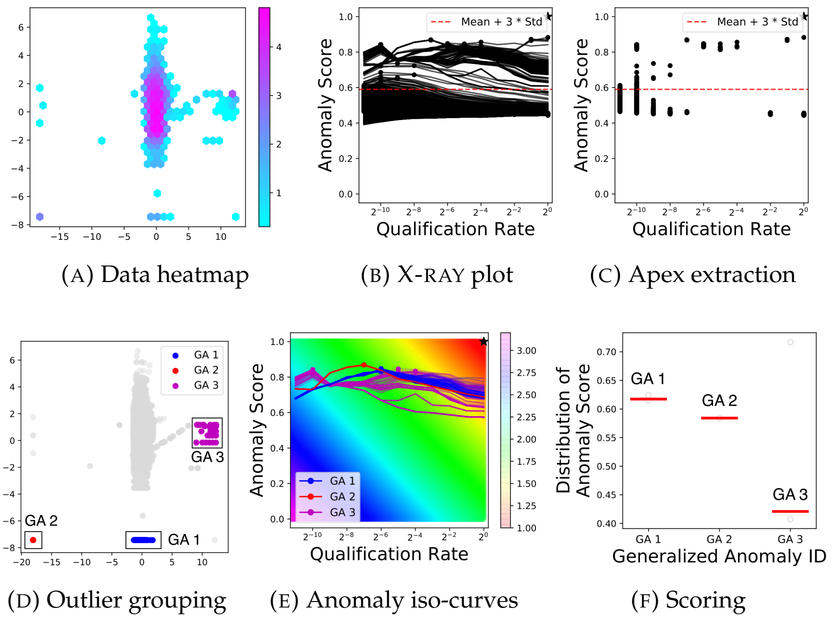
If the points are below the 45 degree line where each point represents a dataset listed in Table 4.4, then it indicates that GEN^{2}OUT$_0$ outperforms the competition in those datasets. As shown in Figure 4.7, for both the evaluation metrics, GEN^{2}OUT$_0$ beats or at least ties with all baselines on majority of the datasets (see Figure 4.1c). The quantitative evaluation demonstrates that GEN^{2}OUT$_0$ is superior to its competitors in terms of evaluation performance as well as obeys all the proposed axioms while none of the competition obeys the axioms.

4.5.3 Group Anomalies

We evaluate the effectiveness of GEN^{2}OUT on real-world intrusion dataset that has attributes describing duration of attack, source and destination bytes. Note that we do not include group anomaly detection methods for comparison as they require group structure information, hence do not apply to our setting. Figure 4.8a shows source bytes plotted against destination bytes for the points. Figures 4.8b – 4.8f shows the X-RAY plot with scores trajectory, APEX with candidate points above the threshold (set at mean + 3 standard deviation of scores in full dataset), identified groups and the generalized anomaly score for each detected group. GEN^{2}OUT matches ground truth as it detects the three anomalous groups as shown in Figure 4.8d. In short GEN^{2}OUT is able to detect groups that correspond to distributed-denial-of-service attack.

4.5.4 Scalability

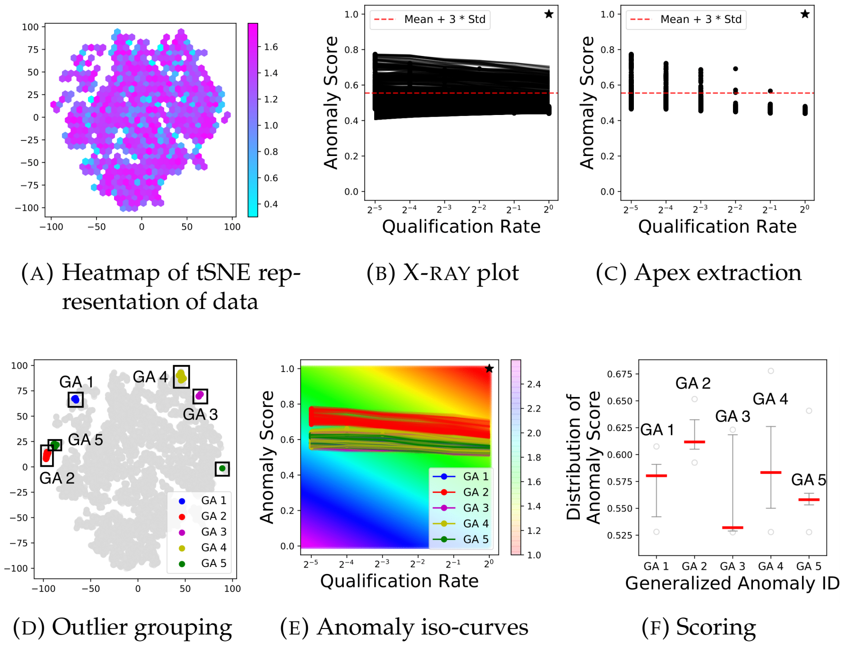
To quantify the scalability, we empirically vary the number of observations in the chosen dataset and plot against the wall-clock running time (on 3.2 GHz 36 core CPU with 256 GB RAM) for the methods. First we compare GEN^{2}OUT$_0$ against the competitors in Figure 4.10a for point-anomalies. The running time curve of GEN^{2}OUT$_0$



(A) Data heatmap (B) X-RAY plot (C) Apex extraction

(D) Outlier grouping (E) Anomaly iso-curves (F) Scoring

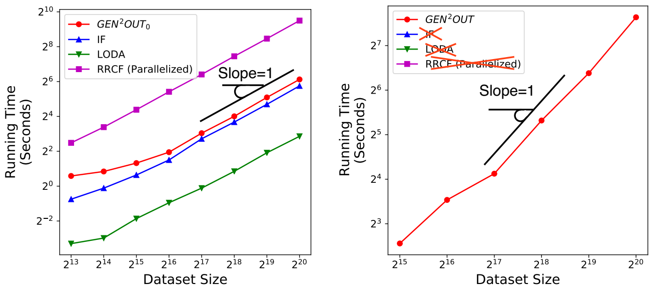
FIGURE 4.8: GEN^{2}OUT detects DDoS attacks on intrusion detection `http` dataset



(A) Heatmap of tSNE representation of data (B) X-RAY plot (C) Apex extraction

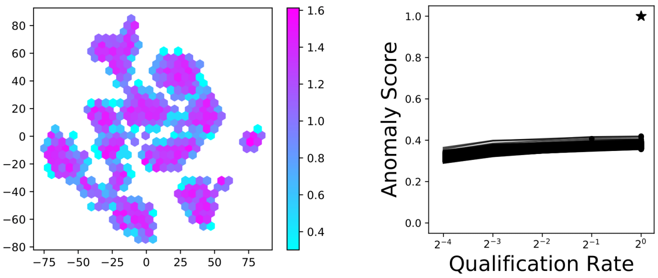
(D) Outlier grouping (E) Anomaly iso-curves (F) Scoring

FIGURE 4.9: GEN^{2}OUT works on real-world EEG data. Assigns highest anomaly score to group anomaly GA2 that corresponds to seizures as we show in Figure 4.1a.



(A) (B)

FIGURE 4.10: (a) GEN^{2}OUT$_0$ is fast and scalable: Evaluation on benchmark datasets show that GEN^{2}OUT (in red) scales linearly (eventual slope=1 in log-log scales). Note that none of the competitors obeys the axioms, and RRCF is significantly slower. (b) GEN^{2}OUT is fast and scalable, linear in size of input.



(A) Data heatmap (B) X-Ray plot

FIGURE 4.11: GEN^{2}OUT works. It correctly flags no anomalies in the `optdigits` dataset

is parallel to the running time curve of IF, which shows that GEN^{2}OUT$_0$ does not increase time complexity except adding a small constant overhead for estimating the depth function $H(.)$. The running time of RRCF is much higher than others even after implementing the trees in parallel. Note that only GEN^{2}OUT$_0$ obeys the axioms. For generalized anomalies, Figure 4.10b reports the wall-clock running time of GEN^{2}OUT as we vary the data size. Notice that GEN^{2}OUT scales linearly with input size. Importantly, competitors do not apply as they require additional information.

4.6 GEN^{2}OUT at Work

4.6.1 No False Alarms.

When applied to datasets containing only normal groups that are relatively equal in size, GEN^{2}OUT correctly identifies them as normal groups i.e. does not flag any set of points as anomalous group. To illustrate this phenomenon, we apply GEN^{2}OUT to `optdigits` dataset which contains the feature representation of numerical digits.

To better visualize the dataset, we embed the points in two dimensional space using tSNE (Maaten and Hinton, 2008) as shown in Figure 4.11a. It is a balanced dataset, where we have equal number of points for each digit, hence no group is present. X-RAY plot (Figure 4.11b) shows that all the score trajectories are below 0.5 (scores close to 1 are anomalous) with mean score at 0.36 in full dataset. Hence, we do not find group and correctly so.

4.6.2 Attention Routing in Medicine.

We apply GEN^{2}OUT on EEG recordings for the epileptic patient (PT1) – PT1 suffered through onset of two seizures in our recording clips; our motivating application. We extract four simple statistical measures from the subsequences of the time series features, namely mean, variance, skewness and kurtosis, by sliding a thirty minute window with two minutes overlap. Figure 4.9a shows $2-$dimensional tSNE representation of the data.

We then compute the generalized anomaly scores over time (within each window) for each detected group. Since the scores generated by GEN^{2}OUT are comparable, we draw attention to the most anomalous time point, where the seizures occurred as the detected groups correspond to seizure time period. The steps of GEN^{2}OUT are illustrated in Figure 4.9 when applied to multi-variate EEG data. Note that we find, several groups as shown in Figure 4.9. Of the detected groups, the group receiving highest score (GA2) is plotted over the raw voltage recordings over time for the patient. The group corresponds to the ground truth seizure duration (see Figure 4.1a). These time points that we direct attention to would assist the domain expert (in this case a clinician) in decision making by alleviating cognitive load of examining all time points.

Part II

Applications

Chapter 5

Public Health Care Fraud Detection

Chapter based on: Shubhranshu Shekhar, Jetson Leder-Luis, and Leman Akoglu (2023). *Unsupervised Machine Learning for Explainable Health Care Fraud Detection*. Tech. rep. National Bureau of Economic Research

5.1 Introduction

Fraud in health care is hard to detect. Insurers face information asymmetries, where physicians and patients both know more about the health care delivered than the insurer responsible for paying for that care. Health care providers such as doctors and hospitals face incentives to maximize their reimbursements from health insurance companies, and insurers must largely rely on documentation from providers themselves. This asymmetric information leads to circumstances where unscrupulous providers can choose to commit fraud.

The scale of health care spending means that even small amounts of fraud can be very expensive. Estimated US health care spending in 2019 was $3.81 Trillion (NHE Fact Sheet, 2021), almost as high as the GDP of Germany, the 4th largest in the world. National health care spending in the US is expected to grow at an average annual rate of 5.4%[1], from 2019 to 2028, outpacing US GDP at 4.3%. Efforts to detect and root out fraud are paramount for limiting the growth of wasteful spending.

These issues are compounded in the federal health care programs, where the government is the insurer. The US federal government spends over a trillion dollars per year on health insurance, largely paid to private firms, and fraud detection is challenging due to the sheer volume of claims being processed. The largest of these programs is Medicare, the federal health insurance program for people of age 65 and older and the disabled. With more than $800 Billion spent on Medicare in 2019, even small shares of waste and abuse lead to large losses, which are ultimately paid for by taxpayers and reduce the capacity of the government to fund valuable social programs. The US Government Accountability Office (GAO) estimates Medicare improper payments, a measure of mistaken or inappropriately documented spending, in 2019 at $46.2 Billion (U.S. Government Accountability Office, 2020). This problem has gained the attention of Medicare administrators faced with the challenge of detecting and deterring waste and fraud to ensure the program stays financially solvent (U.S. Department of Health and Human Services, 2022).

The nature of health care fraud provides insights into how it can be detected. Healthcare providers face incentives to manipulate billing to increase profits. Yet, in general, patients see multiple providers, and there are many providers in the system

[1] The growth was a striking 36.0% in 2020 in response to the COVID-19 pandemic.

that do not commit fraud. Therefore, fraud detection does not rely on the verification of any particular claim, but rather detecting provider-level patterns of care that appear anomalous when considering patient characteristics, medical history, and patterns of behavior by regular non-fraudulent providers.

In this work, we develop new tools to detect health care overbilling or fraud. We build a machine learning (ML) framework to discover patterns and detect anomalous providers using large-scale Medicare claims data. Our method focuses on inpatient hospitalization, the largest category of spending and the highest-intensity health care provided by Medicare, which cost the US government $147 Billion in 2021. The proposed approach identifies anomalous providers based on their billing patterns, using patient-level data including medical history, demographics, and geography. We employ our method to identify anomalous patterns among providers and rank them in order of their suspiciousness *without using any supervision*, that is, not relying on any *a priori* labeled training data. Moreover, our approach is equipped with *explanations* to the suspiciousness of the flagged providers, enabling end users like auditors to use our results to guide further investigation.

Our approach is an ensemble method, utilizing three novel unsupervised detection algorithms that uncover aberrant patterns in care across different data modalities. The first component of the ensemble focuses on providers[2] with large observed expenditures conditioned on patient characteristics and medical history. We use a regression-based analysis to identify providers with large fixed effects that correspond to high spending per patient even controlling for observable medical history and location of the patient. The second component focuses on coding behavior of claims, uncovering rare ICD-10 medical coding patterns employed by providers, which is indicative of manipulation of specific codes a patient is tagged with in order to garner higher reimbursements. The third component is peer based, focusing on identifying aberrant hospital billing code (DRG) patterns among a related group of hospitals, where the group of hospitals share similar patient populations and distributions of types of care.

We assemble the evidence from these three detection methods together to rank providers based on suspiciousness. We utilize instant-runoff voting (Franceschini, Maisano, and Mastrogiacomo, 2022) to reach an aggregate ranking for the suspiciousness of providers. This method follows an iterative procedure to rank the hospital that is most suspicious based on the "vote" across different detectors in each round.

We validate our approach quantitatively with ground-truth data from the Department of Justice (DOJ). Using a corpus of thousands of DOJ press releases about fraud, we tag providers identified as fraudulent and merge these data to compare with our ranking. While only 1 in 20 hospitals nationwide are named in the DOJ Press releases, our ranking substantially improves detection over random sampling: the top 50 providers identified by our method contain 21 providers named in the same DOJ corpus, that an 8-fold lift in detection rate. We note that providers ranked high by our method but not listed by the DOJ are not necessarily false positives; rather, enforcement by the DOJ reflects a combination of opportunity to enforce and capacity constraints, and hence only provides partial ground-truth. The DOJ validation is a form of positive-unlabeled data (Bekker and Davis, 2020), and the overlap with our method is therefore a lower-bound of the amount of fraud successfully detected.

[2] In this work, the words provider and hospital are used interchangeably. While providers can refer to any health care service provider, we specifically study hospitals.

In summary, our proposed approach provides scalable and explainable tools to detecting fraud and abuse in health care systems. Our method does not rely on any supervision or data labeling labor, and thus can be readily employed on massive unlabeled data. As the detectors utilize different data modalities and modeling approaches, the explanations also provide different perspective and reasoning into suspicious behavior. This makes our proposed approach useful in practice, as auditors would be presented with multiple pieces of evidence that support a case and can aid with further investigation. While our analysis focuses on hospitals, this method could be readily adapted for use in detecting overbilling in outpatient claims, doctor's office visits, or other areas of potentially fraudulent care.

We foresee that our method could be particularly effective at auditing of health care providers and guiding future enforcement. While the data set on which we build our method is from Medicare, we anticipate our methods will prove valuable to private insurers as well, who face nearly identical challenges in eliminating fraud from private health insurance systems. As our ranking provides a significant lift in detection rate than one would achieve by random sampling, it can be used to target and prioritize auditing. While our explanations cannot provide legal-standard evidence of bad behavior by providers, they can help sense-making and be used as starting points that guide deeper investigation. Overall, we anticipate that our proposed solution will have value for policymakers, auditors, and enforcers in the health care domain at large.

This chapter proceeds as follows. We describe the background and institutions regarding Medicare payment, health care fraud, and fraud enforcement in Section 6.3, followed by a description of Medicare data in Section 5.3. Then, we present our detection and explanation methodologies, with an overview in Section 6.4. Section 5.5 presents the global expenditure regression-based OD model; Section 5.6 presents the local ICD subspace based OD model; and Section 5.7 presents the local/contextual peer-based excess cost OD model. Section 5.8 reports the ensemble model detection results and multi-view explanations on several case studies. Finally, Section 5.9 provides a post-analysis toward characterizing hospitals with high estimated suspiciousness. We conclude with discussion and takeaways in Section 5.10.

5.2 Background

In this section, we discuss the institutional details of Medicare fraud. First, we describe the Medicare payment system for inpatient hospitalization, which creates incentives for fraud. Second, we discuss the various types of Medicare fraud and the ways in which it is enforced. While many of the institutional details about Medicare claims and enforcement are specific to the federal system, the general nature of health care billing is consistent across both publicly funded and private-payer systems.

5.2.1 Medicare Payment System

Medicare uses a prospective payment system (PPS) for inpatient hospitalization, where providers are paid a fixed amount for each patient's stay, regardless of stay length or cost. Patients are coded with diagnoses and procedure codes based on the International Classification of Diseases (ICD) system, and then based on this coding, each inpatient stay is classified into one Medicare Severity Diagnosis Related Group (DRG). Each DRG is associated with a certain fixed amount per stay, with possible small adjustments (Medpac, 2021). The fixed payment for each DRG is based on the average costs of treating patients under that DRG code nationwide and it is updated annually.

The PPS incentivizes providers to keep the healthcare costs down (Ellis and McGuire, 1986) since the provider's profit is the difference between the fixed DRG payment and the treatment cost. This is in contrast to a reimbursement-based system, where providers would face incentive to incur higher costs for higher reimbursement. However, the PPS may lead to hospitals trying to avoid treating high-cost patients. To address such issues, PPS adjusts the DRG payment (Medpac, 2021) to include provider specific factors such as provider's wage index (geographic factor), patient case-mix to account for patient-population specific treatment cost, teaching and research expenditure, disproportionate share of low-income patients, and number of unusually costly outlier cases.

5.2.2 Health Care Fraud

Hospitals face incentives to miscode patients; when done intentionally or recklessly, this can qualify as fraud. Because the patient's ICD coding dictates their DRG and ultimately the hospital reimbursement amount, hospital coding decisions directly affect hospital profits.

Fraud in inpatient hospitalization takes many forms. One well-studied form is *upcoding*, where hospitals miscode patients to higher severity levels of care in order to receive higher reimbursement (Dafny, 2005; Silverman and Skinner, 2004; Becker, Kessler, and McClellan, 2005). A second common issue is *lack of medical necessity*, where a patient's health conditions do not qualify them for that care (Howard, 2020). Moreover, there is a variety of conduct that can also qualify as health care fraud, such as providing compensation to providers for referring patients, which qualifies as a *kickback*.

In this work, we are largely agnostic to which type of fraud hospitals commit, and instead focus on payment levels. In general, fraud is of greatest concern when it results in wasteful spending. Our method detects hospitals whose anomalous conduct results in higher payments, which is valuable for detecting hospitals where additional auditing is of highest marginal value.

5.2.3 Health Care Anti-Fraud Enforcement

The US government undertakes a number of initiatives to detect and deter waste, fraud and abuse in federally-funded health care spending. Our method, which relies solely on claims data, is complementary to existing methodologies. Private insurers face similar challenges and also work to detect, investigate and enforce against fraudulent providers, although they lack the full weight of the federal investigatory system.

Federal law prohibits Medicare fraud and provides avenues by which fraud can be addressed through criminal and civil enforcement. The federal health care fraud statute provides criminal penalties for those who commit health care fraud, and this enforcement is compounded by criminal enforcement under the anti-kickback statute, as well as the wire fraud and racketeering statutes. Criminal Medicare fraud is prosecuted by the Department of Justice. For a deeper treatment of criminal Medicare fraud, see Eliason, League, Leder-Luis, McDevitt, and Roberts, 2021.

Civil enforcement for Medicare fraud operates through the False Claims Act, which provides an avenue for whistleblowers to come forward with information about fraud and receive compensation. Whistleblowers file their own cases in federal civil court, and the DOJ has an option to support these cases. Leder-Luis, 2020 and Howard, 2020 provide more information about the False Claims Act and show that these whistleblowers provide high deterrence effects.

In addition to litigation, administrators use a variety of policy tools to limit health care waste, fraud and abuse. The Office of the Inspector General of Health and Human Services undertakes administrative actions against firms that overbill Medicare. Medicare also has a variety of auditing programs that seek to detect unnecessary or unjustified spending; see Shi, 2022 for a description of the Recovery Audit Contractors program. Finally, Medicare uses regulations to target unnecessary spending, such as prior authorization requirements. Some of these regulations combat fraud while others combat waste; see Brot-Goldberg, Burn, Layton, and Vabson, 2022 and Eliason, League, Leder-Luis, McDevitt, and Roberts, 2021 for a discussion of these regulations.

In addition to the enforcement actions listed above, Medicare and private insurers undertake some data-driven investigatory work in order to detect fraud. These efforts have received little attention in academic work. Medicare claims processors work with contractors called Unified Program Integrity Coordinators (UPICs) (Noridian Healthcare Solutions, 2022) to audit and detect aberrant payments. In addition, Medicare uses a private-public partnership model through the Healthcare Fraud Prevention Partnership to share data between the federal government and private insurers to detect health care fraud with patterns similar across a variety of types of care and different health insurance programs (Healthcare Fraud Prevention Partnership, 2022). When fraud is identified through these data-driven efforts, investigators can refer those cases to the DOJ for civil or criminal prosecution.

We curate a list of hospitals that have been subject to DOJ actions at both the criminal and civil level, used for quantitative evaluation of our method. While there are many ways in which hospitals could have been investigated or sanctioned, being named in a DOJ press release validates that the hospital was likely committing behavior that rose to the level of criminal or civil fraud, which represents a true positive. A disclaimer, on the other hand, is that the hospitals subjected to DOJ actions likely constitute only a partial list of all fraudulent hospitals, as other unknown fraud and waste may have gone undetected, which represents a false negative.

5.2.4 Related Methodological Work

In addition to the economic studies listed above that discuss health care fraud, several data-centric approaches have been explored in the context of Medicare fraud. We refer the reader to Bauder, Khoshgoftaar, and Seliya, 2017; Kumaraswamy, Markey, Ekin, Barner, and Rascati, 2022; Joudaki et al., 2015 for detailed survey on different methods.

In early work, Rosenberg, Fryback, and Katz, 2000 study upcoding within the claims data. They estimate the probability that a claim has incorrect DRG code, which they further use to identify claims to investigate and audit. Brunt, 2011 study upcoding in the physician office visits data, where they estimate the likelihood of a disease code selected for an office visit. They study the payment differential in the selected code and code used in the data to understand the upcoding practices. **fang2017detecting** find evidence of provider overbilling using inappropriately high number of hours worked to identify outliers.

Recently, Chandola, Sukumar, and Schryver, 2013; Suresh, De Traversay, Gollamudi, Pathria, and Tyler, 2014 introduce methods for provider profile comparison to spot possible misuses or fraud. These works focus on introducing methods and features to represent hospital profiles for comparison, however, do not present any conclusive results. On the other hand, Bauder and Khoshgoftaar, 2018a; Bauder and Khoshgoftaar, 2018b; Herland, Khoshgoftaar, and Bauder, 2018; Bauder and Khoshgoftaar, 2017 utilize publicly available excluded providers to learn models for detection of fraudulent providers. However these approaches rely on availability of human labeled information on fraudulent information, which is often incomplete and hard to obtain for massive Medicare data.

In contrast to earlier methods, unsupervised and explainable methods for the problem, which are more practical in the real world, have received limited attention. Luo and Gallagher, 2010 compare DRG distributions of hospitals providing services for hip replacements and heart health to differences in coding. The underlying assumption is that most hospitals will have similar distribution conditioned on the treatment provided. Recently, Ekin, Lakomski, and Musal, 2019 learn joint distribution of medical procedures and providers using outpatient data. The joint distribution is used to identify provider anomalies based on procedure code and usage frequency by the provider. Most of the work uses only a fraction of massive Medicare data, and often do not incorporate an explanation of results that could be useful to investigators. Our method builds upon these existing studies to provide a precise and explainable detection method that does not rely upon the existence of labeled data.

5.3 Data Description

This study combines data from a variety of sources to detect anomalous provider spending behavior in Medicare and compare it to ground-truth labeling of providers that have faced anti-fraud enforcement.

Our analysis of provider behavior uses a large-scale dataset of Medicare claims. Data were accessed through a data use agreement with the Centers for Medicare and Medicaid Services, facilitated by the Research Data Assistant Center (ResDAC) and the National Bureau of Economic Research (NBER). These hundreds of millions of observations contain extensive data about each hospitalization and patient in the Medicare system, providing an ideal corpus with which to study hospital behavior.

We consider patients hospitalized in 2017, and we use data from 2012 through 2016 to construct the patients' medical history. For these years, we use 100% of samples of Fee-For-Service institutional Medicare data, including inpatient and outpatient claims, and beneficiary [3] information including demographic information and chronic condition indicators from the Chronic Conditions Warehouse. To further understand a beneficiary's medicare history, we use 20% of samples of carrier files, which describe physician office visits.[4]

Table 5.1 describes the sample of inpatient hospitalization claims from 2017. We observe 11.2 million claims from 6.6 million beneficiaries representing 7,661 different providers. Medicare spent in total $131 billion on inpatient care in 2017, out of $710 billion total reported Medicare spending.

Table 5.2 describes our sample used to construct patient medical history from 2012 through 2016. We observe nearly a hundred million physician office visits and

[3] Patient refers to a person receiving health care; beneficiary refers to a person covered by health insurance. Here, they are used interchangeably, as all of our data come from patients who are Medicare beneficiaries.

[4] 20% samples are the largest available for physician office visits.

TABLE 5.1: Inpatient data statistics from year 2017

Spending	
Medicare total expenditure(Statista, 2022)	$710 billion
Medicare inpatient expenditure	$131 billion
Beneficiaries	
Number of inpatient beneficiaries	6.6 million
Number of inpatient claims	11.2 million
Providers	
Number of providers	7,661

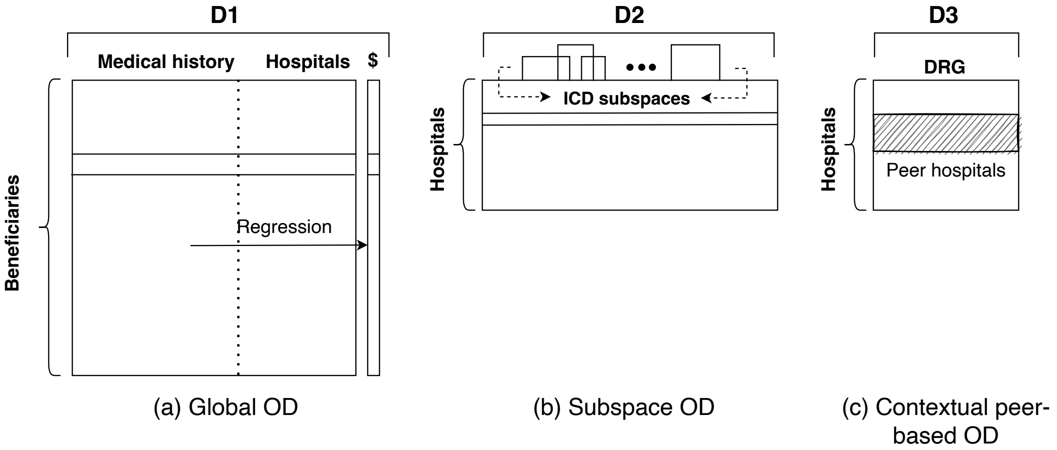
TABLE 5.2: Scale of data from year 2012 to 2016 used to build medical history of patients who are 70 years or older in the inpatient claims from year 2017. The number in each cell is in millions.

	2012	2013	2014	2015	2016
Physician visits	94.7	100.2	102.8	107.7	114.2
Outpatient visits	81.5	87.3	90.9	96.8	104.3
Inpatient visits	4.0	4.2	4.4	5.1	5.8

another hundred million outpatient visits per year, as well as millions of inpatient visits per year. Appendix B.1 provides additional details about the cleaning and use of the Medicare data.

To understand provider characteristics, we use the Medicare Provider-of-Service files, which contain details on providers such as certification number, name, the type of Medicare services that it provides, and type of ownership (private or public). We can identify patients across files using their unique beneficiary identifiers, and we identify providers by their identifiers such as the National Provider Identification (NPI) or CMS Certification Number (CCN). Further, we separately identify Academic Medical Centers based on their membership to Council of Teaching Hospitals (AAMC, 2022). These providers engage in academic research, which could lead them to be ranked as anomalous due to the differences in their claim patterns from other hospitals.

The federal Department of Justice (DOJ) publishes press releases when fraud is identified in order to inform the press and the public as well as deter future fraudulent behavior. To evaluate our automated detection of suspicious providers, we utilize these press releases related to Medicare from the DOJ. To that end, we scraped from the DOJ website thousands of press releases that contain the word ‘Medicare’. Each press release corresponds to a case that the Department of Justice was involved with, often at the time of settlement. Using partial name matching, we tag the hospitals that appear in this corpus. As the DOJ lacks both the capacity and the information to prosecute all Medicare fraud, the press releases provide only a partial list of providers that have engaged in fraudulent behavior. We can consider this a form of positive-unlabeled data: while we can identify firms that have been named in a press release as having likely committed fraud, firms that are *not* named are not necessarily above suspicion. Appendix B.2 provides additional details about the collection and cleaning of the DOJ corpus.



(a) Global OD

(b) Subspace OD

(c) Contextual peer-based OD

FIGURE 5.1: Multi-view anomaly detection on different Medicare data modalities – D1, D2, and D3. Model (a): Global detector based on fixed effects regression model. The coefficient of a hospital is an indicator of excess cost of care at the hospital. Model (b): Local (in ICD codes) detector in the very high dimensional ICD code frequency representation of hospitals. It explains anomalies based on feature importance, i.e. with respect to specific ICD codes. (c): Local and contextual (peer-based) detector based on comparing DRG frequency distributions. It provides a contrastive explanation in terms of excess cost of treatment when compared to peers.

5.4 Method Overview

The Medicare dataset comprises diverse data modalities, which provides an opportunity for modeling the fraud detection problem in various ways. For example, a provider can be represented by the DRG (billing) codes associated with its claims, the frequency of ICD (diagnosis and procedure) codes used in its claims, or by the characteristics of the patient populations that it serves. Each modality presents us with a specific perspective of the data. These different modalities then allow us to learn comprehensive provider behavior which reveal information that cannot be completely uncovered based on only one aspect of the data, since each representation may contain information that is not reflected in others.

In this work, our goal is to estimate a suspiciousness score based on which we rank providers such that anomalous ones are ranked at the top, which may be due to their fraudulent practices. To utilize our different data modalities, we propose an unsupervised *multi-view* anomaly detection approach, suitable for the underlying multi-modal data. Each view (or base detector) presents itself as a different model of the anomalies, operating on a different data representation. As such, each can be seen as providing evidence that corresponds to a particular reason for detection. The explanation provided by each detector provides a unique perspective into suspicious behavior. Collectively, the evidence from these base detectors, i.e. across modalities, can be assembled systematically into an ensemble detection method.

Ensemble methods utilize multiple base detectors, where under certain accuracy and diversity conditions, they are to obtain better performance than the constituent base detector alone and produce more robust results (Aggarwal and Sathe, 2017). Diversity is an important property of ensemble methods, which ensures that the base detectors make independent errors that cancel out when aggregated. Therefore, various approaches have been proposed toward promoting ensemble diversity (Kuncheva and Whitaker, 2003; Nam, Yoon, Lee, and Lee, 2021). In essence, our approach utilizes the diversity of the underlying data representations to induce diversity in the ensemble.

Figure 5.1 shows the different Medicare data modalities we consider and provide a high level description of the corresponding base outlier detection (OD) model that utilizes it. The first model (a) is set up as a global regression onto cost per beneficiary (target variable) from data (denoted **D1** on the figure) reflecting a beneficiary's medical history and the hospitals that they visited. The second OD model (b) performs outlier detection among hospitals as represented by the frequency of ICD codes used in their claims (denoted **D2**). Anomalous coding may be associated with only a few ICD codes (i.e. features) at a time, rather than all. Therefore, the second model is a feature subspace detector, finding outliers locally in subsets of features. Finally, the third OD model (c) performs contextual detection, identifying hospitals that behave differently from their peers. Behavior is captured by the frequency distribution of the DRG codes assigned to each hospital's claims (denoted **D3**). Here, we recognize the heterogeneity among hospitals and compare a hospital's behavior locally, i.e. in the context of its peers with similar characteristics.

In addition to detection, our proposed models can provide explanations for their flagged anomalies. This is especially important in the absence of any ground-truth labels in practice, aiding sense-making, verification and decision making (such as whether to conduct additional investigation or to audit). By capitalization on different data representations, our method leads to different explanations with each OD model, enabling a multi-view reasoning. Specifically, in Figure 5.1, the regression coefficient associated with a hospital in our first OD model (a) would be a direct indicator of excess spending at the hospital. The second OD model (b) quantifies feature (i.e. ICD code) importance, and can explain each flagged anomalous hospital based on the specific ICD codes that they use differently in their claims. The last OD model (c) provides contrastive explanations, through comparing DRG frequencies of a hospital to those of their peers. As the DRG code of a claim dictates cost, differences in the DRG coding distribution can be directly translated to excess cost of treatment. Importantly, the explanation can pinpoint which DRGs are most contributing to large excess cost of a hospital, facilitating auditing.

To arrive at a final anomalous ranking based on different modalities, we combine the rankings from individual detectors such that it captures the agreement among them. In effect, the ensemble approach allows us to gather evidence from multiple models, each leveraging a different data modality. Further, it can be "unrolled" to provide explanations to each flagged anomaly by each detector in the ensemble. Overall, such a multi-view detection and explanation approach takes advantage of corroborating evidences across modalities, and provides a multi-view perspective toward reasoning about suspicious behavior.

The following three sections are organized to present the details of our detection models, in terms of data set up, detection methodology and explanation.

5.5 Expenditure-Based Detection with Massive Fixed-Effect Regression

The goal of a provider-level analysis of expenditure is to understand which providers are associated with high spending on a beneficiary's hospitalization. The incentive of providers who commit fraud is to receive higher reimbursement, and so unexplained high expenditure is potentially suspicious. Our design detects high expenditures that are unexplained by a patient's medical history, which could reflect unnecessary or excessive billing. While any individual patient may receive entirely necessary high levels of care – for example, in response to a severe accident – when a provider's patient population consistently shows expensive, unexplained high expenditure, this may be indicative of fraud or waste.

Our design considers expenditure as a function of a patient's medical history. We collect each beneficiary's medical history, using claims from physicians office visits, hospital outpatient visits, and hospital inpatient visits over a five-year period before the target year. The outcome or target variable is the base claim amount per beneficiary per provider in the current year.

5.5.1 Data Setup

Base payment amount.

For our analysis, we use the base payment amount computed from the Medicare inpatient claims. As explained in Section 6.3, the Medicare Prospective Payment System adjusts the claim payment amount to include expenses due to provider variables such as patient mix, disproportionate share of low-income patients, outlier cases, and expenditure on education and research. These factors are generally external to the provider's coding choice and should be excluded from analysis. Therefore, to understand provider behavior with respect to inpatient encoding, we rely on the base payment amount. The base payment amount is calculated by subtracting the reported adjustment amount from the total claim amount. While payments are also adjusted by provider location through a geographically indexed wage, we do not control for provider wage index adjustments, because the geographical factor will be picked up when controlling for patient location in our regression.

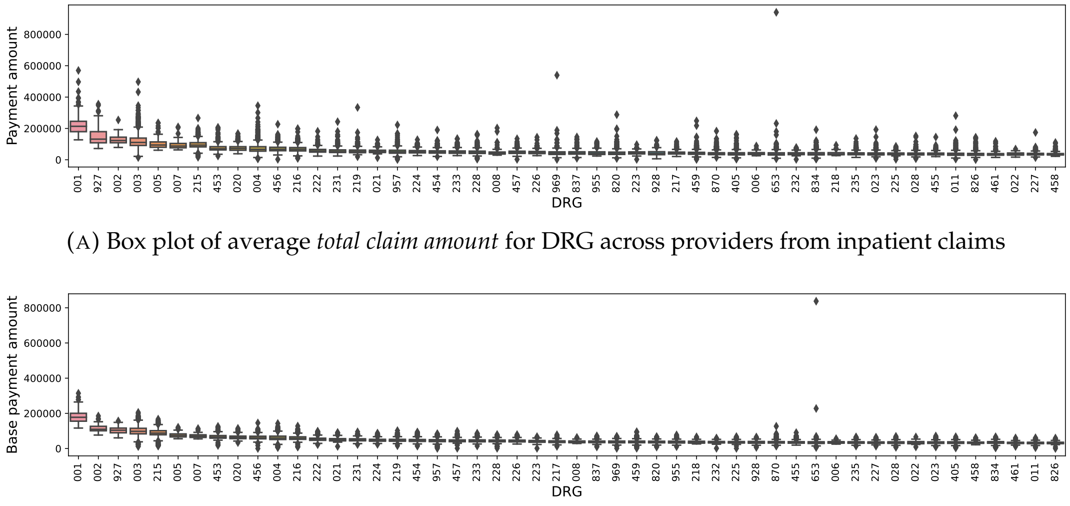
Figure 5.2a shows the box plot of average total claim amount per provider in the inpatient claims data (year 2017) for the top 50 DRGs, sorted by the mean of the box plot. Notice that there is large variation in the average claim amounts for each DRG. This variance across providers is reduced when the box plot instead uses the average base payment amount as shown in Figure 5.2b. However, there remains some variance across providers even when considering the base payment amount.

Patient representation.

We represent each patient by their medical history and their covariates including location.

We consider patients from 2017 who had an inpatient hospitalization claim and are at least 70 years old. Because Medicare is available for individuals aged 65 and older, we include patients aged 70 years or above to ensure we observe a full 5-year history. We construct the medical history based on a patient's provider visits in the previous five years (2012 - 2016). We filter and join patients data from physician visits, outpatient visits, and inpatient hospitalizations in the previous five years. Each patient visit, to a physician or inpatient facility, is assigned codes based on the ICD diagnosis and treatment codes. Thus, for a patient, we collect all the unique codes that were assigned in any of the visits along with their counts.

In addition to the treatment codes, we include the chronic conditions that require regular care, associated with each patient as reported in 2016, the year before the current year. We do not include 2017 chronic conditions as those may be outcomes of the code that the hospitalizations report. Including the 2016 chronic condition of a patient helps understand any comorbidities that may arise due to their medical history and ongoing chronic condition, accounting for the increase in treatment expense. Chronic conditions include diseases such as diabetes, breast cancer, or Alzheimer's



(A) Box plot of average *total claim amount* for DRG across providers from inpatient claims

(B) Box plot of average *base payment amount* for DRG across providers from inpatient claims

FIGURE 5.2: We plot the distribution of total claim and base payment amount across providers from inpatient claims in the year 2017. (a) Distribution of average total claim amount per provider for the top 50 DRGs sorted on the mean of the box plot. There is large variation in the average claim amount for each DRG. (b) Box plot of the average base payment amount across providers for top 50 DRGs sorted on mean of the box plot. The variance across providers is lower for the base payment amount.

disease. Our data provide a comprehensive view of the past treatments received by a patient, and reflects on their health. Further, to account for variation due to a patient's choice of provider, as well as geographic differences in hospital reimbursement rates, we include the patient's location, represented by the first three digits of their zip code.

5.5.2 Detection Model

To estimate expected treatment expense for a patient, we employ a fixed-effects regression model with the outcome or target variable as the total base payment, and the features being the aforementioned patient representation (medical history and location).

We then include as regressors variables corresponding to the count of hospitalizations for that patient at each provider. The coefficients of the provider variables from this regression give the provider fixed effects – in per hospitalization terms– that we use to rank providers.

Note that, because we are interested in capturing the provider-level dependency of cost, we do not include treatment codes from the current year's hospitalization. The codes of the current year's hospitalization reflect the hospital's coding decision, which can be an element in its fraud or overbilling behavior. We address those in Section 5.6. Instead, the providers are added to the model to account for treatment expenses in the current year that are not reflected by the patient's medical profile; see Figure 5.1(a).

Regression model specification for expenditure.

Given (*i*) patient representation $\boldsymbol{X} \in \mathbb{R}^{N\times M}$ for N patients, each with a M-dimensional representation of historical medical profile based on the last five years (2012–2016), and (*ii*) the total base payment Y in year 2017; the specification for expected treatment expenditure prediction is as follows.

$$Y_i \;=\; \beta_0 \;+\; \boldsymbol{X}_i\, \boldsymbol{\beta} \;+\; \sum_j \alpha_j\, H_{j,i} \;+\; \epsilon_i \;\;, \tag{5.1}$$

where Y_i is the total base payment expense for a patient i in 2017; $\boldsymbol{X}_i$ is the patient representation for i, $\boldsymbol{\beta}$ depict regression coefficients associated with patient medical profiles and locations, $H_{j,i}$ is associated with an inpatient Medicare provider j which contains total count of visits to j if patient i visited the provider and 0 otherwise, and α_j's depict the provider fixed effect regression coefficients.

Anomaly scoring.

In the expenditure-based regression, a coefficient α_j can be interpreted as the excess treatment cost due to provider j that cannot be captured by patients' medical profile and location. As such, we can associate the magnitude and sign of this coefficient with the excess spending by a provider, and designate it as its anomaly score.

5.5.3 Model Explanation

The regression model's provider ranking in order of anomalousness is easily explainable through the coefficient values. Specifically, each α_j used for scoring and ranking has the direct interpretation as the excess expenditure on treatment for a patient when visiting the provider j. Therefore, the fixed effects model directly quantifies the excess dollar amount impact of a particular provider, which can be used by an auditor or investigator when deciding which hospitals to investigate.

5.5.4 Evaluation

Figure 5.3 shows the estimated fixed effects, i.e. the α_j coefficients, for providers from our expected expenditure model. The providers with large fixed effects are ranked at the top and flagged as being of suspiciously expensive. In auditing, it is often the case that auditors have a limited budget (time and other resources) for processing red-flags and taking action. Thus, our method allows for targeting of audits towards the most suspicious providers, which corresponds to the highest unexplained spending.

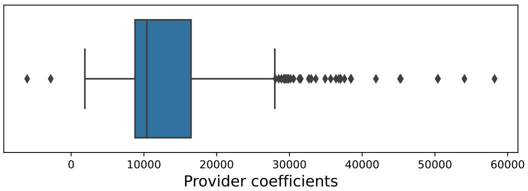
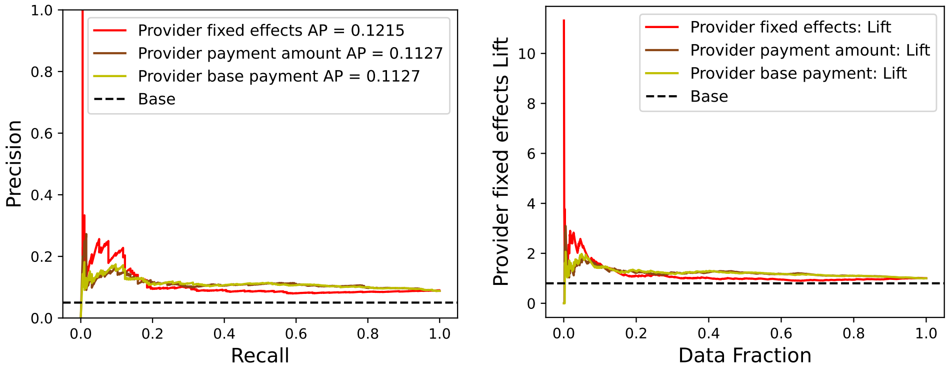


FIGURE 5.3: Distribution of the excess cost of treatment, that is, of α_j's in Eq. (5.1) per provider j. The providers with large excess cost (coefficient) are ranked at the top for audit.

To evaluate the effectiveness of our provider ranking, we use the partial list of known fraudulent providers based on the DOJ press releases described in Section 5.2.3, and we compare our suspicious providers to known fraudulent providers. We quantitatively evaluate the targeting of fraudulent providers using two ranking quality metrics, namely a Precision-Recall (PR) curve, and a Lift curve. The PR curve depicts the positive predictive value (precision) on the y-axis versus the true positive rate (recall) on the x-axis. In audit scenarios with limited budget, a high precision at the top of the ranked list would be useful. Similarly, lift curve measures the targeting effectiveness on y-axis when compared to a random baseline as we move along varying fractions of the ranking on x-axis.

Figure 5.4 reports the PR and Lift curves for our fixed effects model, and compares its performance against two simple intuitive baselines. The baseline methods rank the providers based on average total claim amount and average base payment amount, respectively. Note that our fixed effects model is comparatively more effective at targeting fraudulent hospitals, with relatively higher precision and lift at the top positions.



(A) Precision-Recall curve

(B) Lift curve

FIGURE 5.4: We report (a) Precision-Recall curve (AP: Average Precision denotes area-under-curve) and (b) Lift curve for provider ranking produced by fixed effects coefficients against two simple baselines: ranking of providers based on (1) average total claim amount and (2) average base payment amount. Dashed horizontal line 'Base' depicts the random ranking. Notice that top of the ranking is comparatively better as evidenced by higher precision and lift when recall and top data fraction are low. This is particularly helpful for auditors who would typically process only top ranked providers under limited budget.

Figure 5.5 reports the result of a two-sample test on the fixed effect coefficients as estimated by our model for providers in the DOJ corpus versus the rest of the providers. Notice that the DOJ providers typically have larger fixed effects as compared to others, and their distribution is significantly different as the test rejects the null that the two sets of coefficients are drawn from the same distribution, with $p < 0.001$. We remark that the reported performance is conservative and only the lower limit on our model's targeting ability, since many top ranked providers that are not part of DOJ ground truth may still have been involved in suspicious behavior. We report more qualitative results, and provide case studies through explanations into such flagged providers in Section 5.8.2, after accounting for the evidence from other models in our ensemble.

5.6 ICD Coding-Based Detection with Subspace Analysis

International Classification of Disease (ICD) codes are used by health care providers to characterize a patient's medical condition and treatment. The US uses ICD-10 codes, which were developed by the World Health Organization and can be used to

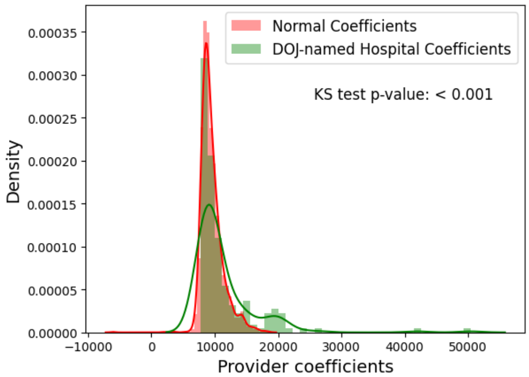


FIGURE 5.5: Comparison of fixed effect coefficients for providers facing anti-fraud lawsuits (known fraudulent entities or outliers) versus the rest of the providers (normal entities). A two-sample test rejects the null hypothesis, implying significantly different distributions statistically.

designate the universe of medical issues and procedures. ICD codes encode provider assessment of a patient based on their reason of visit to the hospital and their medical conditions, and primarily reflect the diagnoses and applied procedures for treatment. For Medicare billing, the assigned ICD codes are then used as input to a "grouper" software used by hospital billers that assigns a diagnostic code (DRG) based on the provider findings as indicated by the assigned ICD codes. As discussed above, in the Medicare PPS, the DRG code determines the reimbursement level. Consequently, ICD coding presents opportunities for miscoding, as providers may try to achieve a more expensive DRG code to obtain higher reimbursement. Therefore, the objective of our ICD coding based analysis is to understand provider coding practices that could reveal the coding patterns applied by providers engaging in fraudulent behavior.

5.6.1 Data Setup

Provider representation.

We use inpatient claims from the year 2017 to understand how providers assign ICD codes to each claim, and represent providers through their reported ICD codes, including diagnostic and procedure codes. This representation captures the coding practices of a provider.

Importantly, since providers have a choice of ICD codes, we also account for ICD *code substitutability*, where a slightly similar ICD code could be used instead to yield higher reimbursements. To capture code substitutability, we estimate the semantic similarity of the description of each code within each chapter of the ICD code hierarchy. Here, the description of each ICD code is constructed by concatenating its text description to the description of its ancestor codes within the ICD hierarchy. Then, pairwise Jaccard distance is computed between the descriptions of the codes and the provider representation is updated using the ICD code similarity.

For example, the description of ICD code J45.20 under chapter X is constructed by concatenating the descriptions of J00-J99 chapter, J40-J47 block, J45, and then the ICD code J45.20 resulting in the description given as "Diseases of the respiratory system – Chronic lower respiratory diseases – Asthma – Mild intermittent asthma uncomplicated. This representation ensures that codes with similar positions in the ICD hierarchy have somewhat similar text descriptions and are therefore near each other in Jaccard distance.

Specifically, let $\boldsymbol{X}^{ICD} \in \mathbb{R}^{N_H \times M_H}$ be the matrix representation of N_H providers in terms of M_H-dimensional ICD codes in which the entries depict the total code usage count by provider, and $\boldsymbol{J} \in \mathbb{R}^{M_H \times M_H}$ be the ICD substitutability matrix consisting of pairwise Jaccard similarities. Then, the provider representation $\boldsymbol{X}^{ICD_{sim}} \in \mathbb{R}^{N_H \times M_H}$ after incorporating the code substitutability is given as $\boldsymbol{X}^{ICD_{sim}} = \boldsymbol{X}^{ICD} \times \boldsymbol{J}$, which re-distributes each code's frequency to substitutable ICD codes that are not directly reported in the claims data.

We note that $\boldsymbol{X}^{ICD_{sim}}$ is very high dimensional ($> 40,000$ features). However, anomalous coding of a claim is likely covert and associate with only a few ICD codes. Therefore, we employ a feature *subspace* based detector for finding outliers locally among subsets of ICD codes. Figure 5.1(b) shows this setup.

5.6.2 Detection Model

We employ a suite of subspace outlier detectors on the high dimensional provider representation $\boldsymbol{X}^{ICD_{sim}}$ to find providers deviating from the majority coding practices within certain ICD subspaces. As we are interested in ICD subspaces that are relevant for a variety of aberrant provider practices, we utilize an ensemble of subspace detection methods that are effective on high dimensional data. In the same spirit as with our overall approach, the ensemble allows us to examine multiple diverse subspaces as each subspace detection method implements a different methodology for exploring candidate subspaces. In particular, our subspace ensemble uses five different state-of-the-art methods that we describe briefly below.

Subspace outlier detection.

While we represent a hospital in the high dimensional ICD space, the abnormal or aberrant behavior may be reflected only in a small, locally relevant subset of codes as pertains to stealthy behavior. Each OD algorithm in the ensemble explores local subspaces differently to provide evidence from diverse subsets. To that end, our OD model consists of the following subspace detectors:

(i) Subspace Outlier Degree (SOD) (Kriegel, Kröger, Schubert, and Zimek, 2009) locally examines each point (hospital) in the data. For each data point, it computes reference points through shared nearest neighbors. The subspace is then characterized by dimensions with low variance, lower than a provided threshold, within the identified reference set. It records the deviation of each data point from the hyperplane spanned by the mean of the identified subspace, where outliers have larger deviation.

(ii) Isolation Forest (IF) (Liu, Ting, and Zhou, 2008b) builds a collection of randomized trees that approximate the density of data points in a random feature subspace characterized by paths in what are called "isolation trees". Each isolation tree is constructed by recursively partitioning data using a randomly chosen point in a randomly selected dimension, until the leaf of the tree contains a single data point. Shorter paths in a tree indicate sparse regions as fewer partitions lead to leaf nodes, and points belonging to each leaf at lower depth indicate outlierness in the subspace characterized by the tree path.

(iii) Robust Random Cut Forest (RRCF) (Guha, Mishra, Roy, and Schrijvers, 2016), like IF, also constructs an ensemble of randomized trees by recursively partitioning the data. It computes the model complexity of each tree as the sum of the bits required to store the depths of each point in the tree. An outlier is defined as a point which increases the model complexity significantly when added to the tree.

(iv) Lightweight on-line detector (LODA) (Pevný, 2016) constructs a collection of histograms on random 1-dimensional projections of the data. Each data point is then associated with the negative log-likelihood based on each histogram, and data points are ranked based on their average likelihood across the 1-D histograms.

(v) RS Hash (RSHASH) (Sathe and Aggarwal, 2016), like LODA, is also an ensemble of histograms; however, it constructs a collection of grid-based histograms in randomly chosen subspaces, and grid sizes vary based on varying sample sizes of data. Each data point is then scored by the number of sampled points sharing the same bin in the histogram. A sparsely populated bin is indicative of outlierness.

We apply the above methods to $\boldsymbol{X}^{ICD_{sim}}$, the ICD representation of providers, and identify the providers that behave abnormally in various subspaces as explored by the algorithms.

Anomaly scoring.

Each subspace algorithm assigns an anomaly score to each provider. The scores have different scale and semantics (path length, likelihood, etc.), and thus are not directly comparable across the methods. Therefore, we aggregate the ranking of providers based on individual scoring of each subspace method. We use the instant-runoff voting technique (details in Section 5.8) for rank aggregation from different subspace algorithms, and provide the final ranking of hospitals by anomalousness across all subspaces.

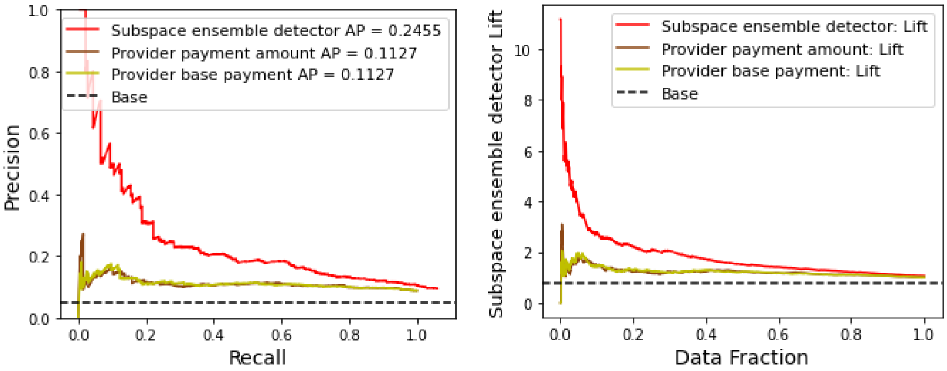
5.6.3 Model Explanation

We explain the ranking of a subspace detector using Shapley Additive Explanation values (SHAP values), introduced in Lundberg and Lee, 2017 and Lundberg et al., 2020. SHAP values estimate feature importance by approximating the effect of removing each feature from the model as the average of differences between the predictions of a model trained with and without the respective feature. We regress the anomaly scores from a subspace detector onto the ICD representation of providers, and then estimate the SHAP values under the regression model. The feature contributions for each observation find the most important codes that affect the anomaly score significantly. This helps us find ICD codes that are contributors to a provider being ranked as an outlier.

Further, we provide dollar amount characterization of important features (ICD codes). Each ICD code is mapped to the most frequent DRG code assigned for the given ICD code within the inpatient claims. Since DRG codes are determinants of the payment for care, through this most-frequent DRG mapping, we associate dollar amount of reimbursement to ICD codes. This lends itself to understanding the dollar amount impact of an important ICD code for an anomalous provider as explained by SHAP feature importance values.

5.6.4 Evaluation

Figure 5.6 reports the performance of our subspace OD model in terms of the PR and Lift curves, using the DOJ ground truth. The subspace model ranking is at least 2×



(A) Precision-recall curve

(B) Lift curve

FIGURE 5.6: We report (a) Precision-Recall curve and (b) Lift curve for provider ranking produced by our ICD-10 subspace outlier detector ensemble against two simple baselines that rank the providers based on (1) average total claim amount and (2) average base payment amount. Dashed horizontal line 'Base' depicts the random ranking.

better at targeting fraudulent providers compared to our two baselines, respectively based on total claim payment and base payment amounts. Our method substantially outperforms random auditing or even detection based strictly on payment amounts.

5.7 Expenditure-Based Detection with Peer Analysis

Our third model is based on peer-based excess spending detection and examines the coding decisions of hospitals as compared to similar "peer" hospitals that treat similar populations. In short, we identify hospitals who are exposed to the same patient population but manage to assign more expensive DRG billing codes.

The objective of the peer-based analysis is uncovering the local patterns of spending behavior among a *related* group of providers called peers, and identifying providers deviating from the group's expected behavior. We utilize the inpatient claims to create a profile for each provider under two complementing data modalities, based on: (1) type of services provided by the hospital, and (2) the patients' chronic condition profiles served by a hospital. We then find groups of related providers based on the similarity of their provider profile representation.

To identify a locally aberrant behavior, each provider is represented in terms of its DRG frequency distribution, which determines spending. Then, the DRG representation of a given provider is compared to the summary DRG distribution of their peers. Figure 5.1(c) visualizes this setup. The providers are then ranked in order of their deviation from group behavior in terms of DRG-based spending.

5.7.1 Data Setup

Provider representation.

We construct hospital profiles to capture the nature of services provided, the characteristics of patient population served, and encoding practices that drive spending for treatment.

Provider profile – Type of services. We first examine a provider's inpatient claims data to understand the type of services provided. Because the DRG codes assigned by providers may be manipulated to accomplish higher reimbursement, we must not represent providers by the exact DRGs they use; instead, we consider the provider's distribution into major diagnostic categories (MDC) (ResDac, 2022). Each MDC corresponds typically to one major body system (circulatory, digestive, etc), and can be associated with a set of medical specialties; each MDC contains a large set of potential DRGs. Therefore, characterizing providers by MDC allows us to consider providers that treat patients with similar types of medical needs, but without relying on the exact DRG codes assigned. For each provider, we record the normalized count of each MDC code in the inpatient claims data in the current year.

Provider profile – Patient population. We create another profile based on patient population characteristics served by a provider. The underlying motivation for this profile is that two providers should be similar if they serve patients with similar medical conditions. To characterize the patient population at a broad level, we use the underlying chronic conditions of the patients. The chronic conditions flag whether a patient has received a previous set of services related to a chronic condition such as diabetes or ischemic heart disease. As a provider's representation, we record the normalized count of the chronic conditions of all the patients treated at the provider.

Provider profile – Spending for care. The spending amount in each claim is directly tied to the assigned DRG code. To capture the DRG encoding practices of a provider, we represent each provider using the normalized counts of DRG codes from its inpatient claims. The DRG frequency representation allows us to compare and contrast the spending between a hospital and its peers that provide similar services or serve similar patients.

5.7.2 Detection Model

Peer identification.

We create peer groups of hospitals that share similarities in the type of services provided or the patient population served.

Let v_j denote the representation for provider j; either based on the type of services profile using MDC codes or based on the patient population profile using chronic conditions of patients. We note that the provider representations are frequency distributions, as they depict normalized counts. Therefore, to measure the similarity between two providers j and k, we use the Hellinger distance for probability distributions, which is an upper bound on the total variation distance (Bar-Yossef, Jayram, Kumar, and Sivakumar, 2004), given as

$$d_{jk} = \frac{1}{\sqrt{2}} \cdot \|\sqrt{v_j} - \sqrt{v_k}\|_2 \tag{5.2}$$

We examine the distribution of pairwise similarity values to decide on a threshold τ to include only the most similar providers in a provider's peer group.

For each provider j, the providers with similarity to j above τ constitute j's peers, denoted $\mathcal{P}_j$. Notice that the peers are specified for each provider separately, rather than using any clustering algorithm. This allows us to create compact peer groups of varying sizes. We note that fixing the peer group size would be a subpar alternative, since j's group may then include distant providers as peers, skewing the representative summary statistics of the group that j is compared to.

Anomaly scoring.

In the Medicare PPS, the reimbursement amount for treatment is directly based on the assigned DRG code to a claim. Therefore, for anomaly scoring, we utilize the provider representations over DRG codes from the inpatient claims, which consist of the normalized counts of the DRG codes used by a provider. In short, this detection mechanism assumes that providers who treat similar patient populations, or provide care for similar illnesses and injuries, should have similar DRG distributions.

For each provider, we have identified a group of providers (peers) with similar characteristics—type of services provided and patient population served—based on which we create a peer group summary in terms of distribution over DRG codes. The summary distribution is created by incorporating DRG frequencies from all the peers, weighted by their similarity to the provider under consideration. Let $\boldsymbol{v}_j^{DRG}$ be the DRG distribution for provider j with n_j claims, and $\boldsymbol{q}_j^{DRG}$ be the summary DRG distribution based on provider j's peers, defined as follows.

$$\boldsymbol{q}_j^{DRG} = \tfrac{1}{Z} \textstyle\sum_{k \in \mathcal{P}_j} n_k \times (1 - d_{jk}) \times \boldsymbol{v}_k^{DRG} \quad \text{where} \;\; \mathcal{P}_j = \{\, k \mid (1 - d_{jk}) \geq \tau \,\} \;\text{ and } Z = \textstyle\sum_{k \in \mathcal{P}_j} n_k \times (1 - d_{jk}) \tag{5.3}$$

Next we tie the DRG usage frequencies to average dollar amount spending by Medicare, as the former dictates the latter. $Cost(c)$ denotes the average base price of DRG code c computed from the inpatient claims data from the year 2017. Then, the excess spending for treatment per claim on average for provider j is given as follows:

$$ExcessSpending_j \;=\; \sum_{c \in DRGs} Cost(c) \;\times\; (\boldsymbol{v}_{j,\,\text{index}(c)}^{DRG} - \boldsymbol{q}_{j,\,\text{index}(c)}^{DRG}) \tag{5.4}$$

where $\boldsymbol{v}_{j,\,\text{index}(c)}^{DRG}$ is the frequency corresponding to DRG code c in the DRG representation $\boldsymbol{v}_j^{DRG}$ for provider j, and $\boldsymbol{q}_{j,\,\text{index}(c)}^{DRG}$ denotes that for DRG code c in the peer group summary representation $\boldsymbol{q}_j^{DRG}$. In short, this amount computes how much more a provider spends because they use a different set of DRG codes than their peers, based on the average price of those DRGs.

The calculated $ExcessSpending$ amount is the anomaly score based on which the providers are ranked, as it depicts the average spending discrepancy for a provider when compared to peers of the given provider. Since we create two different peer groupings – one based on services provided, and another based on patients served – we obtain two rankings, later combined through instant-runoff voting (Section 5.8).

5.7.3 Model Explanation

The peer based OD model's anomaly score is the estimated excess spending, which is directly interpretable as the extra dollar amount a provider charges on each claim on average as compared to what would be expected from other similar providers. Further explanation can be provided for a top-ranked provider by contrasting their frequency distribution over DRG codes against their peers. This allows auditors to have a contrastive understanding of DRG codes used by similar providers, and to pinpoint to specific DRGs with large frequency discrepancies. Direct usage comparison of individual DRGs could point to specific codes that contribute most to the overall spending at a provider, and guide a deeper investigation of the claims associated with those specific DRG codes.

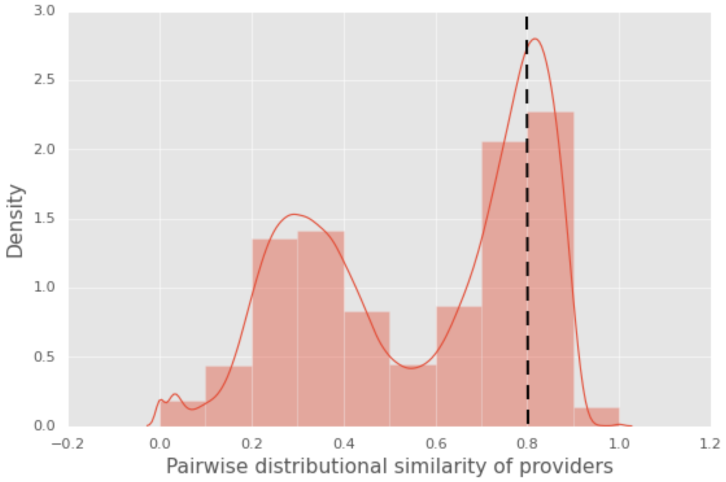


FIGURE 5.7: Distribution of pairwise similarities between provider representations. A provider and its peer hospital pair has similarity $>=0.8$.

5.7.4 Evaluation

Figure 5.7 shows the distribution of pairwise similarities between hospitals, and mark the similarity threshold at $\tau = 0.8$ which is used in our implementation for identifying peers. We exclude providers from our analysis that have less than five peers for the chosen threshold, as the estimation of excess spending could be noisy for these providers due to small peer group. Providers with large excess spending are ranked at the top and are identified as suspicious.

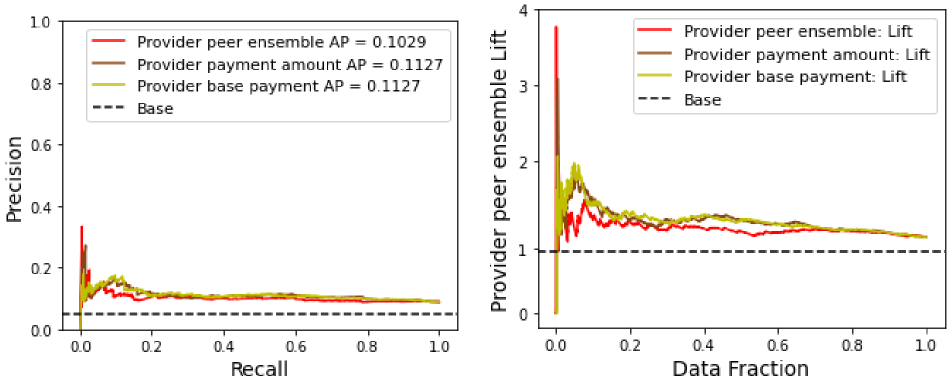
We use the DOJ corpus to evaluate our ranking of the providers based on excess spending. Figure 5.8 reports the PR and Lift curves for our peer analysis. The ranking is also compared to the two baselines, respectively ranking providers by average total claim amount and average base payment amount. Although the peer-based ranking performance is comparable to these simple baselines, we remark that it is the lower bound on the performance. Furthermore, besides a mere ranking and unlike these simple baselines, our model can provide a nuanced explanation through DRG code frequency discrepancies, providing auditors with reasoning for potential factors driving the high spending. Finally, our model fundamentally identifies expensive hospitals as compared to their peers, which may be of interest to auditors interested in waste that may not rise to the level of fraud detected by the DOJ.

Through case studies in Section 5.8, we report further qualitative results and provide peer-based explanations and insights into top flagged providers after aggregating evidences from different OD models.

5.8 Aggregate Provider Ranking

Each outlier detection model presented above is a component of our ensemble method that considers a different data modality and creates a ranked list of providers based on the evidence examined individually. This ensemble method is designed to handle multi-view Medicare data, where different features of the data can be used to evaluate different aspects of suspiciousness. The goal of the ensemble is a single suspiciousness ranking for all providers.

To arrive at the final ranking for auditing, we merge multiple rank lists into a single ranking using instant-runoff voting (IRV). Our goal is to present the aggregate ranking that is most representative of the component models. IRV combines results



(A) Precision-recall curve

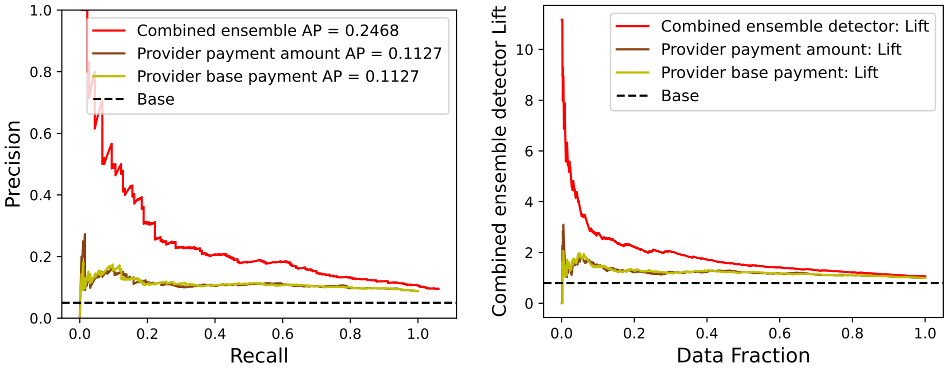
(B) Lift curve

FIGURE 5.8: We report the performance of ranking based on excess spending amount compared to the peers, where peers are identified via similarity based on MDC distributions and patient chronic conditions.

across rankings in a way that best reflects the information contained across multiple models (Franceschini, Maisano, and Mastrogiacomo, 2022).

The rank aggregation proceeds in an iterative manner, where each round utilizes the IRV procedure to find a "winner" (in our case, most suspicious hospital). In each round, votes are counted for each component ranking's first choice, and a hospital with a majority of votes is then ranked at top in our aggregate ranking. The rank lists across models are updated to drop the selected hospital in this round, and the IRV procedure is repeated with updated rank lists in the subsequent rounds to arrive at an aggregate ranking.

In our implementation, we aggregate 8 different rankings across our 3 OD models; one from the regression model, five from different subspace OD algorithms, and two from the peer-based model utilizing two separate similarity measures. Next, we show the effectiveness of our final aggregate ranking for identifying fraudulent hospitals in the Medicare system through quantitative and qualitative evaluations.



(A) Precision-recall curve

(B) Lift curve

FIGURE 5.9: We report the performance of the final ranking of providers as aggregated from 8 rankings based on 3 different OD models. Note that aggregated ranking improves over the ranking by individual constituent experts. The proposed ensemble is on average 4× better than the random targeting of providers for auditing.

5.8.1 Quantitative Evaluation

Figure 5.9 shows the evaluation of our aggregate ranking of hospitals using a PR curve and a Lift curve. The aggregate ranking is compared to intuitive baselines that rank hospitals based on their average reimbursements, or random auditing. Our aggregate ranking is able to target fraudulent providers on average twice as better when compared to the baseline ranking—note the area-under-curve, or average precision (AP) values on legend Figure 5.9(a).

While only 1 in 20 hospitals are named in the DOJ Press releases, the top 50 hospitals identified by our aggregate ranking contain 21 providers named in the DOJ corpus. That is an 8-fold lift in detection rate considering the evaluation at top 50 hospitals, with an average of 4-fold lift over random/by-chance targeting across varying data fractions as seen in Figure 5.9(b). Importantly, our ground-truth consists only of providers named in the DOJ corpus, while there may be others with yet unidentified fraudulent practices – and therefore, our list can be used to find other hospitals not yet identified as fraudulent.

5.8.2 Qualitative Explanation: Case Studies

In this section, we present an analysis of our multi-view detectors, highlighting some of the salient aspects for the fraud detection task. In particular, we discuss how our multi-view detectors can be used to explain the aberrant patterns employed by top ranked flagged hospitals by highlighting parts of data from different views that contributed most to the ranking, which can assist in the process of auditing or deeper investigation.

We examine two top ranked providers from the aggregate ranking (1) the provider at rank 1 that is also named in the DOJ corpus, and (2) the highest-ranked provider which is not in our ground truth (at rank 5, as ranks 1–4 all are part of DOJ ground truth). In the following two case studies, we show how different models contribute evidence toward a better understanding of how each provider stands out.

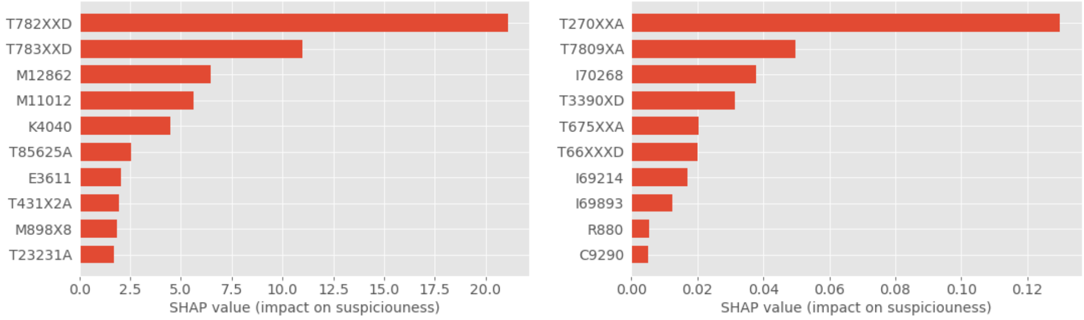
Case 1: Flagged hospital named in DOJ corpus

Our aggregate ranking finds the Cleveland Clinic as the most suspicious hospital under our metrics. Here we present evidence from our 3 Outlier Detection models, where this provider is ranked at #1 by the subspace OD model, ranked at #17 by the peer-based model, and ranked at #27 by our regression-based model.

Notably, the Cleveland clinic settled with the DOJ in the years 2015 and 2021 for \$1.74 million (**cleveland2015settlement**) and \$21 million (**cleveland2021settlement**)[5] respectively. The evidence from our models do not directly match the reason for DOJ settlements; put differently, our exact explanations have not been validated externally by litigation. Moreover, our data do not provide evidence of fraud by the Cleveland Clinic, nor do they substantiate claims from lawsuits against the Clinic. The existence of previous lawsuits by the DOJ against the Clinic validate that this is a provider with past bad behavior, and our metric indicates that this provider engaged in anomalous behavior that can be detected by our algorithm and merits deeper investigation.

Our regression model estimates the excess expenditure on treatment for a patient when visiting the Cleveland Clinic to be \$29,844.33, which is almost 3× the average

[5] This 2021 enforcement was against Akron General Health System, which was acquired by the Cleveland Clinic foundation in 2015.



(A) Top flagged provider that is named in DOJ

(B) Top flagged provider that is not in DOJ

FIGURE 5.10: ICD codes contributing to suspiciousness of top ranked providers based on SHAP values

expenditure (≈$10K) as shown in Figure 5.3. This does not, by itself, indicate that the Cleveland Clinic engaged in bad behavior, as this may reflect that it performs more specialized medical procedures, although our regression accounts for the patient's recent medical history.

One potential concern is that the hospital highlighted in this example, the Cleveland Clinic, as particular aberrant is a unique hospital that serves a particularly sick patient pool, and that therefore, the results are driven by selection of patients into different hospitals, as opposed to the effect of being treated at that hospital on expenditure. We argue this is not the case. Indeed, the two closest peer hospitals to the Cleveland Clinic are New York Presbyterian and Beth Israel Deaconess, both of which are similarly prestigious hospitals involved in specialty care. Therefore, we expect that the results reflect actual divergent coding patterns by the most suspicious providers, rather than detecting hospitals that are engaged in specialty treatment.

Figure 5.10a plots the most important ICD codes that contribute to the anomaly score of the provider from the subspace OD model, based on SHAP values. The top ICD code "T782XXD" is described as "Anaphylactic shock, unspecified, subsequent encounter" which falls under the ancestor "T78" with the description: "Adverse effects, not elsewhere classified"[6]. As such, T78 appears to be a catch-all classification for adverse effects for injuries, poisoning, and other consequences of external causes for visit. Moreover, the code T782XXD is considered exempt from reporting whether the condition is present on admission (POA) to an inpatient facility. The next ICD code "T783XXD" is under the same ancestor, T78, and is also considered exempt from reporting if POA. Similarly, the description of code "M12862" allows non-specific reasons to be used for encoding as the given description is: "Other specific arthropathies, not elsewhere classified, left knee".

We next examine the reimbursement amounts related to these ICD codes, based on their mapping to the DRG they are most frequently associated with. The distribution of the amounts across all ICD codes is given in Figure 5.11. The codes T782XXD and T783XXD can be mapped to two DRG codes: 949 (Aftercare with cc/mcc) and 950 (Aftercare without cc/mcc).[7] The reimbursement amount for DRG code 949 is about 25% more compared to DRG code 950, where T782XXD is reported most frequently against DRG code 949. Further, within the ICD-10 hierarchy, codes T782XXD

[6] ICD codes are available for lookup through ICD10Data. This code is available online at: https://www.icd10data.com/ICD10CM/Codes/S00-T88/T66-T78/T78-

[7] Here, 'cc' and 'mcc' stand for Complication or Comorbidity and Major Complication or Comorbidity, respectively.

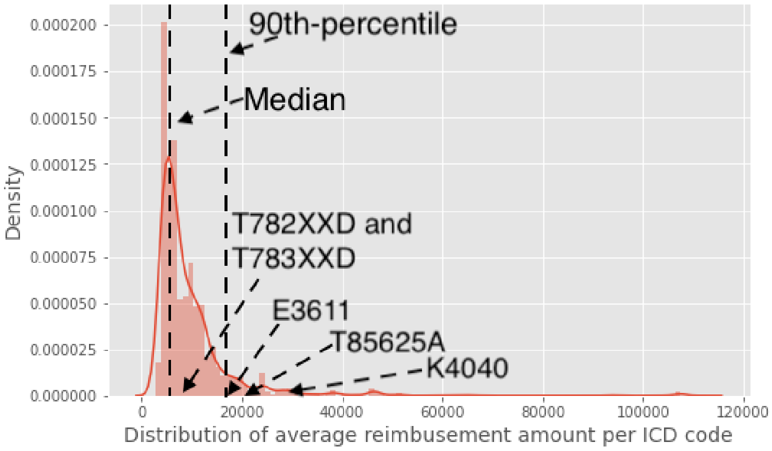
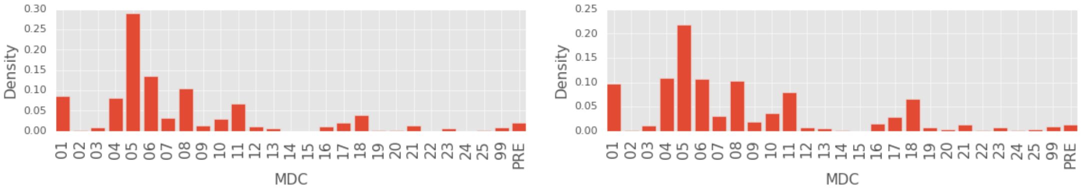


FIGURE 5.11: Distribution of ICD reimbursement amount obtained after mapping ICD code to most frequent DRG code in the inpatient claims data in year 2017. The median reimbursement amount is $6,650.88, and the 90-percentile reimbursement amount is $16,401.04.

and T783XXD are the most expensive and get at least 50% more reimbursement than any other sibling or parent code. Notably, 6 out of top 10 ICD codes contributing to anomaly score (as shown in Figure 5.10) have reimbursement amounts that are more than 50th percentile among all ICD codes, while 3 of them associate with DRG codes with amount above the 90th percentile (see Figure 5.11). All these factors explain, through specific ICD codes, associated DRGs and dollar amounts, the reasoning behind why a flagged provider stands out. This evidence provides starting points for further investigation.

In the peer-based model, the provider is flagged through the peer relation of providers with respect to their MDC representation. Figure 5.12 shows the MDC distribution of the Cleveland Clinic and its nearest peer provider. Notice that in terms of facilities and services provided as encoded by their MDC, the two hospitals are quite similar. We compare the DRG representation of the Cleveland Clinic to the summary DRG representation of all its peer hospitals over the top 50 DRG codes that are selected based on their contribution to excess spending (see Eq. 5.3 for excess spending estimate). As shown in Figure 5.13, Cleveland Clinic uses certain DRG codes more frequently than its peers as indicated by the summary distribution—starting with 219, 220, as well as 309, 310, 330. DRG codes 219 and 220 belong to “Cardiac Valve and Other Major Cardiothoracic Procedures” with reimbursement amount in top 4 most expensive within MDC 05. DRG codes 309, 310 are described as “Cardiac Arrhythmia and Conduction Disorders”, and DRG code 330 is described as “Major small and large bowel procedures with cc”. Note that the description of codes 309, 310 and 330 is specific to a particular condition, while the description for 219–220 allows for ambiguity. Ambiguity may provide opportunities for miscoding to reach for higher reimbursement.

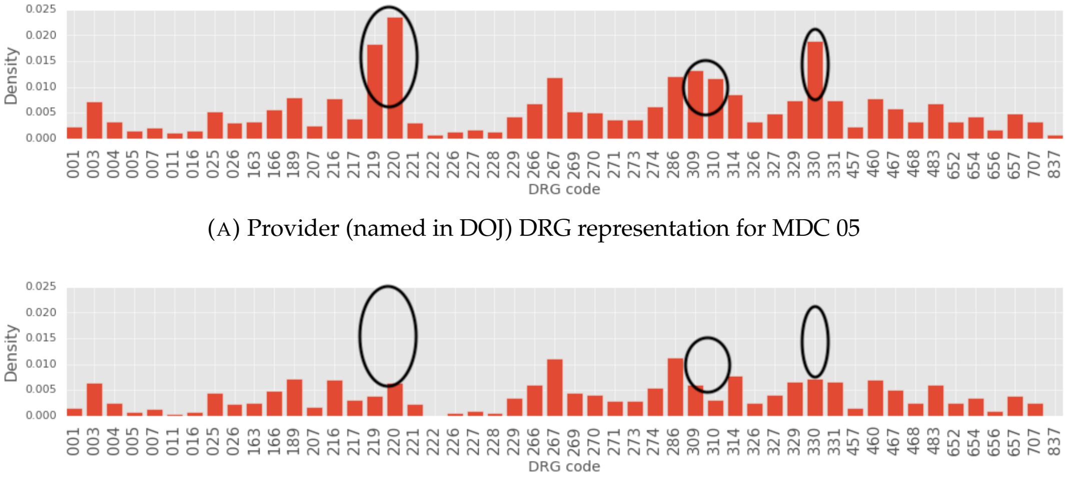
In summary, all three outlier detection models point to evidence from different views of the claims data that makes the top ranked hospital stand out from others, both in terms of local and global analysis. These pieces of evidence explain the ranking by shedding light into certain coding practices that a provider engages in, and may be utilized in further audit processes.



(A) Provider

(B) Nearest peer

FIGURE 5.12: Provider (named in DOJ) and its nearest peer represented in terms of MDC codes indicating provider facilities and services provided.



(A) Provider (named in DOJ) DRG representation for MDC 05

(B) Summary DRG distribution of its peers for MDC 05

FIGURE 5.13: Comparing the DRG distribution of provider (named in DOJ) to the summary distribution created from its peer hospitals.

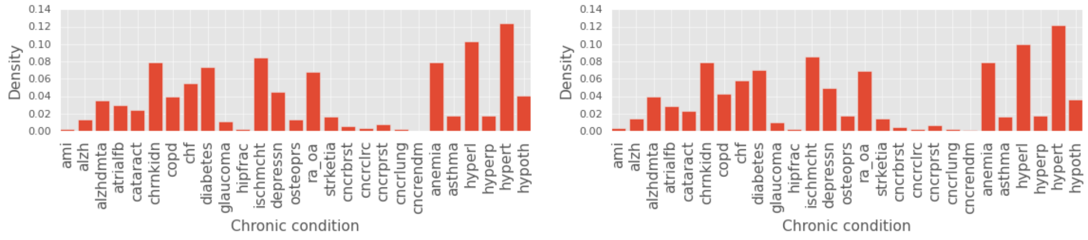
Case 2: Flagged hospital not in DOJ corpus

We now turn to a hospital which is flagged as suspicious by our metric but was never named in a DOJ press release.

In the aggregate ranking, AdventHealth Orlando hospital is ranked at #5 in order of suspiciousness. All 4 hospitals higher in the ranking were named in the DOJ corpus, motivating this case study. This provider is ranked at #5 by the subspace OD model, and ranked at #35 by the peers-based model.

It is important to note that our model does not provide evidence of fraud, nor do we claim that AdventHealth Orlando has committed any fraud. Rather, our ranking of hospital suspiciousness can be used to guide further investigation and audits, and we use this case study to examine how our explainable model can help direct investigatory resources toward the exact claims that make a provider different from its peers.

Figure 5.10b presents the bar plot of the top 10 ICD codes by importance for the provider, based on SHAP values for the anomaly ranking from our subspace OD model. Note that 5 out of these top 10 ICD codes fall under ICD-10 chapter “S00-T88 Injury, poisoning and certain other consequences of external causes”. The ICD code T270XXA is most frequently mapped to DRG code 205 which is described as “Other respiratory system diagnoses with mcc”. The 3rd ranked ICD code “I70268” is described as “Atherosclerosis of native arteries of extremities with gangrene, other extremity”. Based on the descriptions of these top ICD codes, a common thread

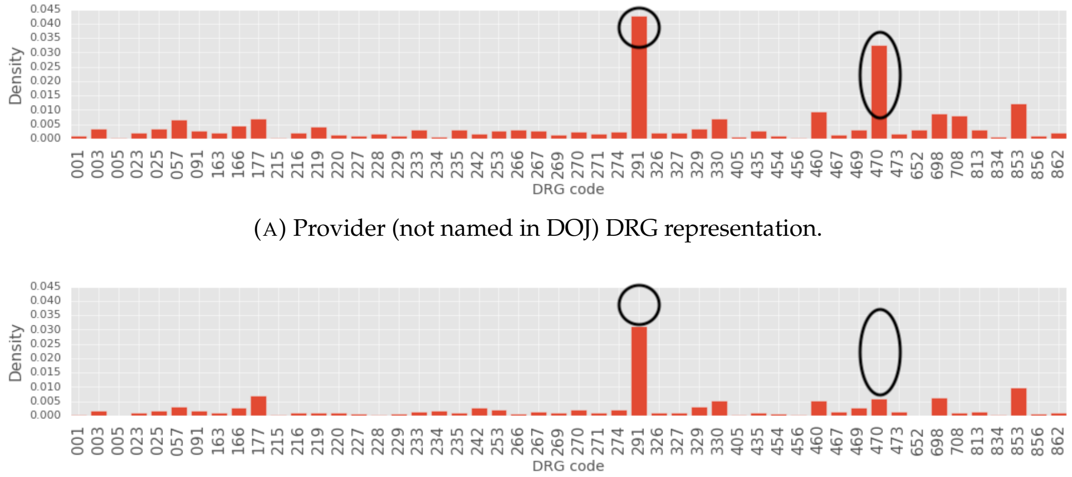


(A) Provider

(B) Nearest peer

FIGURE 5.14: Provider (not in DOJ corpus) and its nearest peer represented in terms of patient population served

appears to be that the codes leave room for ambiguity—due to the catch-all word 'other' in their descriptions. Further, 7 out of 10 ICD codes have reimbursement amount larger than the 50th percentile, and 4 out of 10 have reimbursements larger than 90th-percentile reimbursements across all ICD codes (recall Figure 5.11 for the ICD price distribution).



(A) Provider (not named in DOJ) DRG representation.

(B) Summary DRG distribution of its peers

FIGURE 5.15: Comparing the DRG distribution of provider (not named in DOJ) to the summary distribution created from its peer hospitals.

Next we present evidence from the peer-based OD model, though the provider is not top ranked in this model. Figure 5.14 shows the provider and its nearest peer hospital that serve similar patient populations, represented in terms of chronic conditions of the patients. We note the almost identical distributions of chronic conditions for the provider and its nearest peer hospital. We compare the DRG distribution of the provider to the summary DRG distribution of its peers.

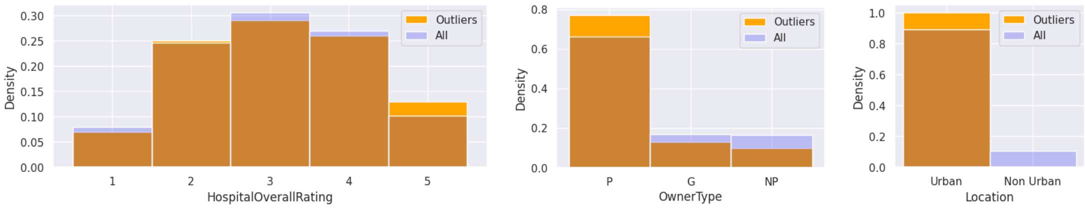
Figure 5.15 shows the distribution over the top 50 DRG codes, where the provider's distribution deviated from the summary distribution the most weighted by DRG reimbursement amount (see Eq. 5.3). We find that excess expenditure is almost entirely driven by two DRG codes, namely 291 (heart failure and shock with mcc) and 470 (major joint replacement or reattachment of lower extremity without mcc) with reimbursement costs larger than the 50th-percentile among DRG codes.

Similar to the earlier case, our models pinpoint specific ICD and DRG codes that can help jump-start further investigation, while highlighting dollar amount discrepancies that provide perspective with respect to monetary value.

5.9 Characterizing Outlier Providers

In this section, we examine the covariates of hospitals to understand the factors that characterize an outlier hospital as detected by our model. The covariates used in the analysis depict various hospital characteristics such as hospital rating, number of unique patients served, ownership type, location, and length of stay for an inpatient visit. These features are derived from publicly available information for all Medicare hospitals and importantly are *not* included in the data used for detection.

Understanding the factors that drive outlier provider behavior is crucial for improving the health care sector. Extensive policy reforms seek to shape the structure of the health care market, increasing regulations on providers deemed to be harmful or inefficient. By characterizing the nature of hospitals deemed suspicious by our metrics, we hope to contribute to the ongoing literature that evaluates how various interventions – for example, those targeting for-profit care – can affect fraudulent behavior.



(A) Hospital rating (scale 1 to 5)

(B) Hospital ownership type: Private, Govt, Non-profit

(C) Hospital location

FIGURE 5.16: Comparison of distributions over categorical covariates for Outlier hospitals and All hospitals

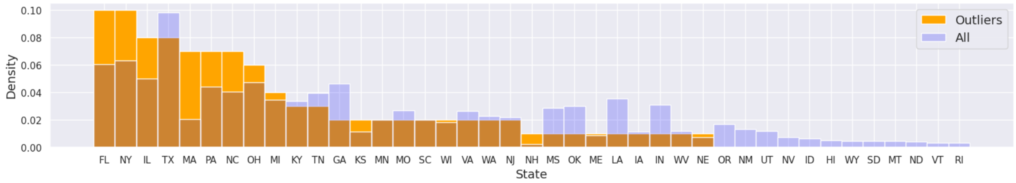
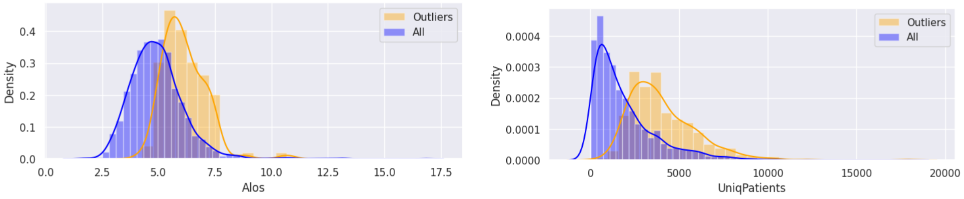


FIGURE 5.17: Comparison of distributions over 'State' for Outlier hospitals and All hospitals

Figure 5.16 shows the normalized histograms for categorical covariates – hospital rating, ownership type, location type – for the providers. We compare the distributions for the top 5% of suspicious providers in aggregate outlier ranking with those over all the providers. The idea is that, assuming fraud is rare, an investigator with limited resources would examine only the top portion of the ranked providers.

We observe in Figure 5.16a that histograms for Hospital Overall Rating largely overlap, indicating that outlier hospitals and all the hospitals are sampled from a similar underlying distribution, i.e. hospital rating is *not* a strong predictor of outlier status. On the other hand, Figures 5.16b and 5.16c show that our top ranked fraudulent providers are more likely to be private (for-profit) urban hospitals, and less likely to be non-urban, government-owned or nonprofit hospitals. This observation agrees with the literature on for-profit care, which has found distortions from this ownership structure (Gupta, Howell, Yannelis, and Gupta, 2021).

Figure 5.17 compares the distributions over states where a hospital is located. Outlier providers are more likely to be from states Florida, New York, Illinois, and Massachusetts, and less likely to be from Texas and Georgia. This is also corroborated by the DOJ cases, where about 15% of the named hospitals are based in Florida.



(A) average Length of Stay (aLOS)

(B) #Unique patients served

FIGURE 5.18: Comparison of distributions over numeric covariates for Outlier hospitals and All hospitals

Figure 5.18 compares the distributions across average length of stay and number of unique patients served. Ranked outlier hospitals keep inpatients longer as compared to other hospitals. This could be to justify the usage of costlier DRGs, or driven by ranked outlier hospitals receiving sicker patients; however, our metrics control for patient health characteristics. Additionally, top ranked fraudulent providers serve more unique patients on average. Since a large fraction of our top ranked providers are also named by the DOJ, it may indicate that a greater number of unique patients may provide more opportunity for perturbations in diagnosis coding resulting in higher reimbursements, or it could reflect the fact that our outliers are largely urban hospitals.

5.10 Discussion

The unsupervised ensemble method introduced in this work provides a new data-driven approach to identifying health care fraud using massive claims data. Our approach uses different data modalities – including patient medical history, provider coding patterns, and provider spending – to detect anomalous behavior consistent with fraud and abuse. Besides detection, the methodology offers interpretability, where qualitative case studies of our results based on model-specific explanations pinpoint specific ICD and DRG codes associated with excess spending at a provider. Finally, our method allows us to characterize the types of providers most likely to be ranked as suspicious, which may be useful for guiding anti-fraud policy more broadly.

Our method substantially outperforms baseline algorithms. We combine evidence from multiple unsupervised outlier detection algorithms that use different types of global and local analysis to create a final ranking of suspiciousness. While only 1 in 20 hospitals are named in our ground truth data as fraudulent, 21 of our top ranked 50 hospitals are in the same corpus, achieving an 8-fold improvement in detection rate.

Our data come from Medicare, the largest federal health care program, and we validate our method quantitatively using Department of Justice (DOJ) press releases that name hospitals. Medicare spends over a hundred billion dollars per year on hospitalizations, and the federal government has limited enforcement capacity. We believe our findings are *per se* interesting, because they help pinpoint fraud by private firms against the government in a way that could be used to improve public spending.

Our method has natural extensions beyond Medicare and beyond hospitalizations. We believe that the same method will prove useful in detecting fraud against private insurers, who face many of the same issues. Private insurers spend hundreds of billions of dollars per year on reimbursing care, and even small shares of fraud can be very expensive. Our detection algorithm can be used to guide auditing by identifying which providers are committing the most egregious behavior. Because our method explains which patterns drive the detection, it can facilitate auditing once a provider is selected by allowing an investigator to focus on certain billing codes and types of care. Our method also has a natural extension to Medicaid, the federal-state partnered low-income subsidy program, which spends an additional $400 Billion per year on health care. With health care spending at 19.7% of US GDP Centers for Medicare & Medicaid Services, 2022, tools for detecting health care fraud can find wide-ranging use.

Chapter 6

Early Prediction of Health Outcomes

Chapter based on: Shubhranshu Shekhar, Dhivya Eswaran, Bryan Hooi, Jonathan Elmer, Christos Faloutsos, and Leman Akoglu (2023). "Benefit-aware early prediction of health outcomes on multivariate eeg time series". In: *Journal of Biomedical Informatics* 139, p. 104296.

6.1 Introduction

Early decision making is critical in a variety of application domains. In medicine, earliness in prediction of health outcomes for patients in ICU (intensive care unit) allows the hospital to redistribute their resources (e.g., ICU bed-time, physician time, etc.) to in-need patients, and potentially achieve better health outcomes overall within the same amount of time, which is also a goal of value-based healthcare (Gray, 2017). Of course, another critical factor in play is the accuracy of such predictions. Hastily but incorrectly predicting unfavorable health outcome (e.g withdrawal of life-sustaining therapies) could hinder equitable decision making in the ICU, and may also expose hospitals to very costly lawsuits.

Clinical Problem Setting Consider resuscitated patients who generally survive to ICU admission in comatose state. For the first several days, it is hard to distinguish patients who will awaken, leading to favorable recovery, from others. Current clinical tests can not ascertain recovery for the initial $24-48$ hours. Therefore, many patients receive aggressive care, however, later learned to have no chance of recovery. As a result, families face tremendous financial and emotional burden with no benefit to patients from such treatments. Early and accurate prognostication in ICU could save lives, allow resource redistribution, avoid long and difficult care, and provide families respite from prolonged uncertainty. A clinician considers patient history, demographics, family support etc. in addition to large amounts of real-time sensor information for taking a decision. Our work is motivated by this real-world application that would help in alleviating the information overload on clinicians and aid them in early and accurate decision making in ICU.

Our Approach. As suggested by the application, the real-time prediction problem necessitates modeling of two competing goals: earliness and accuracy—competing since observing for a longer time, while cuts back from earliness, provides more information (i.e., data) that can help achieve better predictive accuracy. Besides the

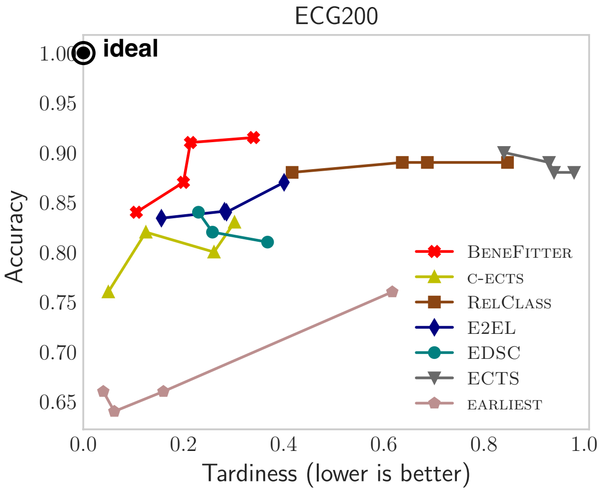


FIGURE 6.1: BENEFITTER **wins**: Note that BENEFITTER (in red) is on the Pareto front (Lotov, Bushenkov, and Kamenev, 2013) of accuracy-vs.-tardiness trade-off on ECG dataset. Each point represents evaluation of a method for a setting of hyper-parameters controlling the trade-off.

earliness-accuracy trade-off, the prediction of health outcomes on electroencephalography (EEG) recordings of ICU patients brings additional challenges. A large number (107) of EEG signal measurements are collected from multiple electrodes constituting high dimensional multivariate time series (our data is 900 GB on disk). Moreover, the series in data can be of various lengths because patients might not survive or be discharged after varying length of stay at the ICU.

To this end, we directly integrate a cost/benefit framework to our proposed solution, BENEFITTER, toward jointly optimizing prediction accuracy and earliness. We do not tackle an explicit multi-objective optimization but rather directly model a *unified* target that infuses those goals. BENEFITTER addresses the additional challenges such as handling (*i*) multi-variate and (*ii*) variable-length signals (i.e., time series), (*iii*) space-efficient modeling, (*iv*) scalable training, and (*v*) constant-time prediction.

We summarize our contributions as follows.

- **Novel, cost-aware problem formulation**: We propose BENEFITTER, which infuses the incurred *savings/gains* $S(t)$ from an early prediction at time t, as well as the *cost* M from each misclassification into a unified target called `benefit` $= S(t) - M$. Unifying these two quantities allows us to directly estimate a *single* target, i.e., `benefit`, and importantly dictates BENEFITTER exactly *when* to output a prediction: whenever estimated `benefit` becomes positive.
- **Efficiency and speed**: The training time for BENEFITTER is linear in the number of input sequences, and it can operate under a streaming setting to update its decision based on incoming observations. Unlike existing work that train a collection of prediction models for each $t = 1, 2, \ldots$ (Dachraoui, Bondu, and Cornuéjols, 2015; Tavenard and Malinowski, 2016; Mori, Mendiburu, Dasgupta, and Lozano, 2017), BENEFITTER employs a *single* model for each possible outcome, resulting in much greater space-efficiency.
- **Multi-variate and multi-length time-series**: Due to hundreds of measurements from EEG signals collected from patients with variable length stays at the ICU, BENEFITTER is designed to handle multiple time sequences, of varying length, which is a more general setting.
- **Effectiveness on real-world data**: We apply BENEFITTER on real-world (a) multi-variate health care data (our main motivating application for this work

is predicting survival/death of cardiac-arrest patients based on their EEG measurements at the ICU), and (b) other 11 benchmark datasets pertaining to various early prediction tasks. On ICU application, BENEFITTER can make decisions with up to 2× early (time-savings) as compared to competitors while achieving equal or better performance on accuracy metrics. Similarly, on benchmark datasets, BENEFITTER provides the best spectrum for trading-off accuracy and earliness (e.g. see Figure 6.1).

6.2 Data and Problem Setting

6.2.1 Data Description

Our use case data are obtained from 725 comatose patients who are resuscitated from cardiac arrest and underwent post-arrest EEG monitoring at a single academic medical center between years 2010–2018.

The raw EEG data are recorded at 256 Hz from 22 scalp electrodes; 11 electrodes in each hemisphere of the brain placed according to 10–20 International System of Electrode Placement (Morley, Hill, and Kaditis, 2016). The raw data is then used to collect quantitative EEG (qEEG) features (LaRoche and Haider, 2018) at an interval of ten seconds that amounts to about 900 GB of disk space for 725 patients. For our experiments, we selected 107 qEEG signals that physicians find informative from the electrode measurements corresponding to different brain regions. The 107-dimensional qEEG measurements from different electrodes on both left and right hemisphere, including the amplitude-integrated EEG (aEEG), burst suppression ratio (SR), asymmetry, and rhythmicity, form our multivariate time-series for analysis. We also record qEEG for each hemisphere as average of qEEG features from 11 electrodes on the given hemisphere.

As part of preprocessing, we normalize the qEEG features in a range $[0, 1]$. The EEG data contains artifacts caused due to variety of informative (e.g. the patient wakes up) or arbitrary (e.g. device unplugged/unavailability of devices) reason. This results in missing values, abnormally high or zero measurements. We filter out the zero measurements, typically, appearing towards the end of each sequence as well as abnormally high signal values at the beginning of each time series from the patient records. The zero measurements towards the end appear because of the disconnection. Similarly, abnormally high readings at the start appear when a patient is being plugged for measurements. The missing values are imputed through linear interpolation.

In this dataset, 225 patients ($\approx$ 31%) out of total 725 patients survived i.e. woke up from coma. Since the length of stay in ICU depends on each individual patient, the dataset contains EEG records of length 24–96 hours. To extensively evaluate our proposed approach, we create 3 versions of the dataset by median sampling (Justusson, 1981) the sequences at one hour, 30 minutes and 10 minutes intervals (as summarized in §6.5, Table 6.4).

6.2.2 Notation

A multi-variate time-series dataset is denoted as $\mathcal{X} = \{(\mathbf{X}_i, l_i)\}_{i=1}^{n}$, consisting of observations and labels for n instances. Each instance i has a label $l_i \in \{1, \ldots C\}$ where C is the number of labels or classes.[1] For example, each possible health outcome at

[1] We use the terms label and class interchangeably throughout the chapter.

the ICU is depicted by a class label as *survival* or *death*. The sequence of observations is given by $\mathbf{X}_i = \{\mathbf{X}_{i1}, \ldots, \mathbf{X}_{it}, \ldots, \mathbf{X}_{iL_i}\}$ for L_i equi-distant time ticks. Here, L_i is the length of time-series i and varies from one instance to another in the general case. It is noteworthy that our proposed BENEFITTER can effectively handle variable-length series in a dataset, whereas most existing early prediction techniques are limited to fixed length time-series, where $L_i = L$ for all $i \in [n]$. Each observation $\mathbf{X}_{it} \in \mathbb{R}^d$ is a vector of d real-valued measurements, where d is the number of variables or signals. We denote $\mathbf{X}_i$'s observations from the start until time tick t by $\mathbf{X}_{i[1:t]}$.

6.2.3 Problem Statement

Early classification of time series seeks to generate a prediction for input sequence $\mathbf{X}$ based on $\mathbf{X}_{[1:t]}$ such that t is small and $\mathbf{X}_{[1:t]}$ contains enough information for an accurate prediction. Formally,

Problem 2 (Early classification). *Given a set of labeled multivariate time series $\mathcal{X} = \{(\mathbf{X}_i, l_i)\}_{i=1}^n$, learn a function $\mathcal{F}_\theta(\cdot)$ which assigns label $\hat{l}$ to a given time series $\mathbf{X}_{[1:t]}$ i.e. $\mathcal{F}_\theta(\mathbf{X}_{[1:t]}) \mapsto \hat{l}$ such that t is small.*

Challenges The challenges in early classification are two-fold: domain-specific and task-specific, discussed as follows.

- *Domain-specific:* Data preprocessing is non-trivial since raw EEG data includes various biological and environmental artifacts. Observations arrive incrementally across multiple signals where the characteristics that are indicative of class labels may occur at different times across signals which makes it difficult to find a decision time to output a label. Moreover, each time series instance can be of different length due to varying length of stay of patients at the ICU which requires careful handling.
- *Task-specific:* Accuracy and earliness of prediction are competing objectives (as noted above) since observing for a longer time, while cuts back from earliness, provides more signals that is likely to yield better predictive performance.

In this work, we propose BENEFITTER (see §6.4) that addresses all the aforementioned challenges.

6.3 Background and Related Work

Time series data has been well-studied in the literature for event detection (Weiss and Hirsh, 1998), anomaly detection (Keogh, Lin, Fu, and Herle, 2006), similarity search (Yeh et al., 2018), visualization (Gao, Li, Li, Lin, and Rangwala, 2017) and more. A comprehensive treatment is provided in (Ralanamahatana, Lin, Gunopulos, Keogh, Vlachos, and Das, 2005). Traditional techniques for time series classification rely on observing whole time series before prediction of its class label. Traditional time series classification draws from a large number of different techniques including near neighbor similarity, interval and phase based feature extraction, recurring short-pattern mining, signal processing, and more recently deep learning based models (Bagnall, Lines, Bostrom, Large, and Keogh, 2017; Susto, Cenedese, and Terzi, 2018; Li, Bissyande, Klein, and Traon, 2016; Ismail Fawaz, Forestier, Weber, Idoumghar, and Muller, 2019). Here, we only survey work most relevant to *early* time-series classification.

Early Classification. The initial mention of early classification of time-series dates back to early 2000s (Rodríguez, Alonso, and Boström, 2001; Bregón, Simón, Rodríguez, Alonso, Pulido, and Moro, 2005) where the authors consider the value in classifying prefixes of time sequences. However, it was formulated as a concrete learning problem only recently (Xing, Pei, Dong, and Yu, 2008; Xing, Pei, and Philip, 2012). Xing, Pei, Dong, and Yu, 2008 mine a set of sequential classification rules and formulate an early-prediction utility measure to select the features and rules to be used in early classification. Later they extend their work to a nearest-neighbor based time-series classifier approach to wait until a certain level of confidence is reached before outputting a decision (Xing, Pei, and Philip, 2012). Parrish, Anderson, Gupta, and Hsiao, 2013 delay the decision until a reliability measure indicates that the decision based on the prefix of time-series is likely to match that based on the whole time-series. Xing, Pei, Yu, and Wang, 2011 advocate the use of interpretable features called shapelets (Ye and Keogh, 2009) which have a high discriminatory power as well as occur earlier in the time-series. Ghalwash and Obradovic, 2012 extend this work to incorporate a notion of uncertainty associated with the decision. Hatami and Chira, 2013 train an ensemble of classifiers along with an agreement index between the individual classifiers such that a decision is made when the agreement index exceeds a certain threshold. As such, none of these methods explicitly optimize for the trade-off between earliness and accuracy.

Dachraoui, Bondu, and Cornuéjols, 2015 propose to address this limitation and introduce an adaptive and non-myopic approach which outputs a label when the projected cost of delaying the decision until a later time is higher than the current cost of early classification. The projected cost is computed from a clustering of training data coupled with nearest neighbor matching. Tavenard and Malinowski, 2016 improve upon Dachraoui, Bondu, and Cornuéjols, 2015 by eliminating the need for data clustering by formulating the decision to delay or not to delay as a classification problem. Mori, Mendiburu, Dasgupta, and Lozano, 2017 take a two-step approach; where in the first step classifiers are learned to maximize accuracy, and in the second step, an explicit cost function based on accuracy and earliness is used to define a stopping rule for outputting a decision. Schäfer and Leser, 2020, instead, utilize reliability of predicted label as stopping rule for outputting a decision. However, these methods require a classification-only phase followed by optimizing for trade-off between earliness and accuracy. Recently, Hartvigsen, Sen, Kong, and Rundensteiner, 2019 employ recurrent neural network (RNN) based discriminator for classification paired with a reinforcement learning task to learn halting policy. The closest in spirit to our work is the recently proposed end-to-end learning framework for early classification (Rußwurm, Lefèvre, Courty, Emonet, Körner, and Tavenard, 2019) that employs RNNs. They use a cost function similar to (Mori, Mendiburu, Dasgupta, and Lozano, 2017) in a fine-tuning framework to learn a classifier and a stopping rule based on RNN embeddings for partial sequences.

Our proposed BENEFITTER is a substantial improvement over all the above prior work on early classification of time series along a number of fronts, as summarized in Table 6.1. BENEFITTER jointly optimizes for earliness and accuracy using a cost-aware `benefit` function. It seamlessly handles multi-variate and varying-length time-series and moreover, leads to explainable early predictions, which is important in high-stakes domains like health care.

Value-based Healthcare Value-based healthcare focuses on maximizing the benefits of provided care. This can be achieved by distributing resources to ensure in-need patients receive care at the right time (Gray, 2017; Bae, 2015). Resource allocation is particularly important in the context of limited provisions (e.g. occupancy rate in life-sustaining therapies) for meeting the medical demand. Prior works (Lee and Porter, 2013; Traoré, Zacharewicz, Duboz, and Zeigler, 2019; Hillary, Justin, Bharat,

TABLE 6.1: Qualitative comparison with prior work. '?' means that the respective method, even though does not exhibit the corresponding property originally, can possibly be extended to handle it.

Property / Method	ECTS (Xing, Pei, and Philip, 2012)	C-ECTS (Dachraoui, Bondu, and Cornuéjols, 2015; Tavenard and Malinowski, 2016)	EDSC (Xing, Pei, Yu, and Wang, 2011)	M-EDSC (Ghalwash and Obradovic, 2012)	RelClass (Parrish, Anderson, Gupta, and Hsiao, 2013)	E2EL (Rußwurm, Lefèvre, Courty, Emonet, Körner, and Tavenard, 2019)	EARLIEST (Hartvigsen, Sen, Kong, and Rundensteiner, 2019)	BENEFITTER
Jointly optimize earliness & accuracy						✓	✓	✓
Distance metric agnostic					✓	✓	✓	✓
Multivariate				✓		✓	✓	✓
Constant decision time						✓	✓	✓
Handles variable length series	✓$^{?}$	✓$^{?}$	✓	✓	✓$^{?}$	✓	✓$^{?}$	✓
Explainable model	✓		✓	✓				✓
Explainable hyper-parameter								✓
Cost aware		✓				✓		✓

and Jitendra, 2016) consider healthcare value in resource allocation from policy perspective. Our proposed BENEFITTER complements the value-based healthcare, and in this work, through `benefit` function, our framework provides clinicians with tools to assist in decision making that aims to achieve better health outcomes overall with limited hospital resources.

6.4 BENEFITTER: Proposed Method

6.4.1 Modeling Benefit

How should an early prediction system trade-off accuracy vs. earliness? In many real-world settings, there is natural misclassification *cost*, denoted M, associated with an inaccurate prediction and certain *savings*, denoted $S(t)$, obtained from early decision-making. We propose to construct a single variable called `benefit` which captures the overall value (savings minus cost) of outputting a certain decision (i.e., label) at a certain time t, given as

$$\texttt{benefit} = S(t) - M \tag{6.1}$$

We directly incorporate `benefit` into our model and leverage it in deciding *when* to output a decision; when the estimate is positive.

Outcome vs. Type Classification

There are two subtly different problem settings that arise in time-series classification that are worth distinguishing between.

- *Outcome classification:* Here, the labels of time-series encode the observed outcome *at the end* of the monitoring period of each instance. Our motivating examples from predictive health care and system maintenance fall into this category. Typically, there are two outcomes: *favorable* (e.g., *survival* or *no-failure*) and *unfavorable* (e.g., *death* or *catastrophic-failure*); and we are interested in knowing when an unfavorable outcome is anticipated. In such cases, predicting an early favorable outcome does not incur any change in course of action, and hence does not lead to any discernible savings or costs. For example, in our ICU application, a model predicting *survive* (as opposed to *death*) simply suggests to the physicians that the patient would survive *provided they continue with their regular procedures of treatment*. That is because $l = survive$ labels we observe in the data are *at the end* of the observed period *only after all regular course of action have been conducted*. In contrast, $l = death$ instances have died *despite* the treatments.

In outcome classification, predicting the favorable class simply corresponds to the 'default state' and therefore we model `benefit` and actively make predictions only for the unfavorable class.

- *Type classification:* Here, the time-series labels capture the underlying process that gives rise to the sequence of observations. In other words, the class labels are *prior* to the time-series observations. The standard time-series early classification benchmark datasets fall into this category. Examples include predicting the type of a bird from audio recordings or the type of a flying insect (e.g., a mosquito) from their wingbeat frequencies (Batista, Keogh, Mafra-Neto, and Rowton, 2011). Here, prediction of *any* label for a time-series at a given time has an associated cost in case of misclassification (e.g., inaccurate density estimates of birds/mosquitoes) as well

TABLE 6.2: Benefit model for ICU outcome prediction.

		Predicted $\widehat{l_i}$	
		survival	*death*
Actual l_i	*survival*	0	$(L_i - t)s - M$
	death	0	$(L_i - t)s$

as potential savings for earliness (e.g., battery life of sensors). In type classification, we separately model `benefit` for each class.

Benefit Modeling for Outcome Classification

Consider the 2-class problem that arises in predictive health care of ICU patients and predictive maintenance of systems. Without loss of generality, let us denote by $l = 0$ the *survival* class where the patient is discharged alive from the ICU at the end of their stay; and let $l = 1$ denote the *death* class where the patient is deceased.

As discussed previously, $l = 0$ corresponds to the 'default state' in which regular operations are continued. Therefore, predicting *survival* would not incur any time savings or misclassification cost. In contrast, predicting *death* would suggest the clinician to intervene to optimize quality of life for the patient. In case of an accurate prediction, say at time t, earliness would bring savings (e.g., ICU bed-time), denoted $S(t)$. Here we use a linear function of time for savings on accurately predicting *death* for a patient i at time t, specifically

$$S(t) = (L_i - t)s \tag{6.2}$$

where s denotes the value of savings per unit time.[2] On the other hand, an inaccurate *death* flag at t, while comes with the same savings, would also incur a misclassification cost M (e.g., a lawsuit).

All in all, the benefit model for the ICU scenario is given as in Table 6.2, reflecting the relative savings minus the misclassification cost for each decision at time t on time-series instance i. As we will detail later in §6.4.3, the main idea behind BeneFitter is to learn a single *regressor* model for the *death* class, estimating the corresponding `benefit` at each time tick t.

Specifying s and M. Here, we make the following two remarks. First, unit-time savings s and misclassification cost M are value estimates that are dictated by the specific application. For our ICU case, for example, we could use $s = \$4,000$ value per unit ICU time, and $M = \$500,000$ expected cost per lawsuit. Note that s and M are domain-specific explainable quantities. Second, the benefit model is most likely to differ from application to application. For example in predictive system maintenance, savings and cost would have different semantics, assuming that early prediction of failure implies a renewal of all equipment. In that case, an early and accurate failure prediction would incur savings from costs of a complete system halt, but also loss of equipment lifetime value due to early replacement plus the replacement costs. On the other hand, early but inaccurate prediction (i.e., a false alarm) would simply incur unnecessary replacement costs plus the loss of equipment lifetime value due to early replacement.

Our goal is to set up a general prediction framework that explicitly models `benefit` based on incurred savings and costs associated with individual decisions, whereas the scope of specifying those savings and costs are left to the practitioner.

[2] Note that BeneFitter is flexible enough to accommodate any other function of time, including nonlinear ones, as the savings function $S(t)$.

We argue that each real-world task should strive to explicitly model `benefit`, where earliness and accuracy of predictions translate to real-world value. In cases where the prediction task is isolated from its real-world use (e.g., benchmark datasets), one could set both $s = M = 1$ for unit savings per unit time earliness and unit misclassification cost per incorrect decision. In those cases where M is not tied to a specific real-world value, it can be used as a “knob” (i.e., hyperparameter) for trading off accuracy with earliness; where, fixing $s = 1$, a larger M nudges BENEFITTER to avoid misclassifications toward higher accuracy at the expense of delayed predictions and vice versa.

Benefit Modeling for Type Classification

Compared to outcome prediction where observations give rise to the labels, in type classification problems the labels give rise to the observations. Without a default class, predictions come with associated savings and cost for each class.

TABLE 6.3: Benefit model for general two-class type prediction.

		Predicted $\widehat{l_i}$	
		type-1	*type*-2
Actual l_i	*type*-1	$(L_i - t)s$	$(L_i - t)s - M_{12}$
	type-2	$(L_i - t)s - M_{21}$	$(L_i - t)s$

Consider the 2-class setting of predicting an insect’s type from wingbeat frequencies. An example benefit model is illustrated in Table 6.3, s capturing the value of battery-life savings per unit time and M depicting the cost of misclassifying one insect as the other. Note that in general, misclassification cost need not be symmetric among the classes.

For type classification problems, we train a total of C `benefit` prediction models, one for each class. Since misclassification costs are already incorporated into `benefit`, we train each (regression) model independently which allows for full parallelism.

6.4.2 Online Decision-making using Benefit

Next we present how to employ BENEFITTER in decision making in real time. Suppose we have trained our model that produces `benefit` estimates per class for a new time-series instance in an online fashion. *How* and *when* should we output predictions?

Thanks to our benefit modeling, the decision-making is quite natural and intuitive: BENEFITTER makes a prediction only when the estimated `benefit` becomes *positive* for a certain class and outputs the label of that class as its prediction i.e. for our ICU scenario the predicted label $\hat{l}$ is given as

$$\hat{l} = \begin{cases} \text{unfavorable}, & \text{if } \texttt{benefit} > 0 \\ \text{favorable, i.e. no action}, & \text{otherwise.} \end{cases}$$

For illustration, in Fig. 6.2 we show `benefit` estimates over time for an input series where $t = 15$ corresponds to decision time.

Note that in some cases BENEFITTER may restrain from making any prediction for the entire duration L of a test instance, that is when estimated `benefit` never goes above zero. For outcome classification tasks, such a case is simply registered as default-class prediction and its prediction time is recorded as L. For the ICU scenario, a non-prediction is where no *death* flag is raised, suggesting survival and regular course of action. For type classification tasks, in contrast, a non-prediction suggests "waiting for more data" which, at the end of the observation period, simply implies insufficient evidence for any class. We refer to those as un-classified test instances. Note that BENEFITTER is different from existing prediction models that always produce a prediction, where un-classified instances may be of independent interest to domain experts in the form of outliers, noisy instances, etc.

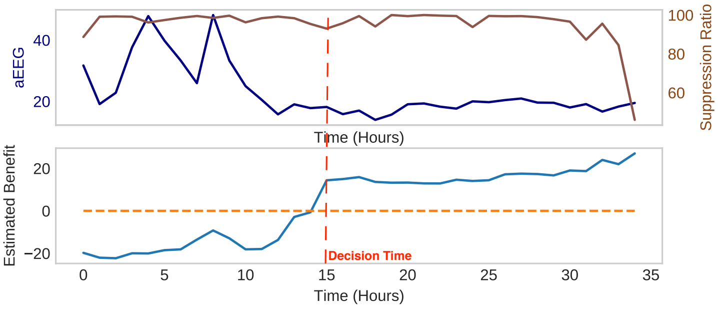


FIGURE 6.2: Benefit estimate over time for a patient from EEG dataset with true $l = 1$ (i.e. *death*). We show two out of all 107 signals used by BENEFITTER: amplitude of EEG (aEEG) and suppression ratio (i.e. fraction of flat EEG epochs).

6.4.3 Predicting Benefit

For each time-series i, we aim to predict the `benefit` at every time tick t, denoted as b_{it}. Consider the outcome classification problem, where we are to train one regressor model for the non-default class, say $no - survival$. For each training series i for which $l_i = 0$ (i.e., default class), `benefit` of predicting *death* at t is $b_{it} = (L_i - t)s - M$. Similarly for training series for which $l_i = 1$ (i.e., *death*), $b_{it} = (L_i - t)s$. (See Table 6.2.) To this end, we create training samples of the form $\{(\mathbf{X}_{i[1:t]}, b_{it})\}_{t=1}^{L_i}$ per instance i. Note that the problem becomes a regression task. For type classification problems, we train a separate regression model per class with the corresponding b_{it} values. (See Table 6.3.)

Model. We set up the task of `benefit` prediction as a sequence regression problem. We require BENEFITTER to ingest multi-variate and variable-length input to estimate `benefit`. We investigate the use of Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997), a variant of recurrent neural networks (RNN), for the sequence (time-series) regression since their recursive formulation allows LSTMs to handle multi-variate variable-length inputs naturally. The recurrent formulation of LSTMs is useful for BENEFITTER to enable real-time predictions when new observations arrive one at a time.

Attention. The recurrent networks usually find it hard to focus on to relevant information in long input sequences. For example, an EEG pattern in the beginning of a sequence may contain useful information about the patient's outcome, however the lossy representation of LSTM would forget it. This issue is mostly encountered in longer input sequences (Luong, Pham, and Manning, 2015). The underlying idea of *attention* (Vaswani et al., 2017) is to learn a context that captures the relevant information from the parts of the sequence to help predict the target. For a sequence of length L, given the hidden state $\mathbf{h}_t$ from LSTM and the context vector $\boldsymbol{c}$, the attention

TABLE 6.4: Summary of the datasets used in this work.

Dataset	Train	Test	Classes	Length	Dimension
EEG-ICU Hour	507	218	2	24–96	107
EEG-ICU 30 Min	507	218	2	48–192	107
EEG-ICU 10 Min	507	218	2	144–576	107
ECG200	100	100	2	96	1
ItalyPowerDemand	67	1029	2	24	1
GunPoint	50	150	2	150	1
TwoLeadECG	23	1139	2	82	1
Wafer	1000	6062	2	152	1
ECGFiveDays	23	861	2	136	1
MoteStrain	20	1252	2	84	1
Coffee	28	28	2	286	1
Yoga	300	3000	2	426	1
SonyAIBO	20	601	2	70	1
Endomondo	99754	42751	2	450	2

step in BENEFITTER combines the information from both vectors to produce a final attention based hidden state as described below:

$$\alpha_t = \frac{\exp\left(\boldsymbol{c}_L \cdot \mathbf{h}_t\right)}{\sum_t \exp\left(\boldsymbol{c}_L \cdot \mathbf{h}_t\right)}; \quad \boldsymbol{c} = \sum_{t=1}^{L} \alpha_t \mathbf{h}_t \tag{6.3}$$

$$\mathbf{h}_{\text{attn}} = \sigma(\mathbf{W}_a[concat(\boldsymbol{c}, \boldsymbol{c}_L)]) \tag{6.4}$$

where $\boldsymbol{c}_L$ is the memory state of the cell at the last time step L, $\mathbf{h}_t$ is the hidden state output of the LSTM at time t, $\mathbf{h}_{\text{attn}}$ is the attention based hidden state, $\sigma(\cdot)$ is the non-linear transformation, and $\mathbf{W}_a$ is the parameter. Intuitively, the attention weights α_t allows the model to learn to focus on specific parts of the input sequence for the task of regression. The `benefit` prediction is given by a single layer neural network such that $\hat{b}_L = \mathbf{h}_{\text{attn}}\mathbf{w} + w_0$ where $\mathbf{w}$ and w_0 are parameters of the linear layer.

BENEFITTER is used in real life decision making, where the attention mechanism could help an expert by highlighting the relevant information that guided the model to output a decision. We present model implementation details and list of tunable parameters in §6.5.

6.5 Experiments

We evaluate our method through extensive experiments on a set of benchmark datasets and on a set of datasets from real-world use cases. We next provide the details of the datasets and the experimental setup, followed by results.

6.5.1 Dataset Description

We apply BENEFITTER on our EEG-ICU datasets (see §6.2.1), and on 11 public benchmark datasets from diverse domains with varying dimensionality, length, and scale. Table 6.4 provides a summary of the datasets used in evaluation. Note that EEG-ICU datasets are variable-length, but benchmarks often used in the literature are not. Detailed description of public datasets are as follows.

- **Benchmark Datasets.** Our benchmark datasets consist of 10 two-class time-series classification datasets from the UCR repository Chen et al., 2015. The datasets cover diverse domains and have diverse range of series length. The UCR archive provides the train/test split for each of these datasets, which we retain in our experiments.

- **Endomondo Dataset.** Endomondo is a social fitness app that tracks numerous fitness attributes of the users. We use the web-scale Endomondo dataset Ni, Muhlstein, and McAuley, 2019 (See Table 6.4) for the early activity prediction task. The data includes various measurements such as heart rate, altitude and speed, along with contextual data such as user id and activity type. For the task of early activity prediction, we use heart rate and altitude signals for early prediction of the type of activity, specifically biking vs. running. (Note that we leave out signals like speed and its derivatives which make the classification task too easy.)

6.5.2 Experimental Setup

Baselines. We compare BENEFITTER to the following six early time-series classification methods (also see Table 6.1), which broadly belong to one of the two types – Machine Learning (ML) (Mitchell and Mitchell, 1997) based and Deep Learning (DL) (LeCun, Bengio, and Hinton, 2015) based methods. ML based methods:

1. ECTS: Early Classification on Time Series (Xing, Pei, and Philip, 2012) uses *minimum prediction length* (MPL) and makes predictions if the MPL of the top nearest neighbor (1-NN) is greater than the length of the test series.
2. EDSC: Early Distinctive Shapelet Classification (Xing, Pei, Yu, and Wang, 2011) extracts local shapelets for classification that are ranked based on the utility score incorporating earliness and accuracy. Multivariate extention of EDSC (M-EDSC) (Ghalwash and Obradovic, 2012) provides a utility function that can incorporate multi-dimensional series.
3. C-ECTS: Cost-aware ECTS (Dachraoui, Bondu, and Cornuéjols, 2015; Tavenard and Malinowski, 2016) trades-off between a misclassification cost and a cost of delaying the prediction, and estimates future expected cost at each time step to determine the optimal time instant to classify an incoming time series.
4. RelClass: Reliable Classification (Parrish, Anderson, Gupta, and Hsiao, 2013) uses a reliability measure to estimate the probability that the assigned label given incomplete data (at time step t) would be same as the label assigned given the complete data.

DL based methods:

5. E2EL: End-to-end Learning for Early Classification of Time Series (Rußwurm, Lefèvre, Courty, Emonet, Körner, and Tavenard, 2019) optimizes a joint cost function based on accuracy and earliness, and provides a framework to estimate a stopping probability based on the cost function.
6. EARLIEST: EARLIEST (Hartvigsen, Sen, Kong, and Rundensteiner, 2019) is a reinforcement learning based method that learns halting policy while optimizing for classification accuracy. It also optimizes a joint function that incorporates objectives for earliness and accuracy.

Hardware. Experiments are run on stock Linux server with 1024 GB RAM, 96 cores 2.10 GHz Intel Xeon CPU.

Model Training Details

We define the outcome prediction problem as a regression task on the `benefit`, as presented in §6.4. The training examples represent the sequences observed up to time t along with their corresponding expected `benefit` at time t. We then split the training examples (as mentioned in Table 6.4) to use 90% of the sequences for training the RNN model and remaining 10% for validation. We select our model parameters based on the evaluation on validation set. We have two sets of hyper-parameters: one corresponding to our benefit formulation that are s and M, and the other for the RNN model. The hyper-parameter grid for the RNN model includes the dimension of the hidden representation $\in \{16, 32\}$ and the learning rate $\eta \in \{0.01, 0.001\}$. For BENEFITTER, we fix s at 1 and vary M/s for model selection. The hyper-parameter grid for our benefit formulation is $M/s \in \{0.5, 1.0, 1.5, 2.0\} \times \max(\{L_i\}_{i=1}^{n})$, where L_i is the length of training series i. For the general multi-class problem we tune for an additional hyper-parameter Δ to predict the class label. We set $\Delta \in \{0.4, 0.5, 0.6, 0.7\}$ of the maximum difference between the expected benefit for the two-class labels for a given training series. BENEFITTER outputs a decision when the predicted benefit is positive and at least Δ higher compared to the predicted benefit of other labels.

For the learning task, we use the mean squared error loss function and Adam optimizer Kingma and Ba, 2014 for learning the parameters. The loss corresponding to class l is given by

$$\mathcal{J} = \frac{1}{|N_{train}|} \sum_{b_l \in \mathcal{B}_l} \sum_{t=1}^{L_i} (\hat{b}_{itl} - b_{itl})^2$$

where $|N_{train|}$ is the number of total count of time steps for all the sequences in the training set, L_i is the length of each sequence, and $\mathcal{B}_l$ denotes the expected benefit for the class l. We use Keras and Pytorch to implement our models.

6.5.3 Evaluation

We design our experiments to answer the following questions:

[Q1] Effectiveness: How effective is BENEFITTER at early prediction on time series compared to the baselines? What is the trade-off with respect to accuracy and earliness? How does the accuracy–earliness trade-off varies with respect to model parameters?

[Q2] Efficiency: How does the running-time of BENEFITTER scale w.r.t. the number of training instances? How fast is the online response time of BENEFITTER?

[Q3] Discoveries: Does BENEFITTER lead to interesting discoveries on real-world case studies?

[Q1] Effectiveness

We compare BENEFITTER to baselines on (1) patient outcome prediction, the main task that inspired our benefit formulation, (2) the activity prediction task on a web-scale dataset, as well as (3) the set of 10 two-class time-series classification datasets. The datasets for the first two tasks are multi-dimensional and variable-length that many of the baselines can not handle. Thus we compare BENEFITTER with E2EL and M-EDSC baselines that can work with such time-series sequences. Comparison with M-EDSC is limited to the smaller one-hour EEG dataset since it does not scale to larger datasets. In order to compare BENEFITTER to all other baselines, we conduct experiments on ten benchmark time-series datasets.

Patient Outcome Prediction. We compare BENEFITTER with the baseline E2EL on two competing criteria: performance (e.g. precision, accuracy) and earliness (tardiness – lower is better) of the decision. We report precision, recall, F1 score, accuracy, tardiness and the total benefit using each method when applied to the test set. EEG dataset is a high dimensional variable-length dataset for which most of the baselines are not applicable. In our experiments, we set the misclassification cost for each of the dataset variants as follows – $M/s = 100$ for dataset variant sampled at an hour, $M/s = 200$ for dataset sampled at 30 minutes, and $M/s = 600$ for dataset sampled at 10 minutes – based on average daily cost of ICU care and the lawsuit cost. For the baseline methods, we report the best results for the earliness-accuracy trade-off parameters. For the baseline methods, we select the best value of accuracy and earliness based on their Euclidean distance to ideal accuracy $= 1$ and ideal tardiness $= 0$.

Table 6.5 reports the evaluation against different performance metrics. Note that predicting 'default state' for a patient does not change the behavior of the system. However, predicting *death* (unfavorable outcome) may suggest clinician to intervene with alternative care. In such a decision setting, it is critical for the classifier to exhibit high precision. Our results indicate that BENEFITTER achieves a significantly higher precision (according to the micro-sign test (Yang, Liu, et al., 1999)) when compared to the baselines. On the other hand, a comparatively lower tardiness indicates that BENEFITTER requires conspicuously less number of observations on average to output a decision (no statistical test conducted for tardiness). We also compare the total `benefit` accrued for each method on the test set where BENEFITTER outperforms the competition. The results are consistent across the three datasets of varying granularity from hourly sampled data to 10 minute sampled data.

TABLE 6.5: Effectiveness of BENEFITTER on EEG datasets. * indicates significance at p-value $\leq$ 0.05 based on the micro-sign test (Yang, Liu, et al., 1999) for the performance metrics. No statistical test conducted for tardiness and total `benefit`.

		Prec.	*Recall*	*F1*	*Acc.*	*Tardiness*	`benefit`
EEG Hour	E2EL	0.70	**0.68**	0.69	0.79	1.0	-2600
	M-EDSC	0.69	0.65	0.67	0.78	**0.52**	2497
	BENEFITTER	**0.80***	**0.68**	**0.73***	**0.83***	0.64	2737
EEG 30Min	E2El	0.64	**0.67**	0.65	0.78	1.0	-4800
	BENEFITTER	**0.68***	0.66	**0.67***	**0.79**	**0.63**	5962
EEG 10Min	E2EL	0.73	**0.69**	0.71	0.82	0.86	-736
	BENEFITTER	**0.76***	**0.69**	**0.72**	**0.83**	**0.48**	18722

For hourly sampled set, we also compare our method to multivariate EDSC baseline (for the 30 min and 10 min EEG dataset M-EDSC does not scale). Though M-EDSC provides better earliness trade-off compared to other two methods, the precision of the outcomes is lowest which is not desirable in this decision setup. In Table 6.5, we indicate the significant results using * that is based on the comparison between BENEFITTER and E2EL.

Benchmark Prediction Tasks. To jointly evaluate the accuracy and earliness (tardiness – lower is better), we plot accuracy against the tardiness to compare the Pareto frontier for each of the competing methods over 10 different benchmark datasets. In Fig. 6.1 and Fig. 6.3, we show the accuracy and tardiness trade-off for 10 benchmark UCR datasets. Each point on the plot represents the model evaluation for a choice of trade-off parameters reported in Table 6.6. Note that BENEFITTER dominates the frontiers of all the baselines in accuracy vs tardiness on five of the datasets. Moreover, our method appears on the Pareto frontier for four out of the remaining five for at least one set of parameters.

TABLE 6.6: Earliness and accuracy trade-off parameters for each of the methods.

Method	Model Training Hyper-parameters
ECTS	support $\in \{0.1, 0.2, 0.4, 0.8\}$
EDSC	Chebyshev parameter $\in \{2.5, 3.0, 3.5\}$
C-ECTS	delay cost $\in \{0.0005, 0.001, 0.005, 0.01\}$
RelClass	reliability $\in 0.001, 0.1, 0.5, 0.9$
E2EL	earliness trade-off $\alpha \in \{0.6, 0.7, 0.8, 0.9\}$
EARLIEST	earliness trade-off $\lambda \in \{0.0, 0.05, 0.1, 0.15\}$
BENEFITTER	$M/s \in \{0.5, 1.0, 1.5, 2.0\} \times \max(\text{series length})$; for benchmark datsets

To further assess the methods, we report quantitative results in Table 6.7 in terms of accuracy at a given tolerance of tardiness. We define an acceptable level of tolerance $\in \{0.50, 0.75\}$ to indicate how much an application domain is indifferent to delay in decision up to the indicated level. For example a tolerance of 0.50 indicates that the evaluation of the decisions is done at $t = 0.50 \times L$, L is the maximum length of sequence, and any decision made up to $t = 0.50 \times L$ are considered for evaluation. In Fig. 6.3, we fix the x-axis at a particular tolerance and report the best accuracy to the left of the fixed tolerance in Table 6.7. The reported tolerance level indicates the average tolerance across the test time-series sequences. BENEFITTER outperforms the competition seven times out of ten for a tolerance level $= 0.50$ indicating that our method achieves best performance using only the half of observations. The remaining three times our method is second best among all the competing methods. Similarly for tolerance $= 0.75$, BENEFITTER is among the top two methods nine out of ten times.

Endomondo Activity Prediction. We run the experiments on full Endomondo dataset (a large scale dataset) to compare BENEFITTER with baselines E2EL and EARLIEST (other baselines do not scale) for one set of earliness-accuracy trade-off parameters. We, first, compare the three methods on a sampled dataset – with 1000 time series instances – evaluated for a choice of trade-off parameters. We select the parameters that yields a performance closest to *ideal* indicated. With the selected parameters, comparison of three methods on large-scale Endomondo activity prediction dataset are reported in Table 6.7 (last row). We report the accuracy of the three methods by fixing their tardiness at ≤ 0.5. Notice that the methods are comparable in terms of the prediction performance while using less than half the length of a sequence for outputting a decision.

The quantitative results suggest a way to choose the best classifier for a specified tolerance level for an application. In critical domains such as medical, or predictive maintenance a lower tolerance would be preferred to save cost. In such domains, BENEFITTER provides a clear choice for early decision making based on the benchmark dataset evaluation.

[Q2] Efficiency

Fig. 6.4 shows the scalability of BENEFITTER with the number of training time-series and number of observations per time series at test time. We use ECG200 dataset from UCR benchmark to report results on runtime.

Linear training time: We create ten datasets from the ECG200 dataset by retaining a fraction $\in \{0.1, 0.2, \ldots, 1.0\}$ of total number of training instances. For a fixed set

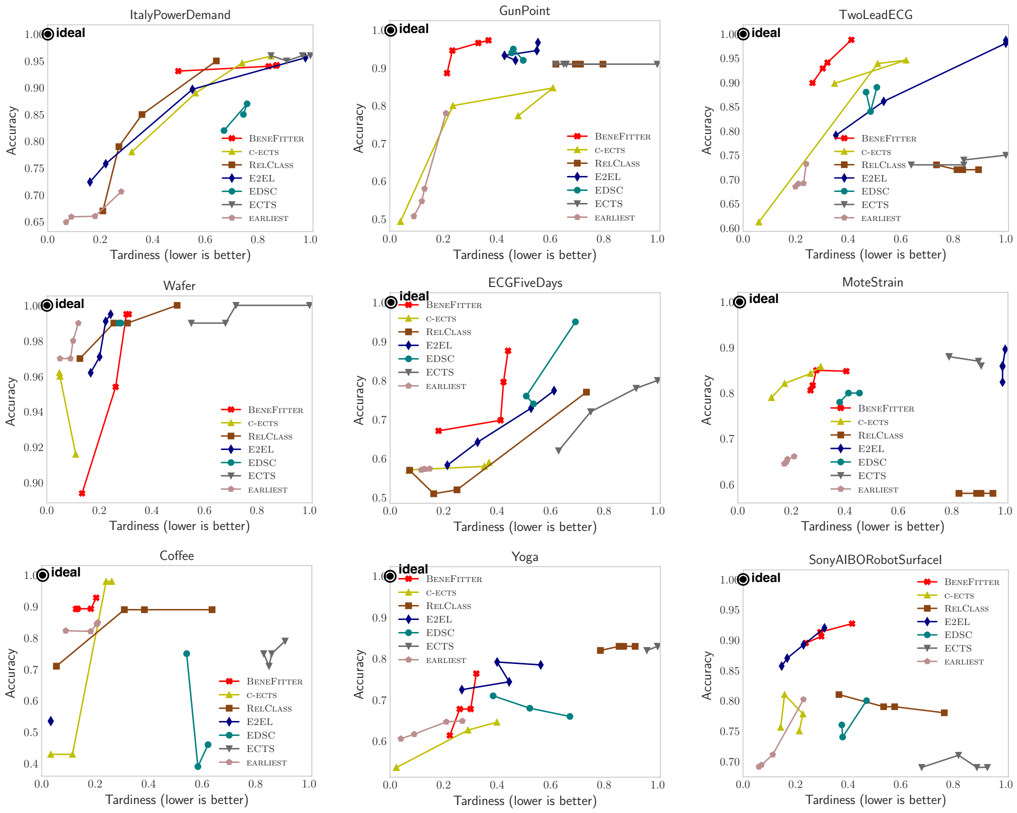


FIGURE 6.3: Comparison of methods based on accuracy versus tardiness trade-off for benchmark prediction tasks (Sec. §6.5.3).

of parameters, we train our model individually for each of the created datasets, and record train time for five independent runs. The average wall-clock running time is reported against the fraction of training sequences in Fig. 6.4 (left). The standard deviation is around the average wall-clock time is indicated by vertical bar in the plot. The points in the plot align in a straight line indicating that BENEFITTER scales linearly with number of sequences.

Constant test time: We now evaluate BENEFITTER runtime by varying the number of observations over time. For this experiment, we retain the hidden state of an input test sequence up to time $(t-1)$. When a new observation at time t arrives, we update the hidden state of the RNN cell using the new observation and compute the predicted benefit based on updated state. Fig. 6.4 (right) plots the wall-clock time against each new observation, and the density estimate is shown on the right. The time is averaged over test set examples. The plot indicates that we get real-time decision in constant time.

The efficiency of our model makes it suitable for deployment for real time prediction tasks.

[Q3] Discoveries

In this section, we present an analysis of BENEFITTER highlighting some of the salient aspects of our proposed framework on ICU patient outcome task. In particular, we discuss how our method explains the `benefit` prediction by highlighting the parts of inputs that contributed most for the prediction, and how our `benefit` formulation assists with model evaluation.

TABLE 6.7: BENEFITTER wins most of the times. Accuracy on benchmark datasets against mean tardiness $\in \{0.5, 0.75\}$. *Bold* represents best accuracy within a given tardiness tolerance, and the *underline* represents the next best accuracy. '-' indicates that on average method requires more observations than the given tardiness tolerance. '✗' specifies non-applicability of a method on the dataset, and 'DNS' shows that a method does not scale for the dataset.

Dataset	Tardiness (≤)	ECTS	EDSC	C-ECTS	RelClass	E2EL	EARLIEST	BENEFITTER
ECG200	0.50	-	0.84	0.83	0.88	0.87	0.66	**0.91**
	0.75	-	0.84	0.83	0.89	0.87	0.76	**0.91**
ItalyPower	0.50	-	-	0.78	0.85	0.89	0.71	**0.93**
	0.75	-	0.85	0.94	**0.95**	0.89	0.71	0.93
GunPoint	0.50	-	0.95	0.80	-	0.93	0.78	**0.97**
	0.75	0.91	0.95	0.84	0.91	0.96	0.78	**0.97**
TwoLeadECG	0.50	-	0.88	0.89	-	0.79	0.73	**0.98**
	0.75	0.73	0.89	0.94	0.73	0.86	0.73	**0.98**
Wafer	0.50	-	0.99	0.96	**1.0**	0.99	0.99	0.99
	0.75	**1.0**	0.99	0.96	**1.0**	0.99	0.99	0.99
ECGFiveDays	0.50	-	-	0.59	0.57	0.64	0.57	**0.87**
	0.75	0.72	**0.95**	0.59	0.77	0.77	0.57	0.87
MoteStrain	0.50	-	0.8	**0.85**	-	-	0.67	**0.85**
	0.75	-	0.8	**0.85**	-	-	0.67	**0.85**
Coffee	0.50	-	-	**0.98**	0.89	0.53	0.85	0.93
	0.75	-	0.75	**0.98**	0.89	0.53	0.85	0.93
Yoga	0.50	-	0.71	0.64	-	**0.79**	0.65	0.76
	0.75	-	0.71	0.64	-	**0.79**	0.65	0.76
SonyAIBO	0.50	-	0.80	0.81	0.81	**0.92**	0.81	**0.92**
	0.75	0.69	0.80	0.81	0.81	**0.92**	0.81	**0.92**
Endomondo	0.50	✗	DNS	✗	✗	**0.68**	0.66	0.66

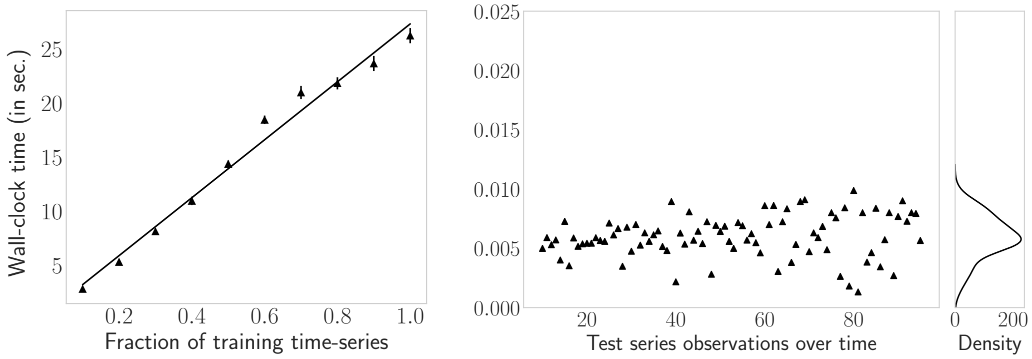


FIGURE 6.4: (left) BENEFITTER scales linearly with the number of time-series, and (right) provides constant-time decision.

Explaining Benefit Estimation Our method utilizes the attention mechanism (see §6.4) in the RNN network for `benefit` regression. The model calculates *weights* corresponding to each hidden state $\mathbf{h}_t$. These weights can indicate which of the time dimensions model focuses on to estimate the benefit for the current input series. In Fig. 6.5, we plot two dimensions of the input time series from EEG dataset. These

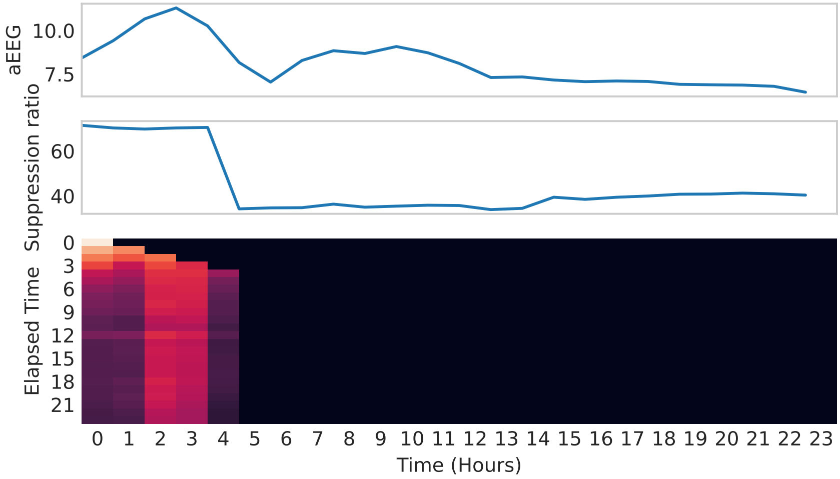


FIGURE 6.5: Input test sequence with corresponding attention weights evaluated at $t = 0, \ldots, 23$. Model decision at $t = 4$.

dimensions correspond to amplitude of the EEG and suppression ratio when measured in left hemisphere of the brain. The input sequence is taken from the hourly sampled dataset. Note that there are sharp rise and fall in the aEEG signal from $t = 1$ to $t = 5$, and from $t = 5$ to $t = 13$. Similarly, we notice sharp changes in SR signal around $t = 4$. The model outputs the attention weights corresponding to each time-dimension of all the inputs (107 dimensional EEG) shown in Fig. 6.5 as a heatmap (dark colors indicate lower weights, lighter colors indicate higher weights). BENEFITTER outputs a decision at $t = 4$, however we evaluate the model at further time steps. Note that the each row of heat map represents evaluation of input at $t = 0, \ldots, 23$. For each evaluation, we obtain a weight signifying the importance of a time dimension which are plotted as heatmap. X-axis of heat map corresponds to time dimension, and y-axis of the heatmap corresponds to the evaluation time step of the input. Observe that the attention places higher weights towards the beginning of the time series where we observe the crests and troughs of in aEEG signal, and abrupt changes in SR signal.

To achieve better outcomes for patients, it is critical to direct attention of the clinician to the time periods with aberrant brain activity. The visualization of importance weights is useful in critical applications where each decision involves a high cost, and particularly helpful in drawing attention of the clinician to important time steps. The advantages to these heatmap visualization are two-fold – (i) it directs the attention of the clinician to time periods in the EEG signals that lead to outcome prediction, and thereby reducing information load on the clinician, and (ii) it allows validation of the predicted outcome where the clinician can cross-check the highlighted time periods; such transparency into model output builds trust, and the feedback may also be used to improve the model. Thus, the estimated benefit along with the visualization of importance weights can assist an expert better for any intervention.

Assisting with Model Evaluation In the clinical setting with comatose patients, there is a natural *cost* associated with an inaccurate prediction and *savings* obtained from knowing the labels early. The `benefit` modeling captures the overall value of outputting a decision. Though, we use this value for learning a regression model, the `benefit` formulation could be used as an evaluation metric to asses the quality of a predictive model. If we know the domain specific unit-time savings s and misclassification cost M, we can then evaluate a model performance for that particular value of s and M. Table 6.8 reports evaluation of BENEFITTER for various values of M/s with model trained using the same values of M/s. We notice that with increasing M/s we improve the precision of the model, however increased M/s also results in higher penalty for any misclassification. For our model trained on hourly sampled EEG data, we observe that values above $M/s = 300$, results in overall negative `benefit` averaged over test data. Assuming unit-time savings $s = \$4000$, we can tolerate lawsuit costs up to \$1.2million for $M/s = 300$. Similarly, any model can be evaluated to assess its usefulness using our `benefit` formulation as an evaluation measure in critical domains.

TABLE 6.8: Taining and evaluation of BENEFITTER for $M/s = \{100, 200, 300, 400\}$ on EEG hourly sampled data.

M/s	Precision	Recall	F1 score	Accuracy	Tardiness	Benefit
100	0.80	0.68	0.73	0.83	0.64	2737
200	0.80	0.67	0.71	0.82	0.68	1032
300	0.82	0.67	0.74	0.84	0.68	156
400	0.82	0.69	0.75	0.84	0.70	-1326

Part III

Conclusions

Chapter 7

Conclusions and Future Directions

7.1 Summary of Contributions

7.1.1 Methodological Contributions

In Chapter 2, we introduced SPI, a new ensemble approach that leverages privileged information (data available only for training examples) for unsupervised anomaly detection. Our work builds on the LUPI paradigm, and to the best of our knowledge, is the first attempt to incorporating PI to improve the state-of-the-art ensemble detectors. We validated the effectiveness of our method on both benchmark datasets as well as three real-world case studies. We showed that SPI, and proposed lightweight SPI-LITE consistently outperform the baselines. Our case studies leveraged a variety of privileged information—"historical future", complex features, expert knowledge—and verified that SPI can unlock multiple benefits for anomaly detection in terms of detection latency, speed, as well as accuracy.

Chapter 3 studies fairness in the context of outlier detection. Although fairness in machine learning has become increasingly prominent in recent years, fairness in the context of unsupervised outlier detection has received comparatively little study. OD is an integral data-driven task in a variety of domains including finance, healthcare and security, where it is used to inform and prioritize auditing measures. Without careful attention, OD as-is can cause unjust flagging of *societal minorities* (w.r.t. race, sex, etc.) because of their standing as *statistical minorities*, when minority status does not indicate positive-class membership (crime, fraud, etc.). This unjust flagging can propagate to downstream supervised classifiers and further exacerbate the issues. Our work tackles the problem of fairness-aware outlier detection. Specifically, we first introduce guiding desiderata for, and concrete formalization of the fair OD problem. We then present FAIROD, a fairness-aware, principled end-to-end detector which addresses the problem, and satisfies several appealing properties: (i) *detection effectiveness:* it is effective, and maintains high detection accuracy, (ii) *treatment parity:* it does not suffer disparate treatment at decision time, (iii) *statistical parity:* it maintains group fairness across minority and majority groups, and (iv) *group fidelity:* it emphasizing flagging of truly high-risk samples within each group, aiming to curb detector "laziness". Finally, we show empirical results across diverse real and synthetic datasets, demonstrating that our approach achieves fairness goals while providing accurate detection, significantly outperforming unsupervised fair representation learning and data de-biasing based baselines.

Chapter 4 presented GEN^{2}OUT – a principled and generalized anomaly detection algorithm that can detect point as well as clustered anomalies. We first design and introduce guiding axioms that a generalized detector should satisfy. Then, we proposed GEN^{2}OUT which has the following desirable properties: (i) Principled and Sound: we propose five axioms that GEN^{2}OUT obeys them, in contrast to top competitors; (ii) Doubly-general: propose doubly general – simultaneously detects point and group anomalies – GEN^{2}OUT. It does not require information on group structure, and ranks detected groups of varying sizes in order of their anomalousness; (iii) Scalability: linear on the input size; requires minutes on 1M dataset on a stock machine; (iv) Effectiveness: applied on real-world data, GEN^{2}OUT wins in most cases over 27 benchmark datasets for point anomaly detection, and agrees with ground truth on seizure detection as well as group detection tasks.

7.1.2 Applications in Decision Support

Chapter 5 introduces a an unsupervised ensemble method to identify health care fraud using massive claims data. Our approach uses different data modalities – including patient medical history, provider coding patterns, and provider spending – to detect anomalous behavior consistent with fraud and abuse. We combine evidence from multiple unsupervised outlier detection algorithms that use different types of global and local analysis to create a final ranking of suspiciousness. Besides detection, the methodology offers interpretability, where qualitative case studies of our results based on model-specific explanations pinpoint specific ICD and DRG codes associated with excess spending at a provider. Finally, our method allows us to characterize the types of providers most likely to be ranked as suspicious, which may be useful for guiding anti-fraud policy more broadly.

In Chapter 6, we consider benefit-aware early prediction of health outcomes for ICU patients and proposed BENEFITTER that is designed to effectively handle multivariate and variable-length signals such as EEG recordings. We made multiple contributions. Novel, cost-aware problem formulation: BENEFITTER infuses the incurred savings from an early prediction as well as the cost from misclassification into a unified target called `benefit`. Unifying these two quantities allows us to directly estimate a *single* target, i.e., `benefit`, and importantly dictates BENEFITTER exactly *when* to output a prediction: whenever estimated `benefit` becomes positive. Efficiency and speed: The training time for BENEFITTER is linear in the number of input sequences, and it can operate under a streaming setting to update its decision based on incoming observations. Multi-variate and multi-length time-series: BENEFITTER is designed to handle multiple time sequences, of varying length, suitable for various domains including health care. Effectiveness on real-world data: We applied BENEFITTER in early prediction of health outcomes on ICU-EEG data where BENEFITTER provides up to 2× time-savings as compared to competitors while achieving equal or better accuracy. BENEFITTER also outperformed or tied with top competitors on other real-world benchmarks.

7.2 Future Directions

The thesis has laid foundations for unsupervised, explainable, and equitable AD, and provided domain specific applications of machine learning as a tool to assist in decision making. This has shepherded way for some important future directions that further empower data -driven decision support.

Human-centered anomaly detection

Currently the thesis mostly focuses on explanations to aid users in decision making, which is crucial for settings in which outliers need to be audited by human analysts. However, the explanations do not take human analysts into account for development of the anomaly detection system. An important aspect in decision support systems is to understand the challenges human experts may encounter in using OD systems in real-life scenarios. A close collaboration with human auditors or investigators who use the OD systems to investigate and validate the flagged anomalies could inform the design for better OD systems. Understanding human perspectives (needs and concerns) can enable building anomaly detection systems that optimize for detection accuracy and ease of use by human experts. Therefore, I am excited to explore human centered OD beyond explanations.

Anomaly detection in presence of an adversary

Anomaly detection finds applications in several crucial domains e.g. health, finance etc. as discussed in the thesis. However, the algorithms introduced herein mostly focus on ranking instances in order of their anomalousness assuming the observed data are *clean*. These detection methods can be vulnerable to deliberate manipulations by adversaries. For example, adversaries might intentionally design fraudulent transactions to resemble legitimate ones, attempting to bypass fraud detection systems. The potential research direction focuses on enhancing anomaly detection methods to be robust against adversarial attacks. The objective is to develop novel techniques that can accurately identify anomalies in data while being resilient to intentional perturbations introduced by malicious adversaries to deceive the system.

Applications

Fraud and waste in public healthcare: Chapter 5 focused on inpatient hospitals, however, there are numerous entities involved in providing care to patients. An interesting and promising direction is to understand the relations between different entities e.g. hospitals, and home health service provider, through shared patients, physicians or medical equipment providers. The data presents itself as a heterogeneous graph, where small and dense communities (group anomalies) could be of particular interest as they may uncover collusion among entities, contributing to fraud and waste in healthcare spending.

Healthcare Informatics: The thesis presented tools to work with EEG data. There is however further need to understand how underlying dynamics of the brain signals evolve before anomalous events such as seizures. There are several promising directions for with respect to understanding neurodynamics of EEG: are there early markers in the signal that indicate onset of seizures? Are there signal idiosyncrasies reflected that lead to awakening of comatose patients? I will consider spatio-temporal analysis of EEG signals that could enable early decision assistance by locating electrodes (corresponding to a region in the brain) of interest for events such as seizures, and other abnormalities in patients.

Part IV

Appendix

Appendix A

Appendix: Fairness-aware Anomaly Detection

A.1 Proofs

A.1.1 Proof of Claim 1

Proof. We want OD to exhibit detection effectiveness i.e. $P(Y=1|O=1) > P(Y=1)$.

$$\begin{aligned}\text{Now,}\quad P(Y=1|O=1) =& P(PV=a|O=1)\cdot\\ & P(Y=1|PV=a,O=1)+\\ & P(PV=b|O=1)\cdot\\ & P(Y=1|PV=b,O=1)\end{aligned}$$

Given SP, we have

$$\begin{aligned}& P(O=1|PV=a) = P(O=1|PV=b)\\ \implies & P(PV=a|O=1) = P(PV=a), \text{ and}\\ & P(PV=b|O=1) = P(PV=b)\end{aligned}$$

Therefore, we have

$$\begin{aligned}\text{Now,}\quad P(Y=1|O=1) =& P(PV=a)\cdot\\ & P(Y=1|PV=a,O=1)+\\ & P(PV=b)\cdot\\ & P(Y=1|PV=b,O=1)\end{aligned} \tag{A.1}$$

Now,

$$\begin{aligned}P(Y=1) =& P(PV=a)\cdot P(Y=1|PV=a)+\\ & P(PV=b)\cdot P(Y=1|PV=b)\end{aligned}$$

Therefore, if we want $P(Y=1|O=1) > P(Y=1)$, then

$$\begin{aligned}
&P(PV=a)\cdot P(Y=1|PV=a,O=1)+\\
&P(PV=b)\cdot P(Y=1|PV=b,O=1)\\
&\quad>\\
&P(PV=a)\cdot P(Y=1|PV=a)+\\
&P(PV=b)\cdot P(Y=1|PV=b)
\end{aligned} \tag{A.2}$$

$$\begin{aligned}
\implies \exists v \in \{a,b\} \quad s.t.\ & P(Y=1|PV=v,O=1)\\
& >\\
& P(Y=1|PV=v)
\end{aligned}$$

□

A.1.2 Proof of Claim 2

Proof. Without loss of generality, assume that $P(Y=1|PV=a,O=1) > P(Y=1|PV=a)$ i.e. (i.e. $P(Y=1|PV=a,O=1) = K\cdot P(Y=1|PV=a); K>1$), and let $\frac{P(Y=1|PV=a)}{P(Y=1|PV=b)} = \frac{P(Y=1|PV=a,O=1)}{P(Y=1|PV=b,O=1)} = \frac{1}{r}$ then
Case 1: When $P(Y=1|PV=b,O=1) < P(Y=1|PV=b)$

$$\begin{aligned}
& P(Y=1|PV=b,O=1) < P(Y=1|PV=b)\\
\implies & P(Y=1|PV=b,O=1) < r\cdot P(Y=1|PV=a)\\
\implies & P(Y=1|PV=b,O=1) < r\cdot P(Y=1|PV=a,O=1),\\
& [\because P(Y=1|PV=a,O=1) > P(Y=1|PV=a)]
\end{aligned}$$

This contradicts our assumption that $P(Y=1|PV=b,O=1) = r\cdot P(Y=1|PV=a,O=1)$, therefore it must be that $P(Y=1|PV=b,O=1) \geq P(Y=1|PV=b)$.

Case 2: When $P(Y=1|PV=b,O=1) = P(Y=1|PV=b)$

$$\begin{aligned}
& P(Y=1|PV=b,O=1) = P(Y=1|PV=b)\\
\implies & P(Y=1|PV=b,O=1) = r\cdot P(Y=1|PV=a)\\
\implies & P(Y=1|PV=b,O=1) < r\cdot P(Y=1|PV=a,O=1),\\
& [\because P(Y=1|PV=a,O=1) > P(Y=1|PV=a)]
\end{aligned}$$

This contradicts our assumption that $P(Y=1|PV=b,O=1) = r\cdot P(Y=1|PV=a,O=1)$, therefore it must be that $P(Y=1|PV=b,O=1) > P(Y=1|PV=b)$.

Case 3: When $P(Y=1|PV=b,O=1) > P(Y=1|PV=b)$ i.e. $(P(Y=1|PV=b,O=1) = L\cdot P(Y=1|PV=b); L>1)$

Now, we know that,

$$
\begin{aligned}
& P(Y=1|PV=a)\cdot P(Y=1|PV=b,O=1) \\
& \quad = P(Y=1|PV=b)\cdot P(Y=1|PV=a,O=1) \\
\implies & P(Y=1|PV=a)\cdot P(Y=1|PV=b,O=1) \\
& \quad = P(Y=1|PV=b)\cdot K\cdot P(Y=1|PV=a) \\
\implies & P(Y=1|PV=b,O=1) = K\cdot P(Y=1|PV=b) \\
\implies & P(Y=1|PV=b,O=1) > P(Y=1|PV=b)
\end{aligned}
$$

And, for ratio to be preserved, it must be that $L = K$.

Hence, enforcing preservation of ratios implies base-rates in flagged observations are larger than their counterparts in the population. □

A.2 Generalizing to Multi-valued and Multiple Protected Attributes

Multi-valued PV. BENEFITTER generalizes beyond binary PV, and easily applies to settings with multi-valued, specifically categorical PV such as race. Recall that $\mathcal{L}_{SP}$ and $\mathcal{L}_{GF}$ are the loss components that depend on PV. For a categorical PV, $\mathcal{L}_{GF}$ in Eq. (3.13) would simply remain the same, where the outer sum goes over all unique values of the PV. For $\mathcal{L}_{SP}$, one could one-hot-encode (OHE) the PV into multiple variables and minimize the correlation of outlier scores with each variable additively. That is, an outer sum would be added to Eq. (3.12) that goes over the new OHE variables encoding the categorical PV.

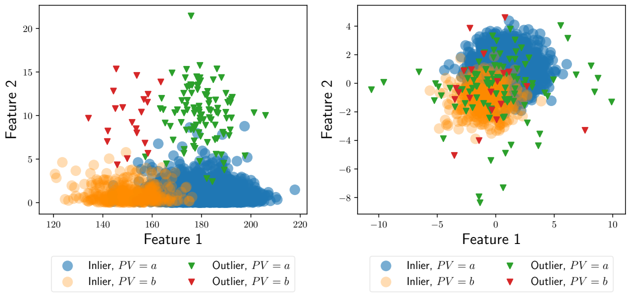
Multiple PVs. BENEFITTER can handle multiple different PVs simultaneously, such as race and gender, since the loss components Eq. (3.12) and Eq. (3.13) can be used additively for each PV. However, the caveat to additive loss is that it would only enforce fairness with respect to each individual PV, and yet may not exhibit fairness for the *joint* distribution of protected variables Kearns, Neel, Roth, and Wu, 2018. Even when additive extension may not be ideal, we avoid modeling multiple protected variables as a single PV that induces groups based on values from the cross-product of available values across all PVs. This is because partitioning of the data based on cross-product may yield many small groups, which could cause instability in learning and poor generalization.

A.3 Data Description

Synthetic data

We illustrate the effectiveness of BENEFITTER on two synthetic datasets, namely `Synth1` and `Synth2` (as illustrated in Fig. A.1). These datasets are constructed to present scenarios that mimic real-world settings, where we may have features which are uncorrelated with respect to outcome labels but partially correlated with PV, or features which are correlated both to outcome labels and PV.

- `Synth1`: In `Synth1`, we simulate a 2-dimensional dataset comprised of samples $X = [x_1, x_2]$ where x_1 is correlated with the protected variable PV, but does not offer any predictive value with respect to ground-truth outlier labels $\mathcal{Y}$, while x_2 is correlated with these labels $\mathcal{Y}$ (see Fig. A.1a). We draw 2400 samples, of which



(A) Synth1 (B) Synth2

FIGURE A.1: Synthetic datasets. See Appendix A.3 for the details of the data generating process.

$PV = a$ (majority) for 2000 points, and $PV = b$ (minority) for 400 points. 120 (5%) of these points are outliers. x_1 differs in terms of shifted means, but equal variances, for both majority and minority groups. x_2 is distributed similarly for both majority and minority groups, drawn from a normal distribution for outliers, and an exponential for inliers. The detailed generative process for the data is below (left), and Fig. A.1a shows a visual.

- Synth2: In Synth2, we again simulate a 2-dimensional dataset comprised of samples $X = [x_1, x_2]$ where x_1, x_2 are partially correlated with both the protected variable PV as well as ground-truth outlier labels $\mathcal{Y}$ (see Fig. A.1b). We draw 2400 samples, of which $PV = a$ (majority) for 2000 points, and $PV = b$ (minority) for 400 points. 120 (5%) of these points are outliers. For inliers, both x_1, x_2 are normally distributed, and differ across majority and minority groups only in terms of shifted means, but equal variances. Outliers are drawn from a product distribution of an exponential and linearly transformed Bernoulli distribution (product taken for symmetry). The detailed generative process for the data is below (right), and Fig. A.1b shows a visual.

Synth1

Simulate samples $X = [x_1, x_2]$ by...

$$PV \sim \text{Bernoulli}(4/5)$$

$$Y \sim \text{Bernoulli}(1/20)$$

$$x_1 \sim \begin{cases} \text{Normal}(-1, 1.44) \quad \text{if} \quad Y = 0,\ PV = 1 \quad \text{[a, majority; inlier]} \\ \text{Normal}(1, 1.44) \quad \text{if} \quad Y = 0,\ PV = 0 \quad \text{[b, minority; inlier]} \\ 2 \times \text{Exponential}(1)(1 - 2 \times \text{Bernoulli}(1/2)) \quad \text{if} \quad Y = 1 \quad \text{[outlier]} \end{cases}$$

$$x_2 \sim \begin{cases} \text{Normal}(-1, 1) \quad \text{if} \quad Y = 0,\ PV = 1 \quad \text{[a, majority; inlier]} \\ \text{Normal}(1, 1) \quad \text{if} \quad Y = 0,\ PV = 0 \quad \text{[b, minority; inlier]} \\ 2 \times \text{Exponential}(1)(1 - 2 \times \text{Bernoulli}(1/2)) \quad \text{if} \quad Y = 1 \quad \text{[outlier]} \end{cases}$$

`Synth2`

$$
\begin{aligned}
&\text{Simulate samples } X = [x_1, x_2] \text{ by...}\\
&PV \sim \text{Bernoulli}(4/5)\\
&\;\;Y \sim \text{Bernoulli}(1/20)\\
&\;\;x_1 \sim \begin{cases} \text{Normal}(180, 10) & \text{if} \quad PV = 1 \quad \text{[a, majority]}\\ \text{Normal}(150, 10) & \text{if} \quad PV = 0 \quad \text{[b, minority]} \end{cases}\\
&\;\;x_2 \sim \begin{cases} \text{Normal}(10, 3) & \text{if} \quad Y = 1 \quad \text{[outlier]}\\ \text{Exponential}(1) & \text{if} \quad Y = 0 \quad \text{[inlier]} \end{cases}
\end{aligned}
$$

Real-world data

We conduct experiments on 4 real-world datasets and select them from diverse domains that have different types of (binary) protected variables, specifically gender, age, and race. Detailed descriptions are as follows.

- **Adult** Lichman et al., 2013 (`Adult`). The dataset is extracted from the 1994 Census database where each data point represents a person. The dataset records income level of an individual along with features encoding personal information on education, profession, investment and family. In our experiments, *gender* $\in$ {*male*, *female*} is used as the protected variable where *female* represents minority group and high earning individuals who exceed an annual income of 50,000 i.e. annual *income* $> 50,000$ are assigned as outliers ($Y = 1$). We further downsample *female* to achieve a *male* to *female* sample size ratio of 4:1 and ensure that percentage of outliers remains the same (at 5%) across groups induced by the protected variable.

- **Credit-defaults** Lichman et al., 2013 (`Credit`). This is a risk management dataset from the financial domain that is based on Taiwan's credit card clients' default cases. The data records information of credit card customers including their payment status, demographic factors, credit data, historical bill and payments. Customer *age* is used as the protected variable where *age* > 25 indicates the majority group and *age* ≤ 25 indicates the minority group. We assign individuals with delinquent *payment status* as outliers ($Y = 1$). The *age* > 25 to *age* ≤ 25 imbalance ratio is 4:1 and contains 5% outliers across groups induced by the protected variable.

- **Abusive Tweets** Blodgett, Green, and O'Connor, 2016 (`Tweets`). The dataset is a collection of Tweets along with annotations indicating whether a tweet is abusive or not. The data are not annotated with any protected variable by default; therefore, to assign protected variable to each Tweet, we employ the following process: We predict the racial dialect — *African-American* or *Mainstream* — of the tweets in the corpus using the language model proposed by Blodgett, Green, and O'Connor, 2016. The dialect is assigned to a Tweet only when the prediction probability is greater than 0.7, and then the predicted *racial dialect* is used as protected variable where *African-American dialect* represents the minority group. In this setting, abusive tweets are labeled as outliers ($Y = 1$) for the task of flagging abusive content on Twitter. The group sample size ratio of *racial dialect* = *African-American* to *racial dialect* = *Mainstream* is set to 4:1. We further sample data points to ensure equal percentage (5%) of outliers across dialect groups.

TABLE A.1: Evaluation measures are reported for the competing methods on the datasets presented in Appendix A.3.

(A) Synth1

	Flag-rate		GroupFidelity		AUC		AP	
Method	$PV = a$	$PV = b$	$PV = a$	$PV = b$	$PV = a$	$PV = b$	$PV = a$	$PV = b$
BASE	0.0262	0.1282	1.0	1.0	0.9594	0.9168	0.8819	0.5849
RW	0.033	0.135	0.9299	0.9309	0.9794	0.9168	0.8819	0.5849
DIR	0.0445	0.0775	0.3953	0.9281	0.9742	0.9138	0.8814	0.7529
LFR	0.0330	0.1350	0.9299	0.9309	0.9794	0.9168	0.8819	0.5849
ARL	0.0520	0.0400	0.9136	0.3955	0.9786	0.5565	0.886	0.1842
BENEFITTER	0.0500	0.0500	0.9639	0.9671	0.9666	0.9634	0.8166	0.7557
FAIROD-L	0.0495	0.0525	0.9149	0.9295	0.9017	0.8714	0.599	0.5214
FAIROD-C	0.0480	0.0600	0.8929	0.9082	0.9499	0.9284	0.7542	0.6501

(B) Synth2

	Flag-rate		GroupFidelity		AUC		AP	
Method	$PV = a$	$PV = b$	$PV = a$	$PV = b$	$PV = a$	$PV = b$	$PV = a$	$PV = b$
BASE	0.0361	0.0811	1.0	1.0	0.6153	0.5464	0.273	0.2335
RW	0.0205	0.1975	0.9242	0.6313	0.7544	0.5586	0.3973	0.2064
DIR	0.0465	0.0675	0.4224	0.9164	0.7892	0.7089	0.3921	0.317
LFR	0.0205	0.1975	0.9242	0.6313	0.7544	0.5586	0.3973	0.2064
ARL	0.0520	0.0400	0.1801	0.1386	0.9786	0.5165	0.886	0.1842
BENEFITTER	0.0500	0.0500	0.9339	0.9201	0.6357	0.6419	0.2726	0.2918
FAIROD-L	0.0500	0.0500	0.8984	0.8843	0.6385	0.6472	0.2742	0.2838
FAIROD-C	0.0450	0.0750	0.8997	0.9095	0.5957	0.5419	0.2665	0.2339

(C) Adult

	Flag-rate		GroupFidelity		AUC		AP	
Method	$PV = a$	$PV = b$	$PV = a$	$PV = b$	$PV = a$	$PV = b$	$PV = a$	$PV = b$
BASE	0.0358	0.0433	1.0	1.0	0.6344	0.6449	0.1105	0.0898
RW	0.0515	0.0391	0.8399	0.8479	0.6323	0.6351	0.1303	0.1141
DIR	0.0515	0.0391	0.9299	0.9309	0.6323	0.6351	0.1303	0.1141
LFR	0.0515	0.0391	0.8099	0.8099	0.6323	0.6351	0.1303	0.1141
ARL	0.0507	0.0444	0.9147	0.5765	0.5951	0.6009	0.0987	0.0848
BENEFITTER	0.0497	0.0511	0.9646	0.9616	0.6374	0.6404	0.1085	0.0912
FAIROD-L	0.0513	0.0403	0.9178	0.9005	0.6425	0.6312	0.1213	0.1048
FAIROD-C	0.0527	0.0302	0.8119	0.7877	0.6533	0.6229	0.1872	0.1435

(D) Credit

	Flag-rate		GroupFidelity		AUC		AP	
Method	$PV = a$	$PV = b$	$PV = a$	$PV = b$	$PV = a$	$PV = b$	$PV = a$	$PV = b$
BASE	0.0445	0.064	1.0	1.0	0.7376	0.7512	0.1938	0.1582
RW	0.0467	0.06627	0.8399	0.8409	0.7376	0.7512	0.1938	0.1582
DIR	0.0467	0.06627	0.6899	0.6809	0.7376	0.7512	0.1938	0.1582
LFR	0.0467	0.06627	0.7299	0.7309	0.7376	0.7512	0.1938	0.1582
ARL	0.0471	0.0645	0.5533	0.6118	0.7242	0.7263	0.1396	0.1054
BENEFITTER	0.0468	0.066	0.9235	0.9421	0.7368	0.7494	0.2134	0.1725
FAIROD-L	0.0475	0.062	0.7147	0.6564	0.7276	0.7394	0.1246	0.1025
FAIROD-C	0.0467	0.0662	0.7871	0.8029	0.7327	0.7484	0.1333	0.1091

(E) Tweets

	Flag-rate		GroupFidelity		AUC		AP	
Method	$PV = a$	$PV = b$	$PV = a$	$PV = b$	$PV = a$	$PV = b$	$PV = a$	$PV = b$
BASE	0.0369	0.1015	1.0	1.0	0.5739	0.5476	0.061	0.0539
RW	0.0479	0.0571	0.2882	0.3312	0.5583	0.582	0.0466	0.0334
DIR	0.0494	0.0507	0.388	0.4178	0.5552	0.5307	0.0454	0.0345
LFR	0.0479	0.0571	0.4082	0.4422	0.5583	0.582	0.0466	0.0334
ARL	0.0482	0.0558	0.5432	0.5762	0.4912	0.5146	0.0504	0.0442
BENEFITTER	0.0488	0.0532	0.9668	0.9671	0.569	0.5699	0.0617	0.0617
FAIROD-L	0.0331	0.1167	0.9137	0.8986	0.5091	0.4237	0.0574	0.0425
FAIROD-C	0.0501	0.0488	0.6753	0.6903	0.5592	0.5891	0.0627	0.1002

(F) Ads

	Flag-rate		GroupFidelity		AUC		AP	
Method	$PV = a$	$PV = b$	$PV = a$	$PV = b$	$PV = a$	$PV = b$	$PV = a$	$PV = b$
BASE	0.0286	0.0318	1.0	1.0	0.7077	0.7234	0.2555	0.2124
RW	0.0491	0.0523	0.8236	0.7813	0.7286	0.7672	0.4227	0.5183
DIR	0.0491	0.0523	0.6236	0.5813	0.7286	0.7672	0.4296	0.5253
LFR	0.0491	0.0523	0.7236	0.6813	0.7286	0.7672	0.4257	0.5253
ARL	0.0499	0.0500	0.5028	0.2181	0.6572	0.6487	0.0885	0.0525
BENEFITTER	0.0499	0.0500	0.9698	0.9699	0.7179	0.7216	0.2592	0.2163
FAIROD-L	0.0683	0.0588	0.5551	0.8684	0.7179	0.7345	0.0005	0.0005
FAIROD-C	0.0499	0.0500	0.6611	0.6966	0.7007	0.7251	0.2636	0.2455

- **Internet ads** Lichman et al., 2013 (Ads). This is a collection of possible advertisements on web-pages. The features characterize each ad by encoding phrases occurring in the ad URL, anchor text, alt text, and encoding geometry of the ad image. We assign observations with class label *ad* as outliers ($Y = 1$) and downsample the data to get an outlier rate of 5%. There exists no demographic information available, therefore we simulate a binary protected variable by randomly assigning each observation to one of two values (i.e. groups) $\in \{0, 1\}$ such that the group sample size ratio is 4:1.

A.4 Hyperparameters

We choose the hyperparameters of BENEFITTER from $\alpha \in \{0.01, 0.5, 0.9\} \times \gamma \in \{0.01, 0.1, 1.0\}$ by evaluating the Pareto curve for fairness and group fidelity criteria. The BASE and BENEFITTER methods both use an auto-encoder with two hidden layers. We fix the number of hidden nodes in each layer to 2 if $d \leq 100$, and 8 otherwise. The representation learning methods LFR and ARL use the model configurations as proposed by their authors. The hyperparameter grid for the preprocessing baselines are set as follows: *repair_level* $\in \{0.0001, 0.001, 0.01, 0.1, 1.0\}$ for DIR, $A_z \in \{0.0001, 0.001, 0.01, 0.1, 0.9\}$ and $A_x = 1 - A_z$ for LFR, and $\lambda \in \{0.0001, 0.001, 0.01, 0.1, 0.9\}$ for ARL. We pick the best model for the preprocessing baselines using Fairness as they only optimize for statistical parity. The best BASE model is selected based on reconstruction error through cross validation upon multiple runs with different random seeds.

A.5 Supplemental Results

In this section, we report Flag-rate, GroupFidelity, AUC and AP (see Table A.1) for the competing methods on a set of datasets (see Appendix A.3) w.r.t groups induced by $PV = v;\ v \in \{a, b\}$ to supplement the experimental results presented in Sec. 6.5. Notice that in most cases (see Table A.1a through Table A.1f), BENEFITTER outperforms the BASE model on label-aware parity metrics (AUC-ratio, AP-ratio) and, furthermore, outperforms BASE on at least one of the performance metrics (e.g. AUC, AP); fairness need not imply worse OD performance.

Appendix B

Appendix: Medicare Data Processing

B.1 Data Preprocessing

Our analysis of provider behavior uses data from each hospitalization and patient in the Medicare system. We consider patients hospitalized in 2017, and we use data from 2012 through 2016 to construct the patients' medical history.

B.1.1 Processing inpatient hospitalizations

We use 100% of samples of Fee-For-Service inpatient claims file from the Medicare data. Annual files contain beneficiary hospitalization details including provider, assigned DRG, assigned ICD codes, and payment reimbursement details including total payment amount, disproportionate payment, education payment, and outlier amount. The raw data is filtered to include claims where the total payment is greater than individual components. For example, if a claim has higher disproportionate payment compared to total payment amount, we exclude such a claim record from our data. These claims may indicate corrupted or noisy data recording. Next, to meet cell-size suppression requirement under our data agreement, we exclude providers along with their claim records, who served 10 or fewer beneficiaries in 2017. We then create lists of unique providers and beneficiaries from the filtered data, which we utilize for merging with other Medicare files.

B.1.2 Provider profile

First, we merge the filtered data with the master beneficiary summary files which contain beneficiary enrollment information including the beneficiary's address, demographics, and chronic conditions. Next, the data are merged with a DRG to MDC mapping.

We then create three types of provider representations. First, we collect the counts for each unique ICD code used by a given provider, creating a representation in terms of ICD codes used. This is a very high dimensional representation, where we apply our subspace based methods.

Next, for each provider, the counts of unique MDC codes are recorded. Since, each MDC typically corresponds to a part of the body, the MDC representation of providers gives a summary distribution in terms of the type of care they provide. Further, we collect counts of chronic conditions for each provider, which represents the distribution of patient population being served by a provider.

We also create the distribution over DRG codes for each provider by collecting the counts of unique DRG codes used by providers. This representation allows us to understand the spending pattern of a provider, since under the PPS system, the DRG code is directly tied to spending amount in each claim.

B.1.3 Beneficiary medical profile

In order to create a beneficiary's medical profile, we stitch through the patient's health care claims across different touchpoints in the Medicare system over the 5 years preceding the 2017 hospitalization (2012 – 2016). Specifically, for these years, we use 100% of samples of Fee-For-Service inpatient and outpatient claims, and 20% of samples of carrier files, which describe physician office visits. 20% is the largest available size of carrier files.

Given the volume of the datasets, we first filter the patient's visits across datasets based on the unique beneficiary list created from inpatient hospitalizations in year 2017. For each type of visit i.e. physician, outpatient, inpatient, we find unique diagnosis codes across five years. Next, for a given beneficiary, we collect the counts over the last five years for each of the unique diagnosis codes. We also include chronic conditions from the year 2016 and the patient's zip code from the master beneficiary summary file. Thus, a beneficiary is represented in terms of assigned codes from past visits, chronic conditions and zip code.

B.2 DOJ Corpus

We scrape and download press releases containing the word 'Medicare' from the central DOJ and the Offices of the United States Attorneys (USAO), which reflect local DOJ branches. The base URLs used in scraping for the DOJ and USAO are https://www.justice.gov/news?keys=medicare and https://www.justice.gov/usao/pressreleases?keys=medicare[1] respectively.

Next, we obtain the list of inpatient hospitals in the Medicare system from 'Medicare Inpatient Hospitals – by Geography and Service' dataset available from the Centers for Medicare and Medicaid Services at https://data.cms.gov/provider-summary-by-type-of-service/medicare-inpatient-hospitals/medicare-inpatient-hospitals-by-geography-and-service. This contains the information on providers including name, CCN (hospital ID), city, and state.

To find providers that are named in DOJ or USAO press releases, we first run a named entity recognizer[2] to obtain the names of all organizations from the press releases. We then run an exact name matching scan for each hospital in the list of Medicare inpatient providers in the recognized organizations from the press releases. Matched hospitals are then recorded as our ground truth. Next, we also run a partial name matching. We obtain tokens for each inpatient hospital in Medicare after dropping the word "hospital" in their name. Then we find organizations from scraped press releases that contain the tokens for Medicare hospitals. Since, we are matching tokens, multiple organizations match for a given Medicare hospital. We manually filter the multiple match and validate the match. The ground truth is augmented with our validated matches, which forms our DOJ corpus for evaluation.

[1] Webpages were accessed on accessed Mar 21, 2022.

[2] We used off-the-shelf entity recognizer Spacy available at https://spacy.io/api/entityrecognizer

Bibliography

AAMC (2022). *Council of Teaching Hospitals and Health Systems (COTH)*. URL: https://www.aamc.org/career-development/affinity-groups/coth.

Adel, Tameem, Isabel Valera, Zoubin Ghahramani, and Adrian Weller (2019). "One-network adversarial fairness". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33, pp. 2412–2420.

Aggarwal, Charu C. (2013). *Outlier Analysis*. Springer.

Aggarwal, Charu C (2015). "Outlier analysis". In: *Data mining*. Springer, pp. 237–263.

Aggarwal, Charu C and Saket Sathe (2017). "Outlier ensembles: An introduction". In: *Springer*.

An, Jinwon and Sungzoon Cho (2015). "Variational autoencoder based anomaly detection using reconstruction probability". In: *Special Lecture on IE* 2.1, pp. 1–18.

Bae, Jong-Myon (2015). "Value-based medicine: concepts and application". In: *Epidemiology and health* 37.

Bagnall, Anthony, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh (2017). "The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances". In: *Data mining and knowledge discovery* 31.3, pp. 606–660.

Bar-Yossef, Ziv, T. S. Jayram, Ravi Kumar, and D. Sivakumar (June 2004). "An information statistics approach to data stream and communication complexity". In: *J. Computer and System Sciences* 68.4. (Preliminary Version in *43rd FOCS*, 2002), pp. 702–732. DOI: 10.1016/j.jcss.2003.11.006.

Barnsley, Michael F. and Alan D. Sloan (Jan. 1988). "A Better Way to Compress Images". In: *BYTE* 13.1, 215–223. ISSN: 0360-5280.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan (2017). "Fairness in machine learning". In: *NIPS Tutorial* 1.

— (2019). *Fairness and Machine Learning*. http://www.fairmlbook.org. fairmlbook.org.

Batista, Gustavo E., Eamonn J. Keogh, Agenor Mafra-Neto, and Edgar Rowton (2011). "SIGKDD Demo: Sensors and Software to Allow Computational Entomology, an Emerging Application of Data Mining". In: *SIGKDD*. New York, NY, USA: ACM.

Bauder, Richard and Taghi Khoshgoftaar (2018a). "Medicare fraud detection using random forest with class imbalanced big data". In: *2018 IEEE international conference on information reuse and integration (IRI)*. IEEE, pp. 80–87.

Bauder, Richard, Taghi M Khoshgoftaar, and Naeem Seliya (2017). "A survey on the state of healthcare upcoding fraud analysis and detection". In: *Health Services and Outcomes Research Methodology* 17.1, pp. 31–55.

Bauder, Richard A and Taghi M Khoshgoftaar (2017). "Medicare fraud detection using machine learning methods". In: *2017 16th IEEE international conference on machine learning and applications (ICMLA)*. IEEE, pp. 858–865.

— (2018b). "The detection of medicare fraud using machine learning methods with excluded provider labels". In: *The Thirty-First International Flairs Conference*.

Becker, David, Daniel Kessler, and Mark McClellan (2005). "Detecting medicare abuse". In: *Journal of Health Economics* 24.1, pp. 189–210.

Bekker, Jessa and Jesse Davis (2020). "Learning from positive and unlabeled data: A survey". In: *Machine Learning* 109.4, pp. 719–760.

Beutel, Alex, Jilin Chen, Zhe Zhao, and Ed H Chi (2017). "Data decisions and theoretical implications when adversarially learning fair representations". In: *arXiv preprint arXiv:1707.00075*.

Beutel, Alex et al. (2019). "Putting fairness principles into practice: Challenges, metrics, and improvements". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 453–459.

Blázquez-García, Ane, Angel Conde, Usue Mori, and Jose A Lozano (2021). "A Review on outlier/Anomaly Detection in Time Series Data". In: *ACM CSUR* 54, pp. 1–33.

Blodgett, Su Lin, Lisa Green, and Brendan O'Connor (2016). "Demographic dialectal variation in social media: A case study of African-American English". In: *arXiv preprint arXiv:1608.08868*.

Bosc, Marcel, Fabrice Heitz, Jean-Paul Armspach, Izzie Namer, Daniel Gounot, and Lucien Rumbach (2003). "Automatic change detection in multimodal serial MRI: application to multiple sclerosis lesion evolution". In: *NeuroImage* 20.2, pp. 643–656.

Boukerche, Azzedine, Lining Zheng, and Omar Alfandi (2020). "Outlier detection: Methods, models, and classification". In: *ACM Computing Surveys (CSUR)* 53.3, pp. 1–37.

Bregón, Aníbal, M Aránzazu Simón, Juan José Rodríguez, Carlos Alonso, Belarmino Pulido, and Isaac Moro (2005). "Early fault classification in dynamic systems using case-based reasoning". In: *CAEPIA*. Springer, pp. 211–220.

Breunig, Markus M, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander (2000). "LOF: identifying density-based local outliers". In: *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 93–104.

Brot-Goldberg, Zarek, Samantha Burn, Timothy Layton, and Boris Vabson (2022). *Rationing medicine through bureaucracy: authorization restrictions in medicare*. Tech. rep. Working Paper.

Brunt, Christopher S (2011). "CPT fee differentials and visit upcoding under Medicare Part B". In: *Health economics* 20.7, pp. 831–841.

Burnaev, Evgeny and Dmitry Smolyakov (2016). "One-Class SVM with Privileged Information and Its Application to Malware Detection." In: *ICDM Workshops*, pp. 273–280. URL: http://dblp.uni-trier.de/db/conf/icdm/icdm2016w.html#BurnaevS16.

Celik, Z Berkay, Patrick McDaniel, Rauf Izmailov, Nicolas Papernot, and Ananthram Swami (2016). "Extending detection with forensic information". In: *arXiv:1603.09638*.

Centers for Medicare & Medicaid Services (2022). *NHE Fact Sheet*. URL: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NHE-Fact-Sheet.

Cevora, George (2020). "Fair Adversarial Networks". In: *arXiv preprint arXiv:2002.12144*.

Chakravarthy, Niranjan, Shivkumar Sabesan, Kostas Tsakalis, and Leon Iasemidis (2009). "Controlling epileptic seizures in a neural mass model". In: *Journal of Combinatorial Optimization* 17.1, pp. 98–116.

Chalapathy, Raghavendra, Aditya Krishna Menon, and Sanjay Chawla (2018). "Anomaly Detection using One-Class Neural Networks". In: *arXiv preprint arXiv:1802.06360*.

Chalapathy, Raghavendra, Edward Toth, and Sanjay Chawla (2018). "Group anomaly detection using deep generative models". In: *ECML-PKDD*, pp. 173–189.

Chandola, Varun, Arindam Banerjee, and Vipin Kumar (2009). "Anomaly detection: A survey". In: *ACM Comp. Surveys* 41.3, p. 15. URL: http://scholar.google.de/scholar.bib?q=info:jAfBmk-9uAcJ:scholar.google.com/&output=citation&hl=de&ct=citation&cd=0.

Chandola, Varun, Sreenivas R Sukumar, and Jack C Schryver (2013). "Knowledge discovery from massive healthcare claims data". In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1312–1320.

Chen, Jinghui, Saket Sathe, Charu Aggarwal, and Deepak Turaga (2017). "Outlier detection with autoencoder ensembles". In: *Proceedings of the 2017 SIAM international conference on data mining*. SIAM, pp. 90–98.

Chen, Jixu, Xiaoming Liu, and Siwei Lyu (2012). "Boosting with Side Information." In: *ACCV*. URL: http://dblp.uni-trier.de/db/conf/accv/accv2012-1.html#ChenLL12.

Chen, Yanping et al. (2015). *The UCR Time Series Classification Archive*. www.cs.ucr.edu/~eamonn/time_series_data/.

Corbett-Davies, Sam and Sharad Goel (2018b). "The measure and mismeasure of fairness: A critical review of fair machine learning". In: *arXiv preprint arXiv:1808.00023*.

— (2018a). "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning." In: *CoRR* abs/1808.00023. URL: http://dblp.uni-trier.de/db/journals/corr/corr1808.html#abs-1808-00023.

Dachraoui, Asma, Alexis Bondu, and Antoine Cornuéjols (2015). "Early classification of time series as a non myopic sequential decision making problem". In: *ECML/PKDD*. Springer, pp. 433–447.

Dafny, Leemore S (2005). "How do hospitals respond to price changes?" In: *American Economic Review* 95.5, pp. 1525–1547. URL: https://www.aeaweb.org/articles?id=10.1257/000282805775014236.

Davidson, Ian and S. S. Ravi (2020). "A Framework for Determining the Fairness of Outlier Detection." In: *ECAI*. Vol. 325, pp. 2465–2472. URL: http://dblp.uni-trier.de/db/conf/ecai/ecai2020.html#DavidsonR20.

Demšar, Janez (2006). "Statistical comparisons of classifiers over multiple data sets". In: *Journal of Machine Learning Research* 7.Jan, pp. 1–30.

Edwards, Harrison and Amos Storkey (2015). "Censoring representations with an adversary". In: *arXiv preprint arXiv:1511.05897*.

Ekin, Tahir, Greg Lakomski, and Rasim Muzaffer Musal (2019). "An unsupervised Bayesian hierarchical method for medical fraud assessment". In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 12.2, pp. 116–124.

Eliason, Paul J, Riley J League, Jetson Leder-Luis, Ryan C McDevitt, and James W Roberts (2021). *Ambulance Taxis: The Impact of Regulation and Litigation on Health Care Fraud*. Tech. rep. National Bureau of Economic Research.

Ellis, Randall P and Thomas G McGuire (1986). "Provider behavior under prospective reimbursement: Cost sharing and supply". In: *Journal of health economics* 5.2, pp. 129–151.

Emmott, Andrew, Shubhomoy Das, Thomas Dietterich, Alan Fern, and Weng-Keen Wong (2013a). “Systematic Construction of Anomaly Detection Benchmarks from Real Data”. In: *KDD ODD*.

Emmott, Andrew F, Shubhomoy Das, Thomas Dietterich, Alan Fern, and Weng-Keen Wong (2013b). “Systematic construction of anomaly detection benchmarks from real data”. In: *KDD - ODD workshop*, pp. 16–21.

Feldman, Michael, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian (2015). “Certifying and removing disparate impact”. In: *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268.

Feyereisl, Jan and Uwe Aickelin (2012). “Privileged information for data clustering.” In: *Inf. Sci.* 194, pp. 4–23. URL: http://dblp.uni-trier.de/db/journals/isci/isci194.html#FeyereislA12.

Fouad, Shereen, Peter Tino, Somak Raychaudhury, and Petra Schneider (2013). “Incorporating Privileged Information Through Metric Learning.” In: *IEEE Neural Net. Learning Sys.* 24.7.

Franceschini, Fiorenzo, Domenico A Maisano, and Luca Mastrogiacomo (2022). “Ranking Aggregation Techniques”. In: *Rankings and Decisions in Engineering*. Springer, pp. 85–160.

Gao, Yifeng, Qingzhe Li, Xiaosheng Li, Jessica Lin, and Huzefa Rangwala (2017). “TrajViz: A Tool for Visualizing Patterns and Anomalies in Trajectory”. In: *ECML/PKDD (3)*. Vol. 10536. Springer, pp. 428–431.

Ghalwash, Mohamed F and Zoran Obradovic (2012). “Early classification of multivariate temporal observations by extraction of interpretable shapelets”. In: *BMC bioinformatics* 13.1, p. 195.

Goel, Naman, Mohammad Yaghini, and Boi Faltings (2018). “Non-discriminatory machine learning through convex fairness criteria”. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 116–116.

Gogoi, Prasanta, DK Bhattacharyya, Bhogeswar Borah, and Jugal K Kalita (2011). “A survey of outlier detection methods in network anomaly identification”. In: *The Computer Journal* 54.4, pp. 570–588.

Gray, Muir (2017). *Value based healthcare*.

Guha, Sudipto, Nina Mishra, Gourav Roy, and Okke Schrijvers (2016). “Robust random cut forest based anomaly detection on streams”. In: *ICML*, pp. 2712–2721.

Gupta, Atul, Sabrina T Howell, Constantine Yannelis, and Abhinav Gupta (2021). *Does Private Equity Investment in Healthcare Benefit Patients? Evidence from Nursing Homes*. Working Paper 28474. National Bureau of Economic Research. DOI: 10.3386/w28474. URL: http://www.nber.org/papers/w28474.

Gupta, Manish, Jing Gao, Charu C Aggarwal, and Jiawei Han (2013). “Outlier detection for temporal data: A survey”. In: *IEEE TKDE* 26.9, pp. 2250–2267.

Hardt, Moritz, Eric Price, and Nati Srebro (2016). “Equality of opportunity in supervised learning”. In: *Advances in neural information processing systems*, pp. 3315–3323.

Hartvigsen, Thomas, Cansu Sen, Xiangnan Kong, and Elke Rundensteiner (2019). “Adaptive-halting policy network for early classification”. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 101–110.

Hatami, Nima and Camelia Chira (2013). “Classifiers with a reject option for early time-series classification”. In: *CIEL*. IEEE, pp. 9–16.

He, Haibo and Edwardo A Garcia (2009). “Learning from imbalanced data”. In: *IEEE Transactions on knowledge and data engineering* 21.9, pp. 1263–1284.

He, Zengyou, Xiaofei Xu, and Shengchun Deng (2003). “Discovering cluster-based local outliers”. In: *Pattern Recognition Letters* 24.9-10, pp. 1641–1650.

Healthcare Fraud Prevention Partnership (2022). *Healthcare Fraud Prevention Partnership*. URL: https://www.cms.gov/hfpp.

Herland, Matthew, Taghi M Khoshgoftaar, and Richard A Bauder (2018). “Big data fraud detection using multiple medicare data sources”. In: *Journal of Big Data* 5.1, pp. 1–21.

Hillary, Wilson, Gole Justin, Mishra Bharat, and Mishra Jitendra (2016). “Value based healthcare”. In: *Advances in management* 9.1, p. 1.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). “Long short-term memory”. In: *Neural computation* 9.8, pp. 1735–1780.

Howard, David (2020). “False claims act liability for overtreatment”. In: *Journal of Health Politics, Policy and Law* 45.3, pp. 419–437.

Hutson, Timothy, Diana Pizarro, Sandipan Pati, and Leon D Iasemidis (2018). “Predictability and Resetting in a Case of Convulsive Status Epilepticus”. In: *Frontiers in neurology* 9, p. 172.

Ismail Fawaz, Hassan, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller (2019). “Deep learning for time series classification: a review”. In: *Data mining and knowledge discovery* 33.4, pp. 917–963.

Järvelin, K. and J. Kekäläinen (2002). “Cumulated gain-based evaluation of IR techniques”. In: *ACM Transactions on Information Systems (TOIS)* 20.4, pp. 422–446. URL: http://scholar.google.de/scholar.bib?q=info:6Bdw8cs-UYMJ:scholar.google.com/&output=citation&hl=de&as_sdt=0,5&ct=citation&cd=0.

Johnson, Justin M and Taghi M Khoshgoftaar (2019). “Medicare fraud detection using neural networks”. In: *Journal of Big Data* 6.1, p. 63.

Jonschkowski, Rico, Sebastian Höfer, and Oliver Brock (2015). “Patterns for learning with side information”. In: *arXiv:1511.06429*.

Joudaki, Hossein et al. (2015). “Using data mining to detect health care fraud and abuse: a review of literature”. In: *Global journal of health science* 7.1, p. 194.

Justusson, BI (1981). “Median filtering: Statistical properties”. In: *Two-Dimensional Digital Signal Prcessing II*. Springer, pp. 161–196.

Kamiran, Faisal and Toon Calders (2012). “Data preprocessing techniques for classification without discrimination”. In: *Knowledge and Information Systems* 33.1, pp. 1–33.

Kearns, Michael, Seth Neel, Aaron Roth, and Zhiwei Steven Wu (2018). “Preventing fairness gerrymandering: Auditing and learning for subgroup fairness”. In: *International Conference on Machine Learning*. PMLR, pp. 2564–2572.

Keogh, Eamonn J., Jessica Lin, Ada Wai-Chee Fu, and Helga Van Herle (2006). “Finding Unusual Medical Time-Series Subsequences: Algorithms and Applications”. In: *IEEE Trans. Information Technology in Biomedicine* 10.3, pp. 429–439.

Kingma, Diederik P and Jimmy Ba (2014). “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980*.

Knight, Carl (2009). *Luck Egalitarianism: Equality, Responsibility, and Justice*. Edinburgh University Press. ISBN: 9780748638697. URL: http://www.jstor.org/stable/10.3366/j.ctt1r2483.

Kodali, Naveen, Jacob Abernethy, James Hays, and Zsolt Kira (2017). “On convergence and stability of gans”. In: *arXiv preprint arXiv:1705.07215*.

Krasanakis, Emmanouil, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris (2018). “Adaptive sensitive reweighting to mitigate bias in fairness-aware classification". In: *Proceedings of the 2018 World Wide Web Conference*, pp. 853–862.

Kriegel, Hans-Peter, Peer Kröger, Erich Schubert, and Arthur Zimek (2009). "Outlier detection in axis-parallel subspaces of high dimensional data". In: *Pacific-asia conference on knowledge discovery and data mining*. Springer, pp. 831–838.

Kriegel, Hans-Peter, Matthias Schubert, and Arthur Zimek (2008). "Angle-based outlier detection in high-dimensional data". In: *KDD*, pp. 444–452.

Krishnan, B et al. (2015). "Epileptic focus localization based on resting state interictal MEG recordings is feasible irrespective of the presence or absence of spikes". In: *Clinical Neurophysiology* 126.4, pp. 667–674.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25, pp. 1097–1105.

Kumaraswamy, Nishamathi, Mia K Markey, Tahir Ekin, Jamie C Barner, and Karen Rascati (2022). "Healthcare Fraud Data Mining Methods: A Look Back and Look Ahead". In: *Perspectives in Health Information Management* 19.1.

Kuncheva, Ludmila I and Christopher J Whitaker (2003). "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy". In: *Machine learning* 51.2, pp. 181–207.

Lapin, Maksim, Matthias Hein, and Bernt Schiele (2014). "Learning using privileged information: SVM+ and weighted SVM." In: *Neural Networks* 53, pp. 95–108.

LaRoche, Suzette M and Hiba Arif Haider (2018). *Handbook of ICU EEG monitoring*. Springer Publishing Company.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning". In: *nature* 521.7553, pp. 436–444.

Leder-Luis, Jetson (2020). *Can whistleblowers root out public expenditure fraud? evidence from medicare*.

Lee, Meng-Chieh, Shubhranshu Shekhar, Christos Faloutsos, T Noah Hutson, and Leon Iasemidis (2021). "Gen2Out: Detecting and ranking generalized anomalies". In: *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, pp. 801–811.

Lee, Meng-Chieh et al. (2020). "AutoAudit: Mining Accounting and Time-Evolving Graphs". In: *arXiv preprint arXiv:2011.00447*.

Lee, Thomas and Michael Porter (2013). *The strategy that will fix healthcare*. Harvard Business Review Boston.

Li, Daoyuan, Tegawende F Bissyande, Jacques Klein, and Yves Le Traon (2016). "Time series classification with discrete wavelet transformed data". In: *International Journal of Software Engineering and Knowledge Engineering* 26.09n10, pp. 1361–1377.

Lichman, Moshe et al. (2013). *UCI machine learning repository*.

Lipton, Zachary, Julian McAuley, and Alexandra Chouldechova (2018). "Does mitigating ML's impact disparity require treatment disparity?" In: *Advances in Neural Information Processing Systems*, pp. 8125–8135.

Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou (2008a). "Isolation Forest." In: *ICDM*.

— (2008b). "Isolation forest". In: *ICDM*. IEEE, pp. 413–422.

— (2012). "Isolation-based anomaly detection". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 6.1, pp. 1–39.

Lotov, Alexander V, Vladimir A Bushenkov, and Georgy K Kamenev (2013). *Interactive decision maps: Approximation and visualization of Pareto frontier*. Vol. 89. Springer Science & Business Media.

Louizos, Christos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel (2015). "The variational fair autoencoder". In: *arXiv preprint arXiv:1511.00830*.

Lundberg, Scott M and Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30.

Lundberg, Scott M et al. (2020). "From local explanations to global understanding with explainable AI for trees". In: *Nature machine intelligence* 2.1, pp. 56–67.

Luo, Wei and Marcus Gallagher (2010). "Unsupervised DRG upcoding detection in healthcare databases". In: *2010 IEEE International Conference on Data Mining Workshops*. IEEE, pp. 600–605.

Luong, Minh-Thang, Hieu Pham, and Christopher D Manning (2015). "Effective approaches to attention-based neural machine translation". In: *arXiv preprint arXiv:1508.04025*.

Ma, Yunlong, Peng Zhang, Yanan Cao, and Li Guo (2013). "Parallel auto-encoder for efficient outlier detection". In: *2013 IEEE International Conference on Big Data*. IEEE, pp. 15–17.

Maaten, Laurens Van der and Geoffrey Hinton (2008). "Visualizing data using t-SNE." In: *JMLR* 9.

Madras, David, Elliot Creager, Toniann Pitassi, and Richard Zemel (2018). "Learning adversarially fair and transferable representations". In: *arXiv preprint arXiv:1802.06309*.

Marcacini, Ricardo Marcondes, Marcos Aurélio Domingues, Eduardo R. Hruschka, and Solange Oliveira Rezende (2014). "Privileged Information for Hierarchical Document Clustering: A Metric Learning Approach." In: *ICPR*, pp. 3636–3641. URL: http://dblp.uni-trier.de/db/conf/icpr/icpr2014.html#MarcaciniDHR14.

Medpac (2021). *Hospital Acute Inpatient Services Payment System*. URL: https://www.medpac.gov/wp-content/uploads/2021/11/medpac_payment_basics_21_hospital_final_sec.pdf.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan (2019). "A survey on bias and fairness in machine learning". In: *arXiv preprint arXiv:1908.09635*.

Mitchell, Tom M and Tom M Mitchell (1997). *Machine learning*. Vol. 1. 9. McGraw-hill New York.

Mori, Usue, Alexander Mendiburu, Sanjoy Dasgupta, and Jose A Lozano (2017). "Early classification of time series by simultaneously optimizing the accuracy and earliness". In: *IEEE Trans. on Neur. Net. Learn. Sys.*, pp. 4569–4578.

Morley, Andrew, Lizzie Hill, and AG Kaditis (2016). "10-20 system EEG Placement". In: *Eur. Respir. Soc*, p. 34.

Muandet, Krikamol and Bernhard Schölkopf (2013). "One-class support measure machines for group anomaly detection". In: *arXiv preprint arXiv:1303.0309*.

Nam, Giung, Jongmin Yoon, Yoonho Lee, and Juho Lee (2021). "Diversity matters when learning from ensembles". In: *Advances in Neural Information Processing Systems* 34, pp. 8367–8377.

NHE Fact Sheet (2021). *Centers for Medicare & Medicaid Services NHE Fact Sheet*. URL: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NHE-Fact-Sheet.

Ni, Jianmo, Larry Muhlstein, and Julian McAuley (2019). "Modeling Heart Rate and Activity Data for Personalized Fitness Recommendation". In: *The Web Conference*. ACM, pp. 1343–1353.

Niu, Li, Wen Li, and Dong Xu (2016). “Exploiting Privileged Information from Web Data for Action and Event Recognition.” In: *Intern. J. of Comp. Vision* 118.2, pp. 130–150. URL: http://dblp.uni-trier.de/db/journals/ijcv/ijcv118.html#NiuLX16.

Noridian Healthcare Solutions (2022). *Unified Program Integrity Contractor (UPIC)*. URL: https://med.noridianmedicare.com/web/jddme/cert-reviews/upic.

Olfat, Matt and Anil Aswani (2019). “Convex formulations for fair principal component analysis”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33, pp. 663–670.

P, Deepak and Savitha Sam Abraham (2020). *Fair Outlier Detection*. arXiv: 2005.09900 [cs.LG].

Pang, Guansong, Chunhua Shen, Longbing Cao, and Anton van den Hengel (2020). “Deep learning for anomaly detection: A review”. In: *arXiv preprint arXiv:2007.02500*.

Parrish, Nathan, Hyrum S Anderson, Maya R Gupta, and Dun Yu Hsiao (2013). “Classifying with confidence from incomplete information”. In: *The Journal of Machine Learning Research* 14.1, pp. 3561–3589.

Pevnỳ, Tomáš (2016). “Loda: Lightweight on-line detector of anomalies”. In: *Machine Learning* 102.2, pp. 275–304.

Phua, Clifton, Vincent Lee, Kate Smith, and Ross Gayler (2010). “A comprehensive survey of data mining-based fraud detection research”. In: *arXiv preprint arXiv:1009.6119*.

Qin, Tao, Tie-Yan Liu, and Hang Li (2010). “A general approximation framework for direct optimization of information retrieval measures”. In: *Information retrieval* 13.4, pp. 375–397.

Ralanamahatana, Chotirat Ann, Jessica Lin, Dimitrios Gunopulos, Eamonn Keogh, Michail Vlachos, and Gautam Das (2005). “Mining time series data”. In: *Data mining and knowledge discovery handbook*. Springer, pp. 1069–1103.

Rayana, Shebuti (2016). *ODDS Library*. URL: http://odds.cs.stonybrook.edu.

ResDac (2022). *Major Diagnostic Category (MDC) Code*. URL: https://resdac.org/cms-data/variables/major-diagnostic-category-mdc-code.

Ribeiro, Bernardete, Catarina Silva, Ning Chen, Armando Vieira, and João Carvalho das Neves (2012). “Enhanced default risk models with SVM+.” In: *Expert Syst. Appl.* 39.11, pp. 10140–10152. URL: http://dblp.uni-trier.de/db/journals/eswa/eswa39.html#RibeiroSCVN12.

Rodríguez, Juan J, Carlos J Alonso, and Henrik Boström (2001). “Boosting interval based literals”. In: *Intelligent Data Analysis* 5.3, pp. 245–262.

Rosenberg, Marjorie A, Dennis G Fryback, and David A Katz (2000). “A statistical model to detect DRG upcoding”. In: *Health Services and Outcomes Research Methodology* 1.3, pp. 233–252.

Ruff, Lukas et al. (2018). “Deep One-Class Classification”. In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80, pp. 4393–4402.

Rußwurm, Marc, Sébastien Lefèvre, Nicolas Courty, Rémi Emonet, Marco Körner, and Romain Tavenard (2019). “End-to-end Learning for Early Classification of Time Series”. In: *arXiv preprint arXiv:1901.10681*.

Sathe, Saket and Charu C Aggarwal (2016). “Subspace outlier detection in linear time with randomized hashing”. In: *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, pp. 459–468.

Schäfer, Patrick and Ulf Leser (2020). “TEASER: early and accurate time series classification”. In: *Data mining and knowledge discovery*, pp. 1336–1362.

Schölkopf, Bernhard, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson (2001). "Estimating the support of a high-dimensional distribution". In: *Neural computation* 13.7, pp. 1443–1471.

Schubert, Erich, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu (2017). "DBSCAN revisited, revisited: why and how you should (still) use DBSCAN". In: *ACM TODS* 42, pp. 1–21.

Shah, Neil, Alex Beutel, Brian Gallagher, and Christos Faloutsos (2014). "Spotting suspicious link behavior with fbox: An adversarial perspective". In: *2014 IEEE International Conference on Data Mining*. IEEE, pp. 959–964.

Sharmanska, Viktoriia, Novi Quadrianto, and Christoph H. Lampert (2013). "Learning to Rank Using Privileged Information." In: *ICCV*, pp. 825–832. URL: http://dblp.uni-trier.de/db/conf/iccv/iccv2013.html#SharmanskaQL13.

— (2014). "Learning to Transfer Privileged Information". In: arXiv:1410.0389. URL: http://arxiv.org/abs/1410.0389.

Shekhar, Shubhranshu and Leman Akoglu (2018). "Incorporating privileged information to unsupervised anomaly detection". In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018*. Springer, pp. 87–104.

Shekhar, Shubhranshu, Dhivya Eswaran, Bryan Hooi, Jonathan Elmer, Christos Faloutsos, and Leman Akoglu (2023). "Benefit-aware early prediction of health outcomes on multivariate eeg time series". In: *Journal of Biomedical Informatics* 139, p. 104296.

Shekhar, Shubhranshu, Jetson Leder-Luis, and Leman Akoglu (2023). *Unsupervised Machine Learning for Explainable Health Care Fraud Detection*. Tech. rep. National Bureau of Economic Research.

Shekhar, Shubhranshu, Neil Shah, and Leman Akoglu (2021). "Fairod: Fairness-aware outlier detection". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 210–220.

Shi, Maggie (2022). *Monitoring for Waste: Evidence from Medicare Audits*. URL: https://mshi311.github.io/website2/Shi_MedicareAudits_2022_09_15.pdf.

Shorvon, S. (2009). *Epilepsy*. Oxford Neurology Library. OUP Oxford. ISBN: 9780199560042. URL: https://books.google.com/books?id=7r2XZWSCJoIC.

Shyu, M-L (2003). "A novel anomaly detection scheme based on principal component classifier". In: *Proc. ICDM Foundation and New Direction of Data Mining workshop, 2003*, pp. 172–179.

Silverman, Elaine and Jonathan Skinner (2004). "Medicare upcoding and hospital ownership". In: *Journal of health economics* 23.2, pp. 369–389.

Statista (2022). *Total Medicare spending from 1970 to 2021 (in billion U.S. dollars)*. URL: https://www.statista.com/statistics/248073/distribution-of-medicare-spending-by-service-type/.

Suresh, Nallan C, Jean De Traversay, Hyma Gollamudi, Anu K Pathria, and Michael K Tyler (2014). *Detection of upcoding and code gaming fraud and abuse in prospective payment healthcare systems*. US Patent 8,666,757.

Susto, Gian Antonio, Angelo Cenedese, and Matteo Terzi (2018). "Time-series classification methods: Review and applications to power systems data". In: *Big data application in power systems*, pp. 179–220.

Tavenard, Romain and Simon Malinowski (2016). "Cost-aware early classification of time series". In: *ECML/PKDD*. Springer, pp. 632–647.

Toth, Edward and Sanjay Chawla (2018). "Group deviation detection methods: a survey". In: *ACM CSUR* 51, pp. 1–38.

Traoré, Mamadou Kaba, Gregory Zacharewicz, Raphaël Duboz, and Bernard Zeigler (2019). "Modeling and simulation framework for value-based healthcare systems". In: *Simulation* 95.6, pp. 481–497.

Tsakalis, Kostas and Leon Iasemidis (2006). "Control aspects of a theoretical model for epileptic seizures". In: *International Journal of Bifurcation and Chaos* 16.07, pp. 2013–2027.

U.S. Department of Health and Human Services (2022). *Annual Report of the Departments of Health and Human Services and Justice*. URL: https://oig.hhs.gov/publications/docs/hcfac/FY2021-hcfac.pdf.

U.S. Government Accountability Office (2020). *Payment Integrity Federal Agencies' Estimates of FY 2019 Improper Payments*. URL: https://www.gao.gov/assets/gao-20-344.pdf.

Van Vlasselaer, Véronique et al. (2015). "APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions". In: *Decision Support Systems* 75, pp. 38–48.

Vapnik, Vladimir and Rauf Izmailov (2015). "Learning with Intelligent Teacher: Similarity Control and Knowledge Transfer". In: *Stat. Learning and Data Sci.* Pp. 3–32.

— (2017). "Knowledge transfer in SVM and neural networks." In: *Ann. Math. Artif. Intell.* 81.1-2, pp. 3–19.

Vapnik, Vladimir and Akshay Vashist (Sept. 4, 2009). "A new learning paradigm: Learning using privileged information." In: *Neural Networks* 22.5-6, pp. 544–557. URL: http://dblp.uni-trier.de/db/journals/nn/nn22.html#VapnikV09.

Vaswani, Ashish et al. (2017). "Attention is all you need". In: *NeurIPS*, pp. 5998–6008.

Verma, Sahil and Julia Rubin (2018). "Fairness definitions explained". In: *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, pp. 1–7.

Vlachos, Ioannis, Balu Krishnan, David M Treiman, Konstantinos Tsakalis, Dimitris Kugiumtzis, and Leon D Iasemidis (2016). "The concept of effective inflow: application to interictal localization of the epileptogenic focus from iEEG". In: *IEEE Transactions on Biomedical Engineering* 64.9, pp. 2241–2252.

Wang, Ziheng and Qiang Ji (2015). "Classifier learning with hidden information." In: *CVPR*.

Weiss, Gary M. and Haym Hirsh (1998). "Learning to Predict Rare Events in Event Sequences". In: *KDD*. AAAI Press, pp. 359–363.

Xing, Zhengzheng, Jian Pei, Guozhu Dong, and Philip S Yu (2008). "Mining sequence classifiers for early prediction". In: *SIAM SDM*, pp. 644–655.

Xing, Zhengzheng, Jian Pei, and S Yu Philip (2012). "Early classification on time series". In: *Knowledge and information systems* 31.1, pp. 105–127.

Xing, Zhengzheng, Jian Pei, Philip S Yu, and Ke Wang (2011). "Extracting interpretable features for early classification on time series". In: *SIAM SDM*, pp. 247–258.

Xiong, Liang, Barnabás Póczos, Jeff Schneider, Andrew Connolly, and Jake VanderPlas (2011). "Hierarchical probabilistic models for group anomaly detection". In: *AISTATS*, pp. 789–797.

Yang, Yiming, Xin Liu, et al. (1999). "A re-examination of text categorization methods". In: *SIGIR*.

Ye, Lexiang and Eamonn Keogh (2009). "Time series shapelets: a new primitive for data mining". In: *SIGKDD*. ACM, pp. 947–956.

Yeh, Chin-Chia Michael et al. (2018). "Time series joins, motifs, discords and shapelets: a unifying view that exploits the matrix profile". In: *Data Min. Knowl. Discov.* 32.1, pp. 83–123.

Yu, Rose, Xinran He, and Yan Liu (2015). "Glad: group anomaly detection in social media analysis". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 10.2, pp. 1–22.

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi (2017). "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment". In: *Proceedings of the 26th international conference on world wide web*, pp. 1171–1180.

Zavrak, Sultan and Murat İskefiyeli (2020). "Anomaly-based intrusion detection from network flow features using variational autoencoder". In: *IEEE Access* 8, pp. 108346–108358.

Zemel, Rich, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork (2013). "Learning fair representations". In: *International Conference on Machine Learning*, pp. 325–333.

Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell (2018). "Mitigating unwanted biases with adversarial learning". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340.

Zhang, Hongjing and Ian Davidson (2020). "Towards Fair Deep Anomaly Detection". In: *arXiv preprint arXiv:2012.14961*.

Zhang, Jiong and Mohammad Zulkernine (2006). "Anomaly based network intrusion detection with unsupervised outlier detection". In: *2006 IEEE International Conference on Communications*. Vol. 5. IEEE, pp. 2388–2393.

Zhou, Chong and Randy C Paffenroth (2017). "Anomaly detection with robust deep autoencoders". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 665–674.