Exploiting Negative Samples: A Catalyst for Cohort Discovery in Healthcare Analytics

Kaiping Zheng [1] Horng-Ruey Chua [2 1] Melanie Herschel [3] H. V. Jagadish [4] Beng Chin Ooi [1]
James Wei Luen Yip [5 1]

Abstract

In healthcare analytics, addressing binary diagnosis or prognosis tasks presents unique challenges due to the inherent asymmetry between positive and negative samples. While positive samples, indicating patients with a disease, are defined based on stringent medical criteria, negative samples are defined in an open-ended manner and remain underexplored in prior research. To bridge this gap, we propose an innovative approach to facilitate cohort discovery within negative samples, leveraging a Shapley-based exploration of interrelationships between these samples, which holds promise for uncovering valuable insights concerning the studied disease, and related comorbidity and complications. We quantify each sample's contribution using data Shapley values, subsequently constructing the Negative Sample Shapley Field to model the distribution of all negative samples. Next, we transform this field through manifold learning, preserving the essential data structure information while imposing an isotropy constraint in data Shapley values. Within this transformed space, we pinpoint cohorts of medical interest via density-based clustering. We empirically evaluate the effectiveness of our approach on the real-world electronic medical records from National University Hospital in Singapore, yielding clinically valuable insights aligned with existing knowledge, and benefiting medical research and clinical decision-making.

1. Introduction

Healthcare analytics leverages diverse healthcare data sources to perform many analytic tasks including diagnoses (Lipton et al., 2016) and prognoses (Mould, 2012). Electronic Medical Records (EMR) are perhaps the most important of these data sources, since they play a crucial role in recording patients' essential information and providing a comprehensive view of their health conditions. The recently increasing availability of EMR data has spawned the development of healthcare analytic models for effective patient management and medical resource allocation.

Without loss of generality, let us delve into a diagnosis or prognosis problem of predicting whether a patient has developed/will develop a certain disease based on the EMR data. This problem is a binary classification, where patients who develop the disease are "positive samples," while those who do not are "negative samples." Notably, we identify the unique nature of such binary classifications in healthcare analytics, as compared to traditional classification tasks. For instance, when classifying cats vs. dogs, both positive and negative samples are based on objective facts. However, in healthcare analytics, positive samples are defined according to rigorous medical criteria, based on medical theories and experience. Contrarily, negative samples are defined in an unrestricted manner, as the complementary set of the positive samples. Consequently, the negative samples may include a wide range of healthy individuals or those outside the studied disease. This leads to an inherent asymmetry: positive samples are well-defined and bounded, while negative samples are diverse and open-ended.

Despite such fundamental asymmetry in healthcare analytics, previous research has not adequately addressed the role of negative samples. One potential reason for this research gap is the enormous challenge posed by investigating an infinitely large negative sample space, which cannot be easily addressed using existing approaches, e.g., it could be difficult to understand why general healthy individuals do not develop a disease. Nonetheless, it is crucial to investigate negative samples to comprehensively study the disease. While the disease may not have developed in these samples, some may exhibit similar symptoms or develop related con-

[1]National University of Singapore, Singapore [2]National University Hospital, Singapore [3]Universität Stuttgart, Germany [4]University of Michigan, USA [5]National University Heart Centre, Singapore. Correspondence to: Kaiping Zheng <kaiping@comp.nus.edu.sg>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

ditions, like comorbidity or complications. Hence, these negative samples urgently need close medical attention, offering clinicians a chance to better understand the disease and facilitate diagnoses, prognoses, and treatment plans.

In this paper, we aim to address this gap by exploring negative samples in healthcare analytics. Given the diversity of negative samples, it may not be meaningful to consider them all as one "group." Instead, we examine the underlying distribution of negative samples to automatically identify medically insightful groups of patients with shared characteristics, i.e., "cohorts" (Mahmood et al., 2014; Zhou et al., 2020). Such cohort discovery among negative samples can provide fresh insights to clinicians on the studied disease, e.g., comprehending the factors contributing to the absence of the disease and the development of related conditions.

Solution. We bring a unique perspective to guide our methodology design in effectively discovering cohorts among negative samples. In Section 3, we elaborate on our approach with three components. Firstly, we propose to quantify each negative sample's contribution to the prediction task using data Shapley values (Ghorbani & Zou, 2019; Rozemberczki et al., 2022). We then construct the Negative Sample Shapley Field, an inherently existing scalar field describing the distribution of all negative samples (Section 3.1). Secondly, to effectively discover cohorts, we transform the original field by manifold learning (Bengio et al., 2013) while preserving the original data structure information and ensuring that changes in data Shapley values are isotropic in all orientations (Section 3.2). Thirdly, in the transformed manifold space, we identify densely-connected clusters among the negative samples with high data Shapley values through DBSCAN (Section 3.3). These clusters identify "hot zones," our target cohorts, exhibiting similar medical characteristics with high data Shapley values.

Novelty. (i) In contrast to mainstream medical cohort studies, we adopt a distinct perspective by focusing on negative samples, and emphasize the significance of cohort discovery among negative samples, as they can reveal future positives, pathological correlations, or similar conditions. This reciprocal relationship between negative and positive samples could contribute to defining positive samples in theoretical medical research. (ii) Existing studies on data Shapley values predominantly measure the value of individual data samples, e.g., for federated learning, or apply them at finer levels for feature explainability (Rozemberczki et al., 2022). What distinguishes our paper is its innovative Shapley-based exploration of interrelationships between samples, extending beyond traditional feature-based similarity methods. It asserts that valuable cohorts should exhibit similar distributions with high data Shapley values.

Contributions. (i) We bridge the research gap caused by the asymmetry between positive and negative samples in healthcare analytics by exploring negative samples for cohort discovery. (ii) We propose an innovative approach for effective cohort discovery: constructing the Negative Sample Shapley Field, transforming the field by manifold learning with structure preservation and isotropy constraint, and discovering cohorts in the manifold space via DBSCAN. (iii) We evaluate the effectiveness of our approach using the EMR data from National University Hospital in Singapore (Section 4). Our approach reveals insights consistent with domain knowledge, verified by clinicians, and has the potential to assist medical practitioners by advancing research and enhancing clinical decision-making.

2. Problem and Our Solution

Distinctiveness of negative samples and the unbounded negative sample space. Let us take hospital-acquired acute kidney injury (AKI), a disease we strive to handle in practice, as an example. In this AKI prediction task, we aim to predict if a patient will develop AKI in the future. A positive sample is a patient who meets the stringent KDIGO criteria (Kellum et al., 2012), and has a closed definition, whereas a negative sample has an open definition without restrictions. Hence, negative samples form an unbounded space, demonstrating an asymmetry compared to positive samples.

Construction of the Negative Sample Shapley Field for cohort discovery. To facilitate the analysis of negative samples, we investigate their distribution and identify those that are most relevant to the prediction task (e.g., AKI prediction task above) and hence worth exploring. In this regard, we propose to measure the valuation of each negative sample to the task by its data Shapley value, which quantifies the sample's contribution to the prediction task. Based on such valuations, we construct a scalar field, the Negative Sample Shapley Field, in which each point is a negative sample, and the point's value is its data Shapley value. This field depicts the distribution of negative samples (see Figure 1(a) for an example). Accordingly, negative samples exhibiting high data Shapley values denote those of considerable importance to the prediction task, thus warranting particular scrutiny in our investigation. We define **"hot zones"** in this field, identified by points with high data Shapley values, as **"cohorts."** Our objective is to automatically detect these cohorts, revealing medically meaningful patterns.

Cohort discovery via manifold learning and density-based clustering. We note that the vast number of negative samples renders an exhaustive search infeasible. Although the Negative Sample Shapley Field is continuously differentiable, the high computational overhead makes it intractable to find local optima via gradient descent. To overcome this obstacle, we make the assumption that a subset of negative samples collected in clinical practice carries significant medical value, e.g., patients who visit hospitals for examinations

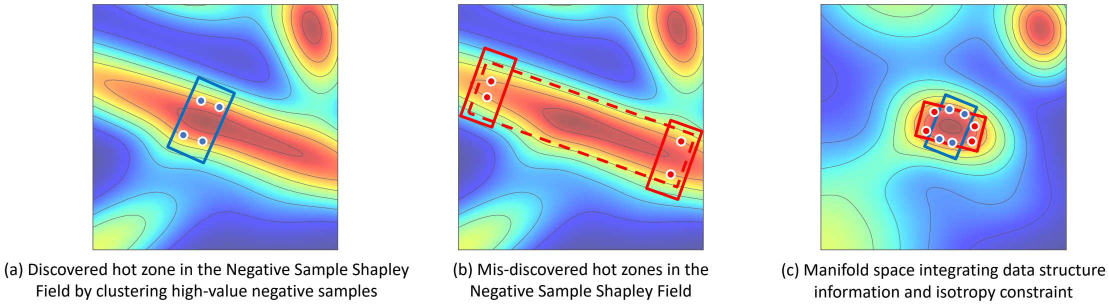

(a) Discovered hot zone in the Negative Sample Shapley Field by clustering high-value negative samples
(b) Mis-discovered hot zones in the Negative Sample Shapley Field
(c) Manifold space integrating data structure information and isotropy constraint

Figure 1: Discovery of hot zones in the Negative Sample Shapley Field.

but do not develop the disease. We posit that these real-world negative samples should be proximate to our desired hot zones in the space and can effectively sample our hot zone boundaries, which are hence of medical interest.

In Figure 1, we exemplify how to discover hot zones in the Negative Sample Shapley Field. Figure 1(a) and (b) demonstrate four points situated on the same contour line, indicating their inclusion in the same hot zone. However, only the former case yields the expected discovered cohort, while the latter leads to mis-discovery. This highlights that the originally constructed Negative Sample Shapley Field is suboptimal for cohort discovery among negative samples, due to its anisotropy in data Shapley values. To overcome this issue, we propose a manifold learning approach. Specifically, we leverage manifold learning to reduce the dimensionality of the raw, sparse EMR data to derive compact representations that not only preserve the underlying data structure information but also benefit subsequent spatial clustering analysis. Further, we introduce an isotropy constraint to ensure uniform changes in data Shapley values across all orientations, preventing the mis-discovery as in Figure 1(b). This transformed space, integrating the data structure information and the isotropy constraint, is more suitable for subsequent cohort discovery as shown in Figure 1(c).

Our objective is then to identify medically meaningful cohorts, specifically dense regions formed by negative samples with high data Shapley values in the manifold space. We set a data Shapley value threshold to extract negative samples with high values and employ the DBSCAN to detect the hot zones among them. The derived cohorts could shed light on the studied disease, its related comorbidity and complications, aiding clinicians in practical healthcare delivery.

3. Methodology

3.1. Negative Sample Shapley Field Construction

Given EMR data $\mathcal{D} = \{d_i\}$, where d_i is a sample with $i \in \{0, \dots, N-1\}$ and N denotes the total sample number. We focus on binary classification, and each d_i consists of input features and a binary label. To investigate negative samples for cohort discovery, we divide $\mathcal{D}$ into $\mathcal{D}^+$ and $\mathcal{D}^-$, representing positive and negative samples. We denote $\mathcal{D}^- = \{d_i^-\}$, where d_i^- is a negative sample with $i \in \{0, \dots, N^- - 1\}$ and N^- is the negative sample number.

Each negative sample $d_i^- = (\mathbf{x}_i, y_i)$ comprises the input features $\mathbf{x}_i$ and its binary label y_i. Our objective is to measure the value of each negative sample by quantifying its contribution to the prediction performance, which we refer to as data valuation. The data Shapley value (Ghorbani & Zou, 2019), stemming from the Shapley value in cooperative game theory, has made significant advances in data valuation (Rozemberczki et al., 2022), which inspires our proposal to calculate the data Shapley value of each negative sample as its value. Specifically, let F denote the prediction model and suppose we are interested in evaluating F's performance on a subset of negative samples $\mathcal{Q} \subseteq \mathcal{D}^-$, along with all the positive samples $\mathcal{D}^+$. We define M as the performance metric function, and then $M(\mathcal{D}^+ \cup \mathcal{Q}, F)$ is the performance achieved on the combined set of $\mathcal{D}^+$ and $\mathcal{Q}$. We define s_i as the data Shapley value for the negative sample d_i^-. s_i satisfies three properties of Shapley values: (i) null player, (ii) symmetry, and (iii) linearity, which are the essential properties of an equitable data valuation (Ghorbani & Zou, 2019). We calculate s_i as follows.

Proposition 3.1. *The data Shapley value s_i for a negative sample d_i^- is defined as:*

$$H \sum_{\mathcal{Q} \subseteq \mathcal{D}^- - \{d_i^-\}} \frac{M(\mathcal{D}^+ \cup \mathcal{Q} \cup \{d_i^-\}, F) - M(\mathcal{D}^+ \cup \mathcal{Q}, F)}{\binom{N^- - 1}{|\mathcal{Q}|}} \tag{1}$$

where H is a constant, and the summation is taken over all subsets of negative samples except d_i^-.

Equation 1 can be re-expressed in the following form:

$$s_i = E_{\pi \sim \Pi}[M(\mathcal{D}^+ \cup A_\pi^{d_i^-} \cup \{d_i^-\}, F) - M(\mathcal{D}^+ \cup A_\pi^{d_i^-}, F)] \tag{2}$$

where Π represents a uniform distribution of all the permutations among $\mathcal{D}^-$, and $A_\pi^{d_i^-}$ denotes all the negative samples before d_i^- in a permutation π. Given the exponential complexity of computing the data Shapley values for negative samples, we further adopt the Monte Carlo permutation sampling technique to approximate the computation of s_i (Castro et al., 2009). By repeating this approximation over multiple Monte Carlo permutations, we efficiently derive the estimated data Shapley value s_i. After computing the data Shapley value of each negative sample, we proceed to define the Negative Sample Shapley Field.

Definition 3.2. (Negative Sample Shapley Field) We define the Negative Sample Shapley Field $\mathcal{S}$ as an inherently existing scalar field representing the distribution of data Shapley values across all negative samples in space. In this field, each point denotes a negative sample d_i^- and is associated with its data Shapley value s_i. Therefore, $\mathcal{S}$ is a mathematical function that maps the input of each negative sample to its corresponding data Shapley value: $\mathbf{x}_i \mapsto s_i$.

With this field $\mathcal{S}$ constructed, cohort discovery among negative samples is reframed as the task of identifying "hot zones," regions within $\mathcal{S}$ with high data Shapley values.

3.2. Manifold Learning with Structure Preservation and Isotropy Constraint

As in Figure 1(a) and (b), although we hope to detect a similarly clustered cohort in the Negative Sample Shapley Field in both scenarios, the anisotropic nature of the space, i.e., the non-uniform distribution of negative samples with similar data Shapley values, presents significant challenges. To mitigate these challenges, we propose to employ manifold learning (Bengio et al., 2013) to transform the original space $\mathcal{S}$ into a new geometric space $\mathcal{S}'$. As elaborated in Section 2, to avoid mis-discovery such as Figure 1(b), we should preserve the underlying structural information in the data while imposing an isotropy constraint on the data Shapley values in $\mathcal{S}'$. The resulting $\mathcal{S}'$ will be more amenable to accurate identification of medically relevant cohorts.

We employ a stacked denoising autoencoder (SDAE) (Vincent et al., 2010) as the backbone model for manifold learning due to its capability of handling input data corruption. Further, we integrate the isotropy constraint while preserving the data structure information in $\mathbf{x}_i$. Consider an SDAE consisting of K denoising autoencoders (DAEs). For the k-th DAE ($k \in \{0, \dots, K-1\}$), the encoder takes $\mathbf{h}_i^{(k)}$ as input, where $\mathbf{h}_i^{(0)} = \mathbf{x}_i$ is the original input. We define $\tilde{\mathbf{h}}_i^{(k)}$ as the corrupted version of $\mathbf{h}_i^{(k)}$ with masking noise generated by a stochastic mapping, $\tilde{\mathbf{h}}_i^{(k)} \sim g_{\mathcal{D}}(\tilde{\mathbf{h}}_i^{(k)}|\mathbf{h}_i^{(k)})$, which randomly sets a fraction of the elements of $\mathbf{h}_i^{(k)}$ to 0. The encoder transforms the corrupted $\tilde{\mathbf{h}}_i^{(k)}$ into an abstract representation $\hat{\mathbf{h}}_i^{(k+1)}$, which is then used by the decoder to recover the uncorrupted $\mathbf{h}_i^{(k)}$. This process enables the DAE to extract useful information for denoising, which is crucial for healthcare analytics given the missing data and noise in real-world EMR data (Lasko et al., 2013). The model architecture is depicted in Figure 7 of Appendix D.1.

Encoder of the k-th DAE. The encoder of the k-th DAE transforms the corrupted representation using an affine transformation followed by a non-linear activation function:

$$\hat{\mathbf{h}}_i^{(k+1)} = f_\theta^{(k+1)}(\tilde{\mathbf{h}}_i^{(k)}) = \sigma(\mathbf{W}_\theta^{(k+1)}\tilde{\mathbf{h}}_i^{(k)} + \mathbf{b}_\theta^{(k+1)}) \quad (3)$$

where $f_\theta^{(k+1)}(\cdot)$ is the encoder with $\mathbf{W}_\theta^{(k+1)}$ and $\mathbf{b}_\theta^{(k+1)}$ as the weight matrix and bias vector, respectively. The rectified linear unit (ReLU) activation function $\sigma(\cdot)$ is used for non-linearity.

Decoder of the k-th DAE. The derived abstract representation $\hat{\mathbf{h}}_i^{(k+1)}$ is subsequently mapped back to the previous latent space in the decoder, with the aim of recovering the uncorrupted representation:

$$\mathbf{z}_i^{(k)} = f_\phi^{(k+1)}(\hat{\mathbf{h}}_i^{(k+1)}) = \sigma(\mathbf{W}_\phi^{(k+1)}\hat{\mathbf{h}}_i^{(k+1)} + \mathbf{b}_\phi^{(k+1)}) \quad (4)$$

where $f_\phi^{(k+1)}(\cdot)$ is the decoder of the k-th DAE, with $\mathbf{W}_\phi^{(k+1)}$, $\mathbf{b}_\phi^{(k+1)}$ and the ReLU activation.

Structure Preservation. To attain a stable and robust abstract representation that is resilient to data corruption, it is crucial to recover the uncorrupted representation as accurately as possible. To achieve this, we adopt a reconstruction loss that preserves the data structure information. For a given batch of negative samples $\mathcal{B}$, the reconstruction loss per sample within this batch is defined as:

$$\mathcal{L}_{rec}^{(k)} = \frac{1}{|\mathcal{B}|}\sum_{i \in \mathcal{B}} \|\mathbf{h}_i^{(k)} - \mathbf{z}_i^{(k)}\|^2 \quad (5)$$

Isotropy Constraint. In addition to the reconstruction loss, it is essential to enforce an isotropy constraint to ensure that data Shapley value changes are uniform across orientations. To achieve this, we introduce a penalty that accounts for the change in data Shapley values relative to the Euclidean distance between two samples:

$$\mathcal{L}_{iso}^{(k)} = \frac{1}{|\mathcal{B}|^2}\sum_{i,j \in \mathcal{B}} (\frac{s_j - s_i}{\mu_{ij}})^2 \quad (6)$$

where i, j are two samples with s_i, s_j as their data Shapley values, μ_{ij} as the distance between $\hat{\mathbf{h}}_i^{(k+1)}$ and $\hat{\mathbf{h}}_j^{(k+1)}$. The overall loss is then a weighted sum of the reconstruction loss and the isotropy penalty, jointly integrating the structural information and the isotropy constraint:

$$\mathcal{L}^{(k)} = \omega_{rec}\mathcal{L}_{rec}^{(k)} + \omega_{iso}\mathcal{L}_{iso}^{(k)} \quad (7)$$

The weights ω_{rec} and ω_{iso} are introduced to address the issue of the two loss terms being on different scales, which ensures that both losses are decreased at similar rates, leading to a better balance between the optimization objectives (Liu et al., 2019; Groenendijk et al., 2021). Specifically, the weights are set to the loss ratio between the current iteration (t) and the previous iteration ($t-1$):

$$\omega_{rec} = \mathcal{L}_{rec}^{(k)}(t)/\mathcal{L}_{rec}^{(k)}(t-1),\ \omega_{iso} = \mathcal{L}_{iso}^{(k)}(t)/\mathcal{L}_{iso}^{(k)}(t-1) \tag{8}$$

We have introduced how to learn the k-th DAE using the loss function in Equation 7. The corrupted input is only used during the initial training to learn robust feature extractors. After the encoder $f_\theta^{(k+1)}(\cdot)$ is trained, it will be applied to the clean input:

$$\mathbf{h}_i^{(k+1)} = f_\theta^{(k+1)}(\mathbf{h}_i^{(k)}) = \sigma(\mathbf{W}_\theta^{(k+1)}\mathbf{h}_i^{(k)} + \mathbf{b}_\theta^{(k+1)}) \tag{9}$$

$\mathbf{h}_i^{(k+1)}$ is used as input for the $(k+1)$-th DAE to continue the repeated training process. When the last DAE is trained, we obtain the encoded representation $\mathbf{h}_i^{(K)}$ in the manifold space $\mathcal{S}'$, which preserves the data structure information in $\mathbf{x}_i$ and integrates the desired isotropy constraint.

3.3. Cohort Discovery Among High Data Shapley Value Negative Samples

We proceed to perform cohort discovery in the encoded manifold space $\mathcal{S}'$, where each negative sample's input $\mathbf{x}_i$ is transformed into $\mathbf{h}_i^{(K)}$. We begin by setting a threshold value τ to filter out negative samples with data Shapley values below τ, which focuses our analysis on negative samples with high data Shapley values, i.e., high contributions to the prediction task. Among the remaining negative samples with high data Shapley values, we target to detect the hot zones in $\mathcal{S}'$, which may represent medically meaningful cohorts of arbitrary shape.

To achieve this, we employ DBSCAN, short for density-based spatial clustering of applications with noise (Ester et al., 1996; Gan & Tao, 2015; Schubert et al., 2017) on such samples. The core idea of DBSCAN is to group samples that are close to each other in the manifold space $\mathcal{S}'$ into clusters, which could locate potential cohorts, while treating the remaining samples as noise or outliers. DBSCAN has three main steps: (i) identify the points within each point's ε- neighborhood and determine the "core points" with over P_{min} neighbors; (ii) detect the connected components of the core points in the neighbor graph, disregarding any non-core points; (iii) assign each non-core point to the clusters which are the ε-neighborhood of the point; otherwise, label the point as noise. This process results in a set of clusters $\{C_1, C_2, \ldots, C_R\}$ and a set of noisy samples Ψ. Given the clusters, we define cohorts as follows.

Definition 3.3. (Cohorts) For a dense cluster C_r identified by DBSCAN, we consider each of its core points and define a spherical space with the core point as its center and ε as its radius. The joint space of all such spherical spaces is the cohort we aim to discover from this cluster.

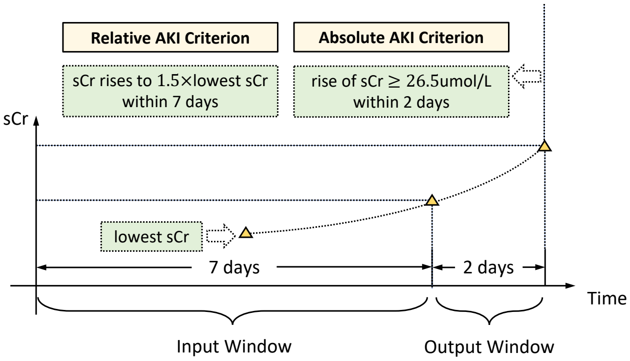


Figure 2: Definition of absolute AKI and relative AKI in hospital-acquired AKI prediction.

Table 1: Key statistics of our dataset.

Statistics	Our Dataset
# of admissions	20732
# of positive samples	911
# of negative samples	19821
# of lab tests	709
Input Window	7 days
Output Window	2 days

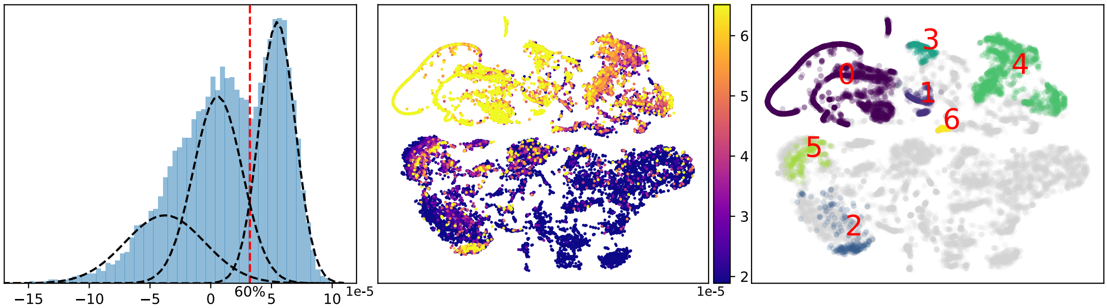
These discovered cohorts provide a promising avenue for further exploration of medically meaningful patterns in EMR data analytics, potentially revealing important insights.

4. Experimental Evaluation

In this section, we first detail the experimental setup. We then evaluate our proposal's capability for cohort discovery in AKI prediction (Section 4.2), delve into the discovered cohorts for in-depth analysis (Section 4.3), and validate the effectiveness of its individual components (Section 4.4). To provide a comprehensive evaluation, we also present supplementary experiments in Appendix G, encompassing an ablation study on the impact of our proposed isotropy constraint (G.1) and comparisons with diverse baselines, including contrastive principal component analysis (G.2), positive-unlabelled learning methods (G.3), influence function-based data valuation (G.4), deep clustering methods (G.5) and clustering all negative samples (G.6). Additionally, we validate the broad applicability of our proposal by conducting cohort discovery analysis on the MIMIC-III public benchmark dataset (Johnson et al., 2016) (G.7).

4.1. Experimental Setup

We focus on hospital-acquired AKI (short for acute kidney injury), which is a disease we strive to handle in our medical practice. According to the KDIGO criteria (Kellum et al., 2012), the definition of AKI is based on the rise of sCr (i.e.,



(a) Data Shapley value histogram among all negative samples

(b) Data Shapley value distribution among all negative samples in the manifold space

(c) Discovered cohorts among high data Shapley value negative samples

Figure 3: Cohort discovery of our proposal for AKI prediction.

serum creatinine), a lab test, beyond a threshold limit within a defined timeline. The definition includes two criteria: absolute AKI and relative AKI, as depicted in Figure 2. Absolute AKI is defined as an increase in sCr of more than 26.5 umol/L within the past two days. Relative AKI, on the other hand, is defined as a rise in sCr of 1.5 times or higher compared to the lowest sCr value within the last seven days.

In hospital-acquired AKI prediction, our goal is to predict whether a patient will develop AKI in hospital with a two-day prediction lead time. We evaluate our approach on the EMR data from National University Hospital in Singapore, containing 709 lab tests as input features. Each hospitalized admission in the data is treated as a sample for analysis. In total, we receive 20,732 admissions, with 911 of them resulting in AKI development. We partition the dataset into 90% training data and 10% testing data.

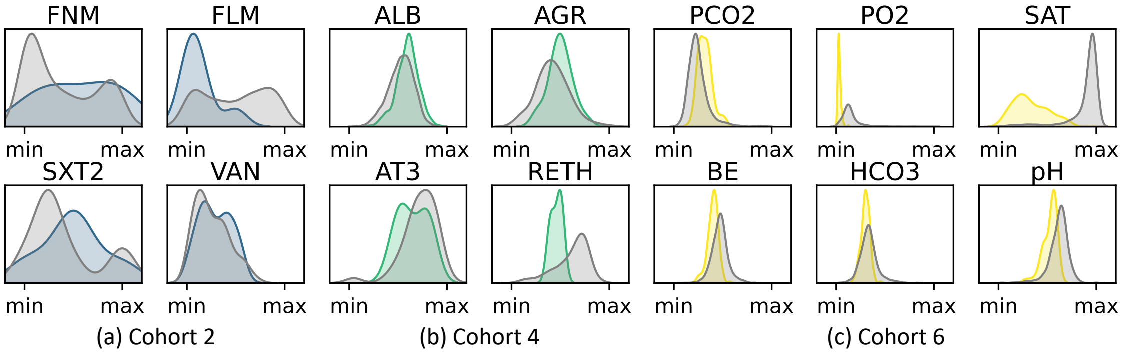
For positive samples where AKI develops during the hospital stay, we record the time of AKI detection and define a two-day window, referred to as the "Output Window," that counts backward from the detection time. This window is not used as input but is crucial in medical practice as it provides a 48-hour lead time, enabling clinicians to take timely interventions following AKI prediction if necessary. The "Input Window," which serves as input for analysis, spans seven days prior to the Output Window. The relationship between the Input Window and the Output Window is depicted in Figure 2. For negative samples, the time of the last recorded lab test is used to determine both the Output Window and the Input Window, respectively. In summary, our approach utilizes 709 lab tests within the Input Window to predict the likelihood of each sample developing AKI after the Output Window. We perform the min-max standardization on the lab test values and then calculate the average to derive input features. Table 1 presents key statistics of our dataset for hospital-acquired AKI prediction.

We employ the logistic regression (LR) model to compute the data Shapley value for each negative sample, using the area under the ROC curve (AUC) as the evaluation metric. More implementation details are elaborated in Appendices C.3, D.2 and E.3. We conduct the experimental evaluation on a server equipped with two Intel Xeon Gold 6248R CPUs, 768GB of memory, and eight NVIDIA V100 GPUs interconnected by NVLINK, using PyTorch 1.12.1.

4.2. Cohort Discovery Results

We present the cohort discovery results in Figure 3, where we first display the data Shapley value histogram among all the negative samples in Figure 3(a). It is noteworthy that this histogram can be well fitted by a Gaussian mixture model, consisting of three distinct and interesting components. The first component on the left represents the negative samples with negative data Shapley values. These samples have a negative impact on the prediction task, meaning that they are detrimental to predicting the AKI occurrence. In prior studies, one generally plausible explanation for the presence of such samples is the existence of mislabeled data (Ghorbani & Zou, 2019). However, for a representative acute disease like AKI, these negative samples are highly likely to be positive samples in the future but have not yet exhibited symptoms of AKI within the monitored time duration. Moving on to the second component in the middle, we observe that its data Shapley values are centered around a mean value close to zero. This implies that these negative samples are generally healthy without any apparent AKI-associated risk factors. Notably, these healthy samples constitute a relatively significant portion of the data, which is commonly observed in clinical practice and aligns with our initial expectations. The third component on the right represents negative samples that are particularly valuable for the prediction task and merit special attention in our study. To further investigate these samples, we introduce a separation line between the second and third components, i.e., a threshold 60% to exclude the lower 60% negative samples based on their data Shapley values while retaining the remaining 40% for further analysis. Our focus is on these 40% samples for identifying the hot zones, as illustrated in Figure 1.

The distribution of all negative samples, in terms of their data Shapley values in the manifold space, is presented in



(a) Cohort 2 (b) Cohort 4 (c) Cohort 6

Figure 4: Lab test patterns of discovered Cohorts 2, 4, and 6. In each cohort, the colored region (blue, green, and yellow) represents the lab test value probability density of the samples in the cohort, while the grey region denotes that of all the other samples outside the cohort.

Figure 3(b). Upon performing DBSCAN on the extracted 40% samples with high data Shapley values (points brighter than dark blue), we identify seven distinct cohorts of interest that are visually displayed using t-SNE plots in Figure 3(c), where grey points are either with low data Shapley values or labeled as noise by DBSCAN. We observe that these discovered cohorts are distinguishable from one another, potentially corresponding to medically meaningful patterns.

4.3. In-depth Analysis of Discovered Cohorts

Cohort 2: inflammatory cohort. Figure 4(a) and (b) indicate an altered neutrophil-to-lymphocyte ratio (NLR) (Zahorec et al., 2001) in this patient group, marked by a more uniform trend of neutrophils (FNM) across feature values and relatively lower feature values of lymphocytes (FLM), compared with those of other negative samples. Altered NLR is often tied to infectious and inflammatory conditions, suggesting an overactive immune response leading to reduced lymphocyte counts (Nathan, 2006; Dhabhar, 2009). An elevated NLR, a reliable inflammatory marker, indicates a propensity for invasive infections (Huang et al., 2020b); yet, the more stable FNM trend in Cohort 2 may suggest less severe inflammatory response and correspondingly no AKI in patients. The feature Cotrimoxazole (SXT2) is the minimal inhibitory concentration of SXT2, an antibiotic, which is the lowest concentration of the antibiotic at which the specific bacterial growth of interest is completely inhibited in-vitro (Kowalska-Krochmal & Dudek-Wicher, 2021). SXT2 MIC is often tested against staphylococcus species and values may therefore reflect the degree of underlying antibiotic resistance by the pathogen that infected respective patients; treatment resistance affects the control of infection which in turn impacts on end-organ or kidney injury. Meanwhile, the level of Vancomycin (VAN), administered to treat infections associated with methicillin-resistant staphylococcus aureus (Holmes & Howden, 2014), are found elevated in the serum of these patients. Severe infections can cause systemic inflammatory response syndrome and kidney injury. Antibiotics like vancomycin can worsen kidney stress and have nephrotoxic properties (Wu & Huang, 2018), potentially leading to kidney dysfunction during treatment. However, modern medical practice can effectively manage these cases. Infections are promptly treated with broad-spectrum antibiotics. Dosage of vancomycin is routinely reduced in response to potentially toxic serum levels and kept within safe limits in clinical practice; thus, the patients may not develop significant AKI (Goldstein et al., 2016).

Cohort 4: hepatic and hematological disorders cohort. As delineated in Figure 3(c), Cohort 4 exhibits an augmented region and an increased quantity of sampling points, indicative of a more expansive patient population. A comprehensive analysis of the lab test indicator distribution for this cohort, portrayed in Figure 4(b), reveals differences in levels of serum proteins. Specifically, derangements in levels of albumin (ALB) and the albumin-globulin ratio (AGR) signify aberrant protein synthesis in patients. Low serum ALB and abnormal AGR are associated with hepatic dysfunction or hematological diseases such as myeloma or monoclonal gammopathy (Spinella et al., 2016; Laudin et al., 2020); comparatively, higher levels of ALB and AGR may be found in negative samples. Hepatic diseases can lead to impaired production of other proteins such as antithrombin III (AT3) (Knot et al., 1984); AT3 levels fall precipitously in the early phases of severe sepsis (Mesters et al., 1996), or undergo accelerated consumption in disseminated intravascular coagulation (Mammen, 1998). Diminished reticulocyte hemoglobin (RETH) is associated with iron deficiency anemia (Auerbach et al., 2021), and could either be linked to hematological disorders or chronic kidney disease. In addition, imbalances in albumin and globulin may also be associated with dehydration. Therefore, our observation derived from Cohort 4 may support the pathophysiological relationship that exists between disorders of the hematological and hepatic systems, which increases the propensity for kidney disease. Clinicians should exercise vigilance in care when managing these cases.

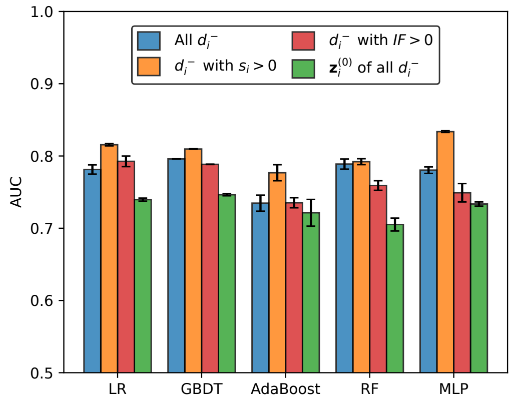


Figure 5: AKI prediction performance for three different settings of our proposal and IF-based data valuation.

Cohort 6: respiratory failure and metabolic acidosis cohort. Figure 4(c) reveals significant metabolic imbalances in patients, leading to an acid-base imbalance. Specifically, increased carbon dioxide pressure (PCO2), reduced oxygen pressure (PO2), and insufficient blood oxygen saturation (SAT) suggest respiratory failure (Breen, 2001). Concurrently, reduced base excess (BE), bicarbonate ion (HCO3) levels, and blood pH values hint at metabolic acidosis, indicating possible acute illnesses causing lactic or ketoacidosis (Kraut & Madias, 2010), or impaired renal acidification in early kidney tubular injury (Winaver et al., 1986). These results suggest potential severe respiratory complications, that arise from advanced pneumonia, heart failure-induced pulmonary edema, or chronic obstructive pulmonary disease (COPD) (Kempker et al., 2020). Alternatively, acute conditions like hypoxia, shock, or severe infection could disrupt aerobic metabolism, leading to anaerobic glucose conversion to lactate, which accumulates in the bloodstream and causes acidosis. The former acute disease states are risk factors for AKI. Severe pneumonia and infections and heart failure could cause end-organ injury including kidney tubular injury, but the latter may remain subclinical and not necessarily manifest with raised serum creatinine and AKI (Yang et al., 2022).

4.4. Validation of Effectiveness of Each Component

We validate the effectiveness of each component in our approach for AKI prediction. Specifically, we evaluate three settings of the negative sample usage in the training data (with positive samples the same): (i) all d_i^-: use all negative samples; (ii) d_i^- with $s_i > 0$: only use the negative samples with positive data Shapley values; (iii) $\mathbf{z}_i^{(0)}$ of all d_i^-: use the decoded representations from the SDAE-based manifold learning. $\mathbf{z}_i^{(0)}$ is in the same dimension as the raw input but is in the decoding space after transformation by SDAE. We further compare with another data valuation baseline: (iv) d_i^- with $IF > 0$: use the negative samples with positive IF values, where IF denotes influence functions measuring how the model changes when a single sample's weight is altered slightly (Weisberg & Cook, 1982). We evaluate several widely adopted classifiers: LR, gradient-boosting decision tree (GBDT), adaptive boosting (AdaBoost), random forest (RF), and multilayer perceptron (MLP). The experimental results in AUC (mean ± std) from five repeats are illustrated in Figure 5.

Effectiveness of Data Shapley Values for Negative Samples. By comparing (i) and (ii), it is clear that after removing negative samples with data Shapley values smaller than 0, all the classifiers exhibit an improvement in AUC. This substantiates the rationale behind our approach of associating samples of significant medical concern with their respective data Shapley values. Further, the efficacy of approximating data Shapley values through Monte Carlo permutation sampling is validated. In comparing data valuation methods, "d_i^- with $IF > 0$" underperforms "d_i^- with $s_i > 0$," as IF-based data valuation deletes informative negative samples, limiting its effectiveness in AKI prediction. Further, IF fails to satisfy equitability conditions due to its inability to consider complex sample interactions, deviating from our focus on cohort discovery and posing robustness issues (Ghorbani et al., 2019). Hence, IF-based data valuation is unsuitable for AKI prediction. Detailed analyses of this comparison are in Appendices G.4.

Effectiveness of Manifold Learning. By changing the input data from the raw space to the decoder's output space after our proposed SDAE-based manifold learning (settings (i) vs. (iii)), we observe a moderate decrease in AUC, approximately 5% in most classifiers. This decrease aligns with our expectations, as the transformation in SDAE introduces a certain level of information loss. However, the performance degradation remains within an acceptable range. These findings demonstrate that our proposed manifold learning manages to preserve the original data structure information and effectively model the original raw data space, despite a significant reduction in data dimension from 709 to 64 (as detailed in Appendix D.2). Thus, this corroborates our design rationale of employing SDAE for manifold learning with structure preservation and isotropy constraint. Further, we compare with contrastive principal component analysis (cPCA) (Abid et al., 2017; 2018) and our proposal without isotropy constraint, with results detailed in Appendices G.2 and G.1. Both methods primarily identify one large cohort and a few smaller cohorts, failing to identify medically meaningful cohorts. This limitation is also observed in deep clustering methods like deep clustering network (DCN) (Yang et al., 2017) and deep embedded K-means clustering (DEKM) (Guo et al., 2021), as discussed in Appendix G.5. Common to these methods is a process entailing dimensionality reduction followed by clustering on the embedded representations. However, their lack of our proposed isotropy constraint—crucial for uniform changes

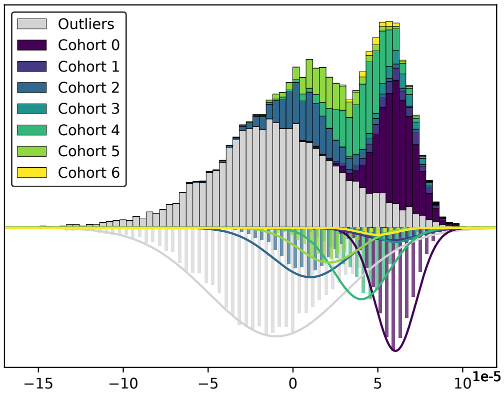


Figure 6: Data Shapley value histogram of the samples within our discovered cohorts.

in data Shapley values across orientations—limits their efficacy. This is because they overlook the key insight that negative samples indicate different types with distinct symptoms, implying diverse cohorts with varied data Shapley value distributions. Consequently, they cannot effectively discover medically relevant cohorts.

Effectiveness of Cohort Discovery. We further validate the ability of our approach to decompose high data Shapley value samples into distinct, medically relevant cohorts. Figure 6 presents the data Shapley value histogram of our identified cohorts, with the upper part aligned with Figure 3(a) but color-coded by cohort proportion. The lower part shows each cohort's data Shapley value distribution. We note seven cohorts effectively partition Figure 3(a)'s third component into Gaussian distributions, implying consistent data Shapley values within each cohort. Cohort 2, identified as the inflammatory group, exhibits relatively lower data Shapley values, as immune abnormalities cannot serve as specific features for kidney injury. Conversely, Cohorts 4 and 6, involving critical metabolic systems, display higher data Shapley values, which indicates their significant medical relevance to AKI prediction. These observations confirm the homogeneity in each cohort due to DBSCAN's detection capability and similarity in data Shapley values, further substantiating our proposed isotropy constraint in manifold learning. Clinically validated by medical professionals, our derived cohort discovery results validate the correctness of the outcomes and the medical utility of our approach.

5. Related Work

The Shapley value, originally introduced in cooperative game theory (Shapley et al., 1953), offers a solution for the equitable distribution of a team's collective value among its individual members (Chalkiadakis et al., 2011). Notable applications of the Shapley value in machine learning encompass data valuation, feature selection, explainable machine learning, etc (Lundberg & Lee, 2017; Ghorbani & Zou, 2019; Williamson & Feng, 2020; Liu et al., 2022; Rozemberczki et al., 2022). Among these, data valuation holds significance in quantifying the contributions of individual data samples toward predictive models. In this research line, the data Shapley value (Ghorbani & Zou, 2019) presents an equitable valuation framework for data value quantification with subsequent research focusing on enhancing computational efficiency (Jia et al., 2019; Ghorbani et al., 2020).

Representation learning is a crucial research area contributing to the success of many machine learning algorithms (Bengio et al., 2013). Among the representation learning methods, manifold learning stands out due to its capability of reducing the dimensionality and visualizing the underlying structure of the data. Traditional manifold learning methods include Isomap (Tenenbaum et al., 2000), locally linear embedding (Roweis & Saul, 2000), and multidimensional scaling (Borg & Groenen, 2005). In recent years, autoencoders (AEs) have gained significant attention, offering efficient and effective representations of unlabeled data. Researchers develop various AE variants for specific application scenarios, among which DAEs and their advanced stacked variant SDAEs (Vincent et al., 2010) are highly suitable to tackle EMR data, where missing and noisy data remains a notorious issue (Lasko et al., 2013).

DBSCAN, short for density-based spatial clustering of applications with noise, is introduced to alleviate the burden of parameter selection for users, facilitate the discovery of arbitrarily shaped clusters, and demonstrate satisfactory efficiency when dealing with large datasets (Ester et al., 1996; Gan & Tao, 2015; Schubert et al., 2017).

6. Conclusion

This paper proposes to examine negative samples for cohort discovery in healthcare analytics, which has not been explored in prior research. In pursuit of this goal, we delve into an innovative, Shapley-based approach to uncover interrelationships among these samples, positing that cohorts of medical significance should manifest similar distributions with high data Shapley values. In particular, we propose to measure each negative sample's contribution to the prediction task via its data Shapley value and construct the Negative Sample Shapley Field to model the distribution of all negative samples. To enhance the cohort discovery quality, we transform this original field into an embedded space using manifold learning, incorporating the original data structure information and isotropy constraint. In the transformed space, we manage to identify medically meaningful cohorts within negative samples by DBSCAN. The experiments on the EMR data from National University Hospital in Singapore demonstrate the effectiveness of our proposal. Further, the medical insights derived from our discovered cohorts are validated by clinicians, underscoring our approach's substantial medical value.

Acknowledgements

We would like to thank the anonymous reviewers for their constructive comments. NUS' research is partially supported by The Lee Foundation in the form of its NUS Lee Kong Chian Centennial Professorship grant. Horng-Ruey Chua is supported by the National Medical Research Council, Singapore, Transition Award 2019 MOH-000100-00. Melanie Herschel is partially funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2075 – 390740016. H. V. Jagadish is supported in part by NSF grants 2106176 and 2312931.

Impact Statement

Our approach, while based on practical clinical observations, is underpinned by profound fundamental insights. Specifically, in classification scenarios where positive labels are defined through domain knowledge, the data valuation of negative samples can serve as a representation of the effective adversarial degree against the positive definition. Such an approach is entirely novel and has not been considered in previous literature. Our study explores the use of this novel representation for clustering analysis of negative samples, significantly aiding in uncovering the distribution characteristics of negative samples collected in real-world scenarios. We have validated its efficacy in medical applications and believe this representation holds considerable potential for further utilization beyond clustering in the healthcare domain. We are confident in its superiority as a state-of-the-art methodology for patient cohort discovery and its potential impact on subsequent related research directions.

References

Abid, A., Bagaria, V. K., Zhang, M. J., and Zou, J. Y. Contrastive principal component analysis. *CoRR*, abs/1709.06716, 2017.

Abid, A., Zhang, M. J., Bagaria, V. K., and Zou, J. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nature communications*, 9(1):2134, 2018.

Alain, G. and Bengio, Y. What regularized auto-encoders learn from the data-generating distribution. *J. Mach. Learn. Res.*, 15(1):3563–3593, 2014.

Auerbach, M., Staffa, S. J., and Brugnara, C. Using reticulocyte hemoglobin equivalent as a marker for iron deficiency and responsiveness to iron therapy. In *Mayo Clinic Proceedings*, volume 96, pp. 1510–1519. Elsevier, 2021.

Bai, T., Zhang, S., Egleston, B. L., and Vucetic, S. Interpretable representation learning for healthcare via capturing disease progression through time. In *KDD*, pp. 43–51. ACM, 2018.

Bengio, Y., Courville, A. C., and Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, 2013.

Borg, I. and Groenen, P. J. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

Breen, P. H. Arterial blood gas and ph analysis: clinical approach and interpretation. *Anesthesiology Clinics of North America*, 19(4):885–906, 2001.

Cai, Q., Zheng, K., Ooi, B. C., Wang, W., and Yao, C. ELDA: learning explicit dual-interactions for healthcare analytics. In *ICDE*, pp. 393–406. IEEE, 2022.

Cai, Q., Zheng, K., Jagadish, H. V., Ooi, B. C., and Yip, J. W. L. Cohortnet: Empowering cohort discovery for interpretable healthcare analytics. *Proc. VLDB Endow.*, 2024.

Cai, S., Zheng, K., Chen, G., Jagadish, H. V., Ooi, B. C., and Zhang, M. Arm-net: Adaptive relation modeling network for structured data. In *SIGMOD Conference*, pp. 207–220. ACM, 2021.

Cao, W., Wang, D., Li, J., Zhou, H., Li, L., and Li, Y. BRITS: bidirectional recurrent imputation for time series. In *NeurIPS*, pp. 6776–6786, 2018.

Castro, J., Gómez, D., and Tejada, J. Polynomial calculation of the shapley value based on sampling. *Comput. Oper. Res.*, 36(5):1726–1730, 2009.

Chalkiadakis, G., Elkind, E., and Wooldridge, M. J. *Computational Aspects of Cooperative Game Theory*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2011.

Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

Chen, C., Liang, J., Ma, F., Glass, L., Sun, J., and Xiao, C. UNITE: uncertainty-based health risk prediction leveraging multi-sourced data. In *WWW*, pp. 217–226. ACM / IW3C2, 2021.

Chua, H.-R., Zheng, K., Vathsala, A., Ngiam, K.-Y., Yap, H.-K., Lu, L., Tiong, H.-Y., Mukhopadhyay, A., MacLaren, G., Lim, S.-L., et al. Health care analytics with time-invariant and time-variant feature importance to predict hospital-acquired acute kidney injury: observational longitudinal study. *Journal of Medical Internet Research*, 23(12):e30805, 2021.

Dhabhar, F. S. Enhancing versus suppressive effects of stress on immune function: implications for immunoprotection and immunopathology. *Neuroimmunomodulation*, 16(5): 300–317, 2009.

Elkan, C. and Noto, K. Learning classifiers from only positive and unlabeled data. In *KDD*, pp. 213–220. ACM, 2008.

Ester, M., Kriegel, H., Sander, J., and Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pp. 226–231. AAAI Press, 1996.

Gan, J. and Tao, Y. DBSCAN revisited: Mis-claim, unfixability, and approximation. In *SIGMOD Conference*, pp. 519–530. ACM, 2015.

Ghorbani, A. and Zou, J. Y. Data shapley: Equitable valuation of data for machine learning. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2242–2251. PMLR, 2019.

Ghorbani, A., Abid, A., and Zou, J. Y. Interpretation of neural networks is fragile. In *AAAI*, pp. 3681–3688. AAAI Press, 2019.

Ghorbani, A., Kim, M. P., and Zou, J. A distributional framework for data valuation. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3535–3544. PMLR, 2020.

Goldstein, S. L., Mottes, T., Simpson, K., Barclay, C., Muething, S., Haslam, D. B., and Kirkendall, E. S. A sustained quality improvement program reduces nephrotoxic medication-associated acute kidney injury. *Kidney international*, 90(1):212–221, 2016.

Grimes, D. A. and Schulz, K. F. Cohort studies: marching towards outcomes. *The Lancet*, 359(9303):341–345, 2002.

Groenendijk, R., Karaoglu, S., Gevers, T., and Mensink, T. Multi-loss weighting with coefficient of variations. In *WACV*, pp. 1468–1477. IEEE, 2021.

Guo, W., Lin, K., and Ye, W. Deep embedded k-means clustering. In *ICDM (Workshops)*, pp. 686–694. IEEE, 2021.

Holmes, N. E. and Howden, B. P. What's new in the treatment of serious mrsa infection? *Current opinion in infectious diseases*, 27(6):471–478, 2014.

Huang, Y., Lyu, X., Li, D., Wang, L., Wang, Y., Zou, W., Wei, Y., and Wu, X. A cohort study of 676 patients indicates d-dimer is a critical risk factor for the mortality of covid-19. *PloS one*, 15(11):e0242045, 2020a.

Huang, Z., Fu, Z., Huang, W., and Huang, K. Prognostic value of neutrophil-to-lymphocyte ratio in sepsis: A meta-analysis. *The American journal of emergency medicine*, 38(3):641–647, 2020b.

Jia, R., Dao, D., Wang, B., Hubis, F. A., Hynes, N., Gürel, N. M., Li, B., Zhang, C., Song, D., and Spanos, C. J. Towards efficient data valuation based on the shapley value. In *AISTATS*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1167–1176. PMLR, 2019.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Kellum, J. A., Lameire, N., Aspelin, P., Barsoum, R. S., Burdmann, E. A., Goldstein, S. L., Herzog, C. A., Joannidis, M., Kribben, A., Levey, A. S., et al. Kidney disease: improving global outcomes (kdigo) acute kidney injury work group. kdigo clinical practice guideline for acute kidney injury. *Kidney international supplements*, 2(1): 1–138, 2012.

Kempker, J. A., Abril, M. K., Chen, Y., Kramer, M. R., Waller, L. A., and Martin, G. S. The epidemiology of respiratory failure in the united states 2002–2017: A serial cross-sectional study. *Critical Care Explorations*, 2(6), 2020.

Knot, E., Ten Cate, J., Drijfhout, H., Kahlé, L., and Tytgat, G. Antithrombin iii metabolism in patients with liver disease. *Journal of clinical pathology*, 37(5):523–530, 1984.

Kowalska-Krochmal, B. and Dudek-Wicher, R. The minimum inhibitory concentration of antibiotics: Methods, interpretation, clinical relevance. *Pathogens*, 10(2):165, 2021.

Kraut, J. A. and Madias, N. E. Metabolic acidosis: pathophysiology, diagnosis and management. *Nature Reviews Nephrology*, 6(5):274–285, 2010.

Lasko, T. A., Denny, J. C., and Levy, M. A. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PloS one*, 8(6):e66341, 2013.

Laudin, G. E., Levay, P. F., and Coetzer, B. Globulin fraction and albumin: globulin ratio as a predictor of mortality in a south african multiple myeloma cohort. *International Journal of Hematologic Oncology*, 9(3):IJH27, 2020.

Lee, C., Luo, Z., Ngiam, K. Y., Zhang, M., Zheng, K., Chen, G., Ooi, B. C., and Yip, J. W. L. Big healthcare data analytics: Challenges and applications. In *Handbook of*

Large-Scale Distributed Computing in Smart Healthcare, pp. 11–41. Springer, 2017.

Lipton, Z. C., Kale, D. C., Elkan, C., and Wetzel, R. C. Learning to diagnose with LSTM recurrent neural networks. In *ICLR (Poster)*, 2016.

Liu, S., Johns, E., and Davison, A. J. End-to-end multi-task learning with attention. In *CVPR*, pp. 1871–1880. Computer Vision Foundation / IEEE, 2019.

Liu, Z., Chen, Y., Yu, H., Liu, Y., and Cui, L. Gtg-shapley: Efficient and accurate participant contribution evaluation in federated learning. *ACM Trans. Intell. Syst. Technol.*, 13(4):60:1–60:21, 2022.

Lundberg, S. M. and Lee, S. A unified approach to interpreting model predictions. In *NIPS*, pp. 4765–4774, 2017.

Mahmood, S. S., Levy, D., Vasan, R. S., and Wang, T. J. The framingham heart study and the epidemiology of cardiovascular disease: a historical perspective. *The lancet*, 383(9921):999–1008, 2014.

Makhzani, A. and Frey, B. J. k-sparse autoencoders. In *ICLR (Poster)*, 2014.

Mammen, E. F. Antithrombin: its physiological importance and role in dic. In *Seminars in thrombosis and hemostasis*, volume 24, pp. 19–25. Copyright© 1998 by Thieme Medical Publishers, Inc., 1998.

Mesters, R. M., Mannucci, P. M., Coppola, R., Keller, T., Ostermann, H., and Kienast, J. Factor viia and antithrombin iii activity during severe sepsis and septic shock in neutropenic patients. 1996.

Mordelet, F. and Vert, J. A bagging SVM to learn from positive and unlabeled examples. *Pattern Recognit. Lett.*, 37:201–209, 2014.

Mould, D. Models for disease progression: new approaches and uses. *Clinical Pharmacology & Therapeutics*, 92(1): 125–131, 2012.

Nathan, C. Neutrophils and immunity: challenges and opportunities. *Nature reviews immunology*, 6(3):173–182, 2006.

Ooi, B. C., Tan, K., Wang, S., Wang, W., Cai, Q., Chen, G., Gao, J., Luo, Z., Tung, A. K. H., Wang, Y., Xie, Z., Zhang, M., and Zheng, K. SINGA: A distributed deep learning platform. In *ACM Multimedia*, pp. 685–688. ACM, 2015.

Roweis, S. T. and Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500): 2323–2326, 2000.

Rozemberczki, B., Watson, L., Bayer, P., Yang, H., Kiss, O., Nilsson, S., and Sarkar, R. The shapley value in machine learning. In *IJCAI*, pp. 5572–5579. ijcai.org, 2022.

Schubert, E., Sander, J., Ester, M., Kriegel, H., and Xu, X. DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Trans. Database Syst.*, 42(3): 19:1–19:21, 2017.

Shapley, L. S. et al. A value for n-person games. 1953.

Spinella, R., Sawhney, R., and Jalan, R. Albumin in chronic liver disease: structure, functions and therapeutic implications. *Hepatology international*, 10:124–132, 2016.

Szklo, M. Population-based cohort studies. *Epidemiologic reviews*, 20(1):81–90, 1998.

Tenenbaum, J. B., Silva, V. d., and Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P. Extracting and composing robust features with denoising autoencoders. In *ICML*, volume 307 of *ACM International Conference Proceeding Series*, pp. 1096–1103. ACM, 2008.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11:3371–3408, 2010.

Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., Wang, B., Xiang, H., Cheng, Z., Xiong, Y., et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in wuhan, china. *jama*, 323(11):1061–1069, 2020.

Wang, X. and Kattan, M. W. Cohort studies: design, analysis, and reporting. *Chest*, 158(1):S72–S78, 2020.

Weisberg, S. and Cook, R. D. Residuals and influence in regression. 1982.

Williamson, B. D. and Feng, J. Efficient nonparametric statistical inference on population feature importance using shapley values. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10282–10291. PMLR, 2020.

Winaver, J., Agmon, D., Harari, R., and Better, O. S. Impaired renal acidification following acute renal ischemia in the dog. *Kidney international*, 30(6):906–913, 1986.

Wu, H. and Huang, J. Drug-induced nephrotoxicity: pathogenic mechanisms, biomarkers and prevention strategies. *Current drug metabolism*, 19(7):559–567, 2018.

Wu, J., Huang, J., Zhu, G., Wang, Q., Lv, Q., Huang, Y., Yu, Y., Si, X., Yi, H., Wang, C., et al. Elevation of blood glucose level predicts worse outcomes in hospitalized patients with covid-19: a retrospective cohort study. *BMJ Open Diabetes Research and Care*, 8(1):e001476, 2020.

Xu, Y., Biswal, S., Deshpande, S. R., Maher, K. O., and Sun, J. RAIM: recurrent attentive and intensive model of multimodal patient monitoring data. In *KDD*, pp. 2565–2573. ACM, 2018.

Yang, B., Fu, X., Sidiropoulos, N. D., and Hong, M. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3861–3870. PMLR, 2017.

Yang, H. S., Hur, M., Lee, K. R., Kim, H., Kim, H. Y., Kim, J. W., Chua, M. T., Kuan, W. S., Chua, H. R., Kitiyakara, C., et al. Biomarker rule-in or rule-out in patients with acute diseases for validation of acute kidney injury in the emergency department (brava): a multicenter study evaluating urinary timp-2/igfbp7. *Annals of Laboratory Medicine*, 42(2):178, 2022.

Zahorec, R. et al. Ratio of neutrophil to lymphocyte counts-rapid and simple parameter of systemic inflammation and stress in critically ill. *Bratislavske lekarske listy*, 102(1): 5–14, 2001.

Zhang, X., Li, S., Chen, Z., Yan, X., and Petzold, L. R. Improving medical predictions by irregular multimodal electronic health records modeling. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 41300–41313. PMLR, 2023.

Zheng, K., Gao, J., Ngiam, K. Y., Ooi, B. C., and Yip, J. W. L. Resolving the bias in electronic medical records. In *KDD*, pp. 2171–2180. ACM, 2017a.

Zheng, K., Wang, W., Gao, J., Ngiam, K. Y., Ooi, B. C., and Yip, J. W. L. Capturing feature-level irregularity in disease progression modeling. In *CIKM*, pp. 1579–1588. ACM, 2017b.

Zheng, K., Cai, S., Chua, H. R., Wang, W., Ngiam, K. Y., and Ooi, B. C. TRACER: A framework for facilitating accurate and interpretable analytics for high stakes applications. In *SIGMOD Conference*, pp. 1747–1763. ACM, 2020.

Zheng, K., Chen, G., Herschel, M., Ngiam, K. Y., Ooi, B. C., and Gao, J. PACE: learning effective task decomposition for human-in-the-loop healthcare delivery. In *SIGMOD Conference*, pp. 2156–2168. ACM, 2021.

Zheng, K., Cai, S., Chua, H. R., Herschel, M., Zhang, M., and Ooi, B. C. Dyhealth: Making neural networks dynamic for effective healthcare analytics. *Proc. VLDB Endow.*, 15(12):3445–3458, 2022.

Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., Gu, X., et al. Clinical course and risk factors for mortality of adult inpatients with covid-19 in wuhan, china: a retrospective cohort study. *The lancet*, 395(10229):1054–1062, 2020.

APPENDIX

A. Notation Table

In this paper, scalars are denoted by symbols such as x, vectors are represented by boldface symbols such as $\mathbf{x}$, and matrices are described by uppercase boldface symbols such as $\mathbf{X}$. To provide a comprehensive overview of the notations used throughout the paper, we present a summary of notations in Table 2.

Table 2: Notations.

Notation	Description
$\mathcal{D}$, d_i	EMR data, each sample in EMR data
$\mathcal{D}^+$, $\mathcal{D}^-$	Positive samples, negative samples
d_i^-	Each negative sample
$\mathbf{x}_i$, y_i	Input features of d_i^- , binary label of d_i^-
F	Prediction model
$\mathcal{Q}$	A subset of negative samples
M	Performance metric function
s_i	Data Shapley value for d_i^-
π	A Monte Carlo permutation
$A_\pi^{d_i^-}$	All the negative samples before d_i^- in π
$\mathcal{S}$	The Negative Sample Shapley Field
$\mathcal{S}'$	Transformed space after SDAE-based manifold learning
K	Number of DAEs in SDAE
k	Each DAE in SDAE, $k \in \{0, \ldots, K-1\}$
$\mathbf{h}_i^{(k)}$	Input to the encoder of the k-th DAE
$\tilde{\mathbf{h}}_i^{(k)}$	Corrupted version of $\mathbf{h}_i^{(k)}$ with masking noise
$f_\theta^{(k+1)}(\cdot)$	Encoder of the k-th DAE
$\hat{\mathbf{h}}_i^{(k+1)}$	Output from the encoder of the k-th DAE
$f_\phi^{(k+1)}(\cdot)$	Decoder of the k-th DAE
$\mathbf{z}_i^{(k)}$	Output from the decoder of the k-th DAE
$\mathcal{L}_{rec}^{(k)}$	Reconstruction loss in the k-th DAE
$\mathcal{L}_{iso}^{(k)}$	Isotropy constraint in the k-th DAE
$\mathcal{L}^{(k)}$	Overall loss in the k-th DAE
$\mathbf{h}_i^{(k+1)}$	Input to the encoder of the $(k+1)$-th DAE
$\mathbf{h}_i^{(K)}$	Input for medical cohort discovery

B. Extended Related Work

B.1. Cohort Studies

Healthcare analytics capitalizes on the rich insights derived from patients' EMR data to facilitate a wide spectrum of analytic tasks, ranging from diagnostic (Lipton et al., 2016) to prognostic applications (Mould, 2012). Current research in this field primarily focuses on addressing the intrinsic challenges associated with EMR data, such as irregularity and missing data (Lee et al., 2017; Zheng et al., 2017b; Che et al., 2018; Cao et al., 2018; Zhang et al., 2023), as well as bias (Zheng et al., 2017a). Additionally, significant efforts are directed towards enhancing the capabilities of EMR analytic models by improving interpretability (Bai et al., 2018; Zheng et al., 2020; Cai et al., 2021; Chua et al., 2021; Cai et al., 2022), augmenting reliability (Chen et al., 2021; Zheng et al., 2021), and integrating multimodal EMR processing (Xu et al., 2018; Zheng et al., 2022; Zhang et al., 2023). This field is experiencing rapid growth, fueled by the increasing availability of EMR data and continuous advancements in computational methodologies, notably deep learning (Ooi et al., 2015). Within the realm of healthcare analytics, cohort studies represent a compelling research avenue (Grimes & Schulz, 2002; Wang & Kattan, 2020; Cai et al., 2024). As a specific subtype of longitudinal studies, cohort studies focus on selecting a group of patients who share a common defining characteristic in order to investigate a particular outcome of interest. Cohort studies are well-suited for identifying potential risk factors and causes and monitoring the progression of diseases in patients' health conditions. For instance, in Mahmood et al. (2014), significant medical insights into the epidemiology of cardiovascular disease and its associated risk factors are provided through a cohort study. Another notable example pertains to the coronavirus disease 2019 (COVID-19), where multiple cohort studies have consistently demonstrated the critical role of D-dimer as a risk factor contributing to the mortality of COVID-19 patients (Huang et al., 2020a; Wang et al., 2020; Zhou et al., 2020). Leveraging cohort studies, researchers acquire the capacity to meticulously scrutinize various medical conditions, yielding invaluable medical insights. This, in turn, has the potential to drive substantial advancements in patient management and healthcare delivery (Szklo, 1998; Wu et al., 2020).

B.2. Related Work on Each Component of Our Approach

The Shapley value, originally introduced in cooperative game theory (Shapley et al., 1953), offers a solution for the equitable distribution of a team's collective value among its individual members (Chalkiadakis et al., 2011). This value allocation mechanism embodies key principles such as fairness, symmetry, and efficiency (Chalkiadakis et al., 2011), rendering it widely applicable across various machine learning applications (Rozemberczki et al., 2022). Notable applications of the Shapley value in machine learning encompass data valuation, feature selection, explainable machine learning, etc (Lundberg & Lee, 2017; Ghorbani & Zou, 2019; Williamson & Feng, 2020; Liu et al., 2022; Rozemberczki et al., 2022). Among these applications, data valuation holds particular significance in quantifying the contributions of individual data samples toward predictive models. In this research line, the data Shapley value (Ghorbani & Zou, 2019) presents an equitable valuation framework for data value quantification. Subsequent research efforts primarily focus on enhancing the computational efficiency of the data Shapley value through the application of specific techniques (Jia et al., 2019; Ghorbani et al., 2020).

Representation learning, i.e., learning data representations that benefit downstream tasks, is a crucial research area contributing to the success of many machine learning algorithms (Bengio et al., 2013). Among the representation learning methods, manifold learning, which operates under the assumption that the probability mass of the original data tends to concentrate in lower-dimensional regions compared to the original space, stands out due to its capability of reducing the dimensionality and visualizing the underlying structure of the data. Traditional manifold learning methods include Isomap (Tenenbaum et al., 2000), locally linear embedding (Roweis & Saul, 2000), and multi-dimensional scaling (Borg & Groenen, 2005). In recent years, AEs have garnered substantial interest in representation learning. AEs excel in capturing underlying data structures by reconstructing input data, thereby providing efficient and effective representations of unlabeled data. Researchers develop various AE variants for specific application scenarios, e.g., regularized AEs (Alain & Bengio, 2014), sparse AEs (Makhzani & Frey, 2014), DAEs (Vincent et al., 2008). For example, regularized AEs (Alain & Bengio, 2014) are proposed to prevent AEs from learning trivial identity mappings and to enhance their ability to capture comprehensive information from data. More specifically, sparse AEs (Makhzani & Frey, 2014), inspired by the sparse coding hypothesis in neuroscience, aim to learn sparse representations, and DAEs (Vincent et al., 2008) are introduced to learn representations robust to noise and outliers, hence effectively handling input data corruption. Specifically, DAEs and their advanced stacked variant SDAEs (Vincent et al., 2010) are highly suitable to tackle EMR data, in which missing and noisy data remains a notorious issue (Lasko et al., 2013). These models could effectively address the complexities associated with EMR data and contribute to improved representation learning.

DBSCAN, short for density-based spatial clustering of applications with noise, is introduced to alleviate the burden of parameter selection for users, facilitate the discovery of arbitrarily shaped clusters, and demonstrate satisfactory efficiency when dealing with large datasets (Ester et al., 1996; Gan & Tao, 2015; Schubert et al., 2017). A subsequent study, ρ-approximate DBSCAN further advances the quality of cluster approximation and computational efficiency (Gan & Tao, 2015). Then in Schubert et al. (2017), it is shown that the original DBSCAN algorithm, when equipped with appropriate indexes and parameters, can achieve performance comparable to that of the ρ-approximate DBSCAN algorithm. Up till now, DBSCAN remains one of the most widely adopted clustering algorithms.

B.3. Related Work on Baseline Methods

Contrastive principal component analysis (cPCA) is a generalized variant of the standard PCA. Its primary purpose is to visualize and investigate patterns specific to a target dataset in contrast to an existing background dataset. In this manner, cPCA excels at identifying crucial dataset-specific patterns that might be missed by PCA (Abid et al., 2017; 2018).

In the setting of learning from positive and unlabeled data, generally referred to as PU learning, we only have access to positive examples and unlabeled data for analytics. PU learning has garnered increasing interest within the machine learning community, and among the notable research endeavors, three influential PU learning methods have emerged: Classic Elkanoto (Elkan & Noto, 2008), Weighted Elkanoto (Elkan & Noto, 2008), and Bagging-based PU-learning (Mordelet & Vert, 2014). The first two are founded on the assumption of samples being “selected completely at random,” while the latter, Bagging-based PU-learning, leverages bootstrap aggregating (bagging) techniques to achieve improved performance.

Influence functions, as introduced in Weisberg & Cook (1982), present an alternative approach for assessing the value of individual samples. Influence functions quantitatively measure how the prediction model changes when the weight of a single sample is perturbed slightly. To compute the influence function value per sample, it is common practice to employ the leave-one-out (LOO) method, a well-established technique in the field (Ghorbani & Zou, 2019; Rozemberczki et al., 2022).

Recently, there has been a growing interest in deep clustering methods that simultaneously optimize representation learning and clustering. One noteworthy example is the deep clustering network (DCN) (Yang et al., 2017). DCN combines dimensionality reduction and K-means clustering, where the dimensionality reduction component is accomplished via learning a deep neural network. Another noteworthy approach is deep embedded K-means clustering (DEKM) (Guo et al., 2021). DEKM alternately employs an autoencoder to learn a deep embedding space, and identifies clusters within this space, thereby revealing valuable cluster-structure information.

C. Negative Sample Shapley Field Construction

C.1. Proof of Data Shapley Values for Negative Samples

We establish the proof of data Shapley values for negative samples by relating our problem to the original context of the Shapley value in game theory (Shapley et al., 1953), i.e., reducing it to a cooperative game (Chalkiadakis et al., 2011; Ghorbani & Zou, 2019; Rozemberczki et al., 2022).

Specifically, our problem is framed as a negative sample valuation game for a fair distribution of the collective performance achieved by the prediction model to each participating negative sample in the training data (with positive samples the same), while maintaining consistency with the three fundamental properties of an equitable data valuation: (i) null player, (ii) symmetry, and (iii) linearity.

Null player. We define a negative sample d_i^- as a "null player" and set its data Shapley value to zero if its inclusion in any subsets of the negative sample set in training data does not influence the performance of the prediction model. Formally, for a negative sample $d_i^- = (\mathbf{x}_i, y_i)$ and $\forall \mathcal{R} \subseteq \mathcal{D}^- \setminus d_i^-$, if the performance remains unchanged by adding d_i^-, i.e., $M(\mathcal{D}^+ \cup \mathcal{R}, F) = M(\mathcal{D}^+ \cup \mathcal{R} \cup \{d_i^-\}, F)$, then $s_i = 0$. In this negative sample valuation game, the null player property ensures that the negative samples with no impact on the prediction performance are assigned zero values for their data Shapley values.

Symmetry. Two negative samples, d_i^-, and d_j^-, are assigned the same value if they consistently influence the performance of the prediction model when added to any subsets of the negative sample set in training data. This property arises from the concept of symmetry. Formally, for two negative samples $d_i^- = (\mathbf{x}_i, y_i)$ and $d_j^- = (\mathbf{x}_j, y_j)$, and $\forall \mathcal{R} \subseteq \mathcal{D}^- \setminus \{d_i^-, d_j^-\}$, if the prediction performance remains the same after adding d_i^- or d_j^-, i.e., $M(\mathcal{D}^+ \cup \mathcal{R} \cup \{d_i^-\}, F) = M(\mathcal{D}^+ \cup \mathcal{R} \cup \{d_j^-\}, F)$, then $s_i = s_j$. This property ensures that the negative samples with equivalent marginal contributions are assigned the same data Shapley values.

Linearity. The influence of a negative sample d_i^- on the overall pooled data is equivalent to its influence on constituent sub-datasets. We could denote s_i as $s_i(d_i^-, \mathcal{D}_{test})$, representing the data Shapley value of the negative sample d_i^- evaluated on all test data $\mathcal{D}_{test}$. The linearity property states that for two sets of test data, $\mathcal{D}_{test}^1$ and $\mathcal{D}_{test}^2$, the following holds:

$$s_i(d_i^-, \mathcal{D}_{test}^1 \cup \mathcal{D}_{test}^2) = s_i(d_i^-, \mathcal{D}_{test}^1) + s_i(d_i^-, \mathcal{D}_{test}^2) \tag{10}$$

This linearity property ensures that the data Shapley value of a negative sample on the pooled test dataset is equal to the sum of its data Shapley values on the two individual test datasets, in this negative sample valuation game.

Proof. We prove Proposition 3.1 by establishing the connection between our negative sample valuation game and the cooperative game theory context (Chalkiadakis et al., 2011). In a cooperative game, there exists a set of n players and a characteristic function $m : 2^{[n]} \mapsto \mathbb{R}$ that assigns a payment value to each selected player (Rozemberczki et al., 2022). In our case, the players correspond to individual negative samples, and the characteristic function $m(\mathcal{Q})$ represents the performance obtained when the subset of negative samples $\mathcal{Q}$ ($\mathcal{Q} \subseteq \mathcal{D}^-$) is included in the prediction model. By leveraging the three properties discussed above, our negative sample valuation game ensures the fair distribution of collective performance to the participating negative samples. Therefore, each negative sample acts as a player, and the prediction model F incorporates all the participating negative samples $\mathcal{Q}$ (along with the positive samples $\mathcal{D}^+$) to achieve the overall performance $m = M(\mathcal{D}^+ \cup \mathcal{Q}, F)$. Consequently, the data Shapley value of each negative sample corresponds to the payment received by each player in this cooperative game analogy. □

C.2. Monte Carlo Permutation Sampling

We adopt Monte Carlo permutation sampling to approximate the data Shapley values for negative samples. The detailed procedure of each Monte Carlo iteration is presented in Algorithm 1. The algorithm begins by initializing the necessary variables for computation in lines 1-8. Subsequently, for a given permutation, we calculate the marginal contribution of each negative sample in the current Monte Carlo iteration towards its overall data Shapley value, as described in lines 9-24.

In particular, for each indexed negative sample, we include it in the training data and retrain the classifier. Then, we measure its marginal contribution by calculating the difference in the AUC metric (lines 10-14). Additionally, in line 15, we compute the absolute difference between the full AUC (which uses all the training data and evaluates the trained model on the test data) and the new AUC (which includes the current negative sample). If this difference falls below a predefined threshold,

Algorithm 1 Data Shapley Value Computation for Negative Samples by Monte Carlo Sampling

Input: Negative training data $(\mathbf{X}^-_{train}, \mathbf{y}^-_{train})$, Positive training data $(\mathbf{X}^+_{train}, \mathbf{y}^+_{train})$, Test data $(\mathbf{X}_{test}, \mathbf{y}_{test})$.
Output: Marginal contribution of each negative sample in current Monte Carlo iteration to its overall data Shapley value.

```
 1: Initialize permutation of indices of $\mathbf{X}^-_{train}$: *perm* $\leftarrow$ random permutation
 2: Initialize marginal contributions of $\mathbf{X}^-_{train}$ with zeros: *marginal_contribs* $\leftarrow$ zeros
 3: Initialize truncation counter: *truncation_counter* $\leftarrow 0$
 4: Initialize new score with a random score: *new_score* $\leftarrow$ random_score        ▷ 0.5 for AUC
 5: Initialize a classifier: *clf* $\leftarrow$ create a new classifier
 6: Fit the classifier with all training data: *clf*.fit($\mathbf{X}^-_{train} \cup \mathbf{X}^+_{train}, \mathbf{y}^-_{train} \cup \mathbf{y}^+_{train}$)
 7: Evaluate the classifier on test data: *full_score* $\leftarrow$ AUC(*clf*, $\mathbf{X}_{test}$, $\mathbf{y}_{test}$)
 8: Initialize training data: $(\mathbf{X}', \mathbf{y}') \leftarrow (\mathbf{X}^+_{train}, \mathbf{y}^+_{train})$
 9: **for** idx **in** perm **do**
10:     Set old score to the current new score: *old_score* $\leftarrow$ *new_score*
11:     Update training data with current negative sample: $(\mathbf{X}', \mathbf{y}') \leftarrow (\mathbf{X}' \cup \mathbf{X}^-_{train}[idx], \mathbf{y}' \cup \mathbf{y}^-_{train}[idx])$
12:     Create a new classifier and train it: *clf* $\leftarrow$ new classifier, *clf*.fit($\mathbf{X}', \mathbf{y}'$)
13:     Update new score: *new_score* $\leftarrow$ AUC(*clf*, $\mathbf{X}_{test}$, $\mathbf{y}_{test}$)
14:     Calculate marginal contribution of the current negative sample: *marginal_contribs*$[idx] \leftarrow$ *new_score* $-$ *old_score*
15:     Calculate the distance to the full score: *distance_to_full_score* $\leftarrow$ |full_score - new_score|
16:     **if** distance_to_full_score $\leq$ truncation_tolerance $\times$ full_score **then**
17:         Increment truncation counter: *truncation_counter* $\leftarrow$ *truncation_counter* $+ 1$
18:         **if** truncation_counter $> 5$ **then**
19:             **break**
20:         **end if**
21:     **else**
22:         Reset truncation counter: *truncation_counter* $\leftarrow 0$
23:     **end if**
24: **end for**
25: **return** *marginal_contribs*
```

specifically "$truncation\_tolerance$" times the full AUC, for more than five consecutive negative samples, we terminate the current Monte Carlo iteration by early stopping (lines 16-20). This early stopping criterion is based on the observation that further inclusion of negative samples is unlikely to yield a significant improvement in AUC.

After calculating the marginal contribution of each negative sample in each Monte Carlo iteration, the overall data Shapley value of a particular negative sample is derived by taking the mean of its marginal contributions across different iterations.

C.3. Implementation Details

In our experiments, we employ the LR model to approximate the data Shapley values for negative samples. We use AUC as the evaluation metric with the aforementioned early stopping criterion. Specifically, we set the threshold $truncation\_tolerance$ to 0.025. This means that if the absolute difference between the full AUC and the new AUC remains within 0.025 times the full AUC for more than five consecutive negative samples, the current Monte Carlo iteration will be terminated.

D. Manifold Learning with Structure Preservation and Isotropy Constraint

D.1. Model Architecture of SDAE-based Manifold Learning

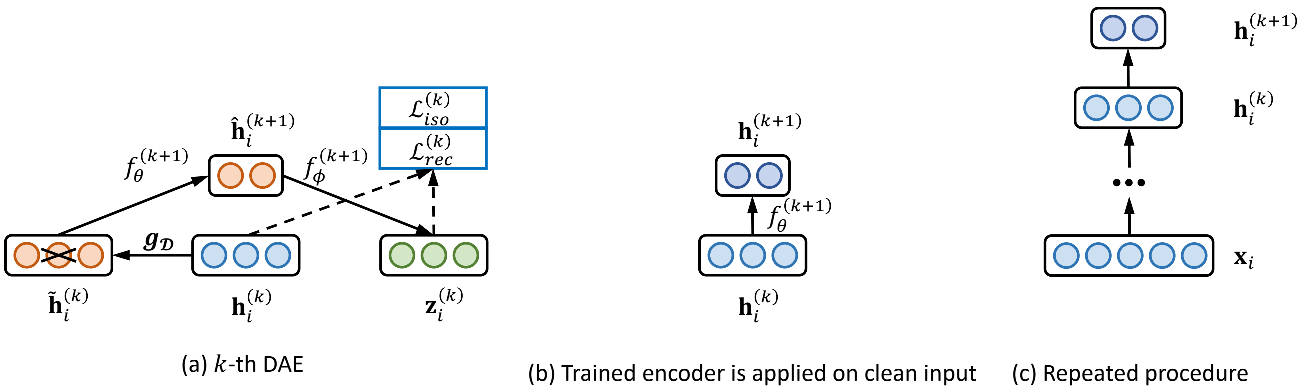


Figure 7: Model architecture of SDAE-based manifold learning.

The architecture of our SDAE-based manifold learning model is presented in Figure 7. Specifically, Figure 7 (a) illustrates the computation details of the k-th DAE (denoising autoencoder). Both the reconstruction loss and our proposed isotropy constraint are employed in this process, and hence contribute to the training of robust feature extractors, utilizing clean and corrupted inputs simultaneously. Subsequently, after training the encoder of the k-th DAE, it is applied to clean inputs for transformation, as depicted in Figure 7 (b). This iterative training process continues, as shown in Figure 7 (c), until the last DAE generates the final encoded representation within the manifold space. This encoded representation serves as the foundation for subsequent cohort discovery.

D.2. Implementation Details

We utilize an SDAE comprising 3 DAEs. These DAEs serve to transform the 709-dimension input data, which corresponds to 709 lab tests (refer to Section 4.1), using encoders with dimensions of 256, 128, and 64, respectively. We utilize the Adam optimizer to train the SDAE in an unsupervised manner, using the loss function described in Equation 7. Our objective is to obtain an optimal manifold space to support subsequent density-based clustering for automatic cohort identification. To determine the optimal learning rate for training, we perform a grid search over a range of values, i.e., $[0.1, 0.05, 0.01, 0.005, 0.001, 0.0005]$, and run 10 repeats per learning rate. The model run with the lowest loss is selected as the optimal model for subsequent cohort discovery, which corresponds to a learning rate of 0.005. Other parameters are held constant during the training process, including a batch size of 1024, a mask probability of 0.2 for the denoising process, and a total of 100 epochs. These parameters provide stability and ensure sufficient training iterations to learn meaningful representations in the SDAE.

E. Cohort Discovery Among High Data Shapley Value Negative Samples

E.1. Details for DBSCAN

The DBSCAN (density-based spatial clustering of applications with noise) algorithm follows a specific process to perform clustering. It requires two essential parameters: (i) ε, which specifies the maximum distance between two samples for them to qualify as neighbors, and (ii) P_{min}, which defines the minimum number of samples required to form a dense region.

The detailed description of the DBSCAN algorithm is as follows. (i) Start by selecting an unvisited sample arbitrarily. (ii) Retrieve its ε-neighborhood, consisting of all samples within a distance of ε from the selected sample. (iii) If the ε-neighborhood contains more than P_{min} samples, initiate a new cluster and designate the selected sample as a "core point." The core point is a sample that has a sufficient number of neighbors within its ε-neighborhood to form a dense region. (iv) If the ε-neighborhood has fewer than P_{min} samples, label the selected sample as noise. However, note that this sample may later fall within the ε-neighborhood of another sample, causing it to be assigned to a different cluster. (v) For each sample that is determined to belong to a dense region within a cluster, consider its ε-neighborhood as part of the same cluster. Add all the samples found within this neighborhood to the cluster and check if these samples' respective ε-neighborhoods are also dense (if so, add them to the cluster as well). This process continues recursively until the entire densely connected cluster is detected. (vi) Proceed to the next unvisited sample and repeat steps (ii) to (v) until all samples have been assigned to a cluster or labeled as noise. By following this process, DBSCAN identifies densely connected clusters $\{C_1, C_2, \ldots, C_R\}$ and recognizes noisy samples Ψ.

E.2. Clusters vs. Cohorts

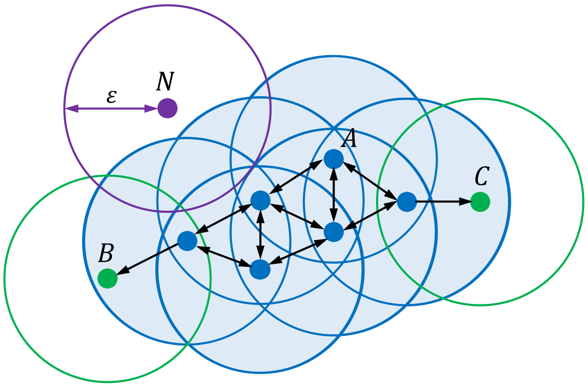


Figure 8: Relationship between clusters and cohorts in a DBSCAN example.

We illustrate the relationship between clusters and cohorts using an example within the DBSCAN algorithm, as depicted in Figure 8. In this example, we set P_{min} to 4, and the value of ε is indicated in the figure as the radius of the circles.

As shown in the figure, Point A and all the other blue points are core points because their ε-neighborhoods contain at least P_{min} points. Therefore, they form a single cluster. Additionally, Point B and Point C are reachable from Point A via existing paths, making them belong to the same cluster as well. However, Point N is labeled as noise since it does not meet the criteria to be a core point and is not reachable from any core points.

According to Definition 3.3, for this identified cluster by the DBSCAN algorithm, we consider each core point (i.e., all the blue points) and define a spherical space with the core point as its center and ε as its radius. The combined area covered by all such spherical spaces, depicted in blue, represents the cohort that we aim to discover from this cluster.

E.3. Implementation Details

The choice of parameters in the DBSCAN algorithm, specifically the search radius (ε) and the minimum number of points (P_{min}), has a significant impact on the quality of the clustering results. To determine the optimal parameter combination, we start by exploring various values for P_{min}.

Given a specific P_{min} value, we calculate the 75th percentile of the distribution of $(P_{min}/2)$-nearest distances for the extracted 40% samples with high data Shapley values (as described in Section 4.2). We consider this calculated value as the appropriate ε for the clustering process. The underlying rationale is that regions with local densities exceeding twice the upper bound of the global density represent distinct high-density areas.

By iterating over different values of P_{min} and adjusting the corresponding ε values, we assess the clustering quality achieved by each parameter combination using the Silhouette score, which measures the cohesion and separation of clusters to evaluate their quality. After a thorough evaluation, we determine that a value of P_{min} equal to 100 yields the most suitable parameter choice for our DBSCAN clustering. This method ensures that the clustering process considers the distribution of distances within high-density areas and selects an appropriate value for ε, leading to improved clustering results based on the Silhouette score assessment.

F. In-depth Analysis of Our Proposed Approach

F.1. Design Choices of Components

K-means clustering results. We first present the K-means clustering results that comprise two subsets: the first involves K-means clustering analysis of the constructed Negative Sample Shapley Field **without** our proposed isotropy constraint, and the second **applies** the isotropy constraint to K-means clustering. Each subset includes two figures: one displaying the t-SNE plots of the K-means clustering results, incorporating a heatmap of data Shapley values along with multiple K-means outcomes for various K values[1]; the other, based on a trained K-means model with $K = 7$, presents a histogram of data Shapley values for samples within the corresponding identified cohorts. The choice of $K = 7$ as our primary focus stems from its emergence as the cohort number via DBSCAN analysis as shown in Figure 3, alongside a detailed medical examination based on these clustering results, affirming their medical relevance.

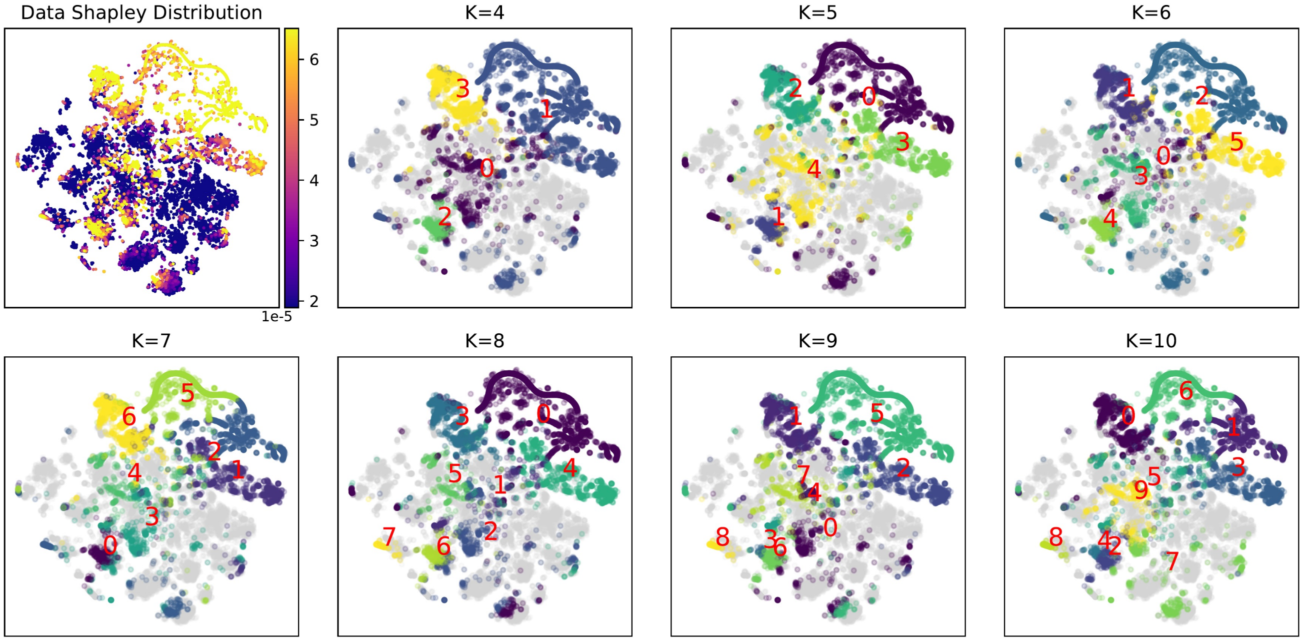


Figure 9: Cohort discovery results of our proposal using K-means without isotropy constraint.

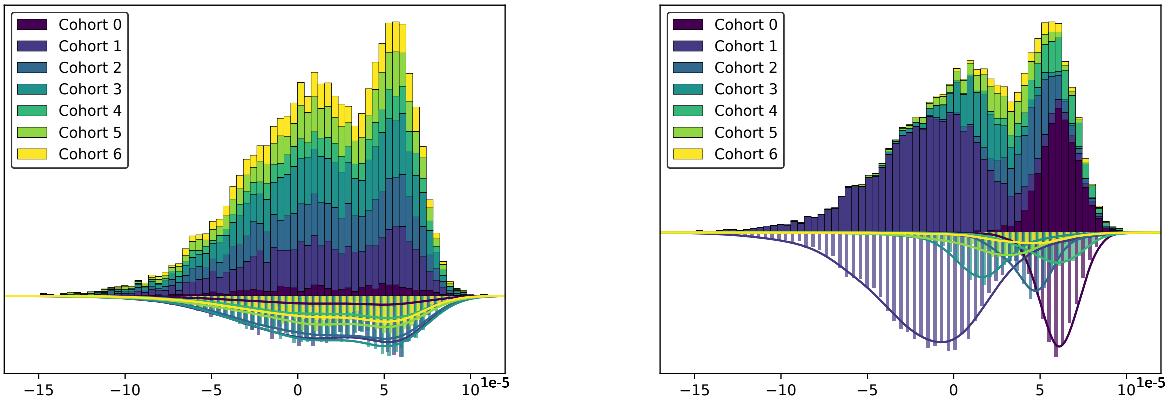
Let us first focus on the K-means outcomes without the isotropy constraint. As observed in Figure 9, the degree of differentiation between the clusters formed is not particularly ideal. Further, Figure 10a starkly demonstrates that spatial clustering of high-temperature points[2] within the Negative Sample Shapley Field, aimed at automatically identifying hot zones, is completely ineffective in the absence of the isotropy constraint. This is explicitly manifested in the undifferentiated distribution of data Shapley values, which appears to be unrelated to the data Shapley values themselves, resembling a nearly random grouping. This scenario precisely encapsulates the situation we are facing at the beginning of our work.

Following further investigation, we have identified the cause of the difficulty as the irregular shapes of the hot zones in the unconstrained high-dimensional Negative Sample Shapley Field, as illustrated in Figure 1. Traditional clustering methods based on spatial distance struggle to function effectively in this context. To address this issue, we propose an isotropy constraint as a solution.

Figures 11 and 10b present the second subset of results, specifically the K-means clustering results under the isotropy constraint. As evident from Figure 11, the quality of the clustering results has significantly improved. Moreover, Figure 10b further supports this assessment, demonstrating that the obtained clusters exhibit a clear similarity in data Shapley values within each cluster, approximating a normal distribution. This indicates that K-means clustering holds promise when the isotropy constraint is in place.

[1] Disambiguation: "K" herein denotes the number of clusters to partition in K-means clustering.

[2] We use "temperature" to refer to data Shapley values, where high-temperature points correspond to data samples with high data Shapley values.



(a) K-means without isotropy constraint.

(b) K-means with isotropy constraint.

Figure 10: Data Shapley value histograms of samples within discovered cohorts using K-means clustering.

We emphasize that the comparison between clustering methods is orthogonal to the core concerns of this paper, and the related discussion has already been sufficiently addressed in the existing literature. The prevailing view suggests that DBSCAN, by adopting a method of expanding and merging high-density regions, offers a more flexible clustering result. This approach is deemed potentially superior in handling high-dimensional scenarios where the boundaries of clusters are irregular, compared to direct spatial partitioning based on kernels as performed by K-means.

Such assertions are observable in our results, for instance, in the subplot at the lower left corner of Figure 11 ($K = 7$), where a portion of Cohort 4 (green) is clustered into Cohort 2 (blue). This clustering evidently stems from the rigid spatial segmentation inherent to kernel-based partitioning, as intuitively, these blue points should belong to a dense cluster of green points. Contrastingly, our proposed approach utilizing DBSCAN does not cluster this small subset of blue points into Cohort 2 (i.e., the hepatic and hematological disorders cohort, that is, Cohort 4 in Figure 3(c)), attributing to a more refined analysis of the cases associated with these points. We believe that the DBSCAN-based clustering is more accurate, indicating that these blue points indeed do not share pathological characteristics with the cohort. Thus, in this specific case, DBSCAN demonstrates superior performance.

Moreover, K-means assigns every point to a cluster, whereas DBSCAN, by expanding and merging based on high-density regions, allows discrete, low-density points to remain un-clustered. This approach is more suitable for our specific application, where attention can be deferred from isolated points to concentrate on high-density areas, often indicating significant medical insights. Furthermore, unlike K-means, DBSCAN does not require the a priori specification of a cluster number K, a parameter challenging to determine in complex, uncertain analysis scenarios. These considerations lead us to conclude that DBSCAN is a preferable choice for our application.

Choice of clustering methods. Our experience during the development of our cohort discovery approach informs our choice of clustering methods. Initially, we employ K-means but find it inadequate for cohort discovery due to its assumption of evenly distributed data samples around centroids, which may not hold in high-dimensional medical data contexts (as discussed above and also noted by Yang et al. (2017)).

To address these limitations, we further explore deep clustering methods such as DCN (Yang et al., 2017) and DEKM (Guo et al., 2021), which jointly optimize representation learning and clustering, with the underlying assumption that the latent representations derived from deep neural networks will be inherently well-suited for clustering. Nonetheless, both methods fail to identify medically meaningful cohorts, resulting in only one large cohort and a few smaller ones (with the corresponding results presented in Appendix G.5). Further investigation reveals that this unsatisfactory outcome is due to the anisotropic nature of the constructed Negative Sample Shapley Field, indicating the non-uniform distribution of negative samples with similar data Shapley values.

To mitigate this issue, we propose the isotropy constraint to ensure uniform data Shapley value changes across orientations, rendering our approach more amenable to subsequent spatial clustering. We further adopt DBSCAN as a mature and proven spatial clustering method with broad applicability across various scenarios. DBSCAN is capable of identifying high-density connected subspaces of arbitrary shape without requiring a pre-assumed number of clusters K (Ester et al., 1996; Gan & Tao, 2015; Schubert et al., 2017). Furthermore, our approach seeks to identify high-temperature connected subspace in the

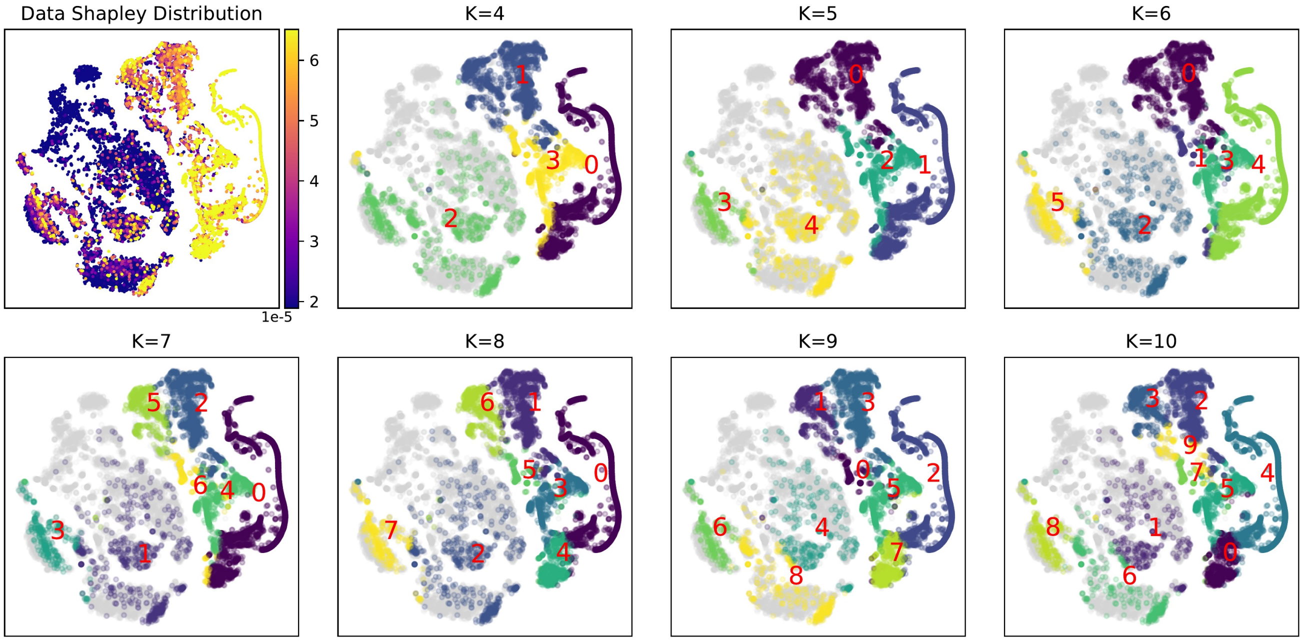


Figure 11: Cohort discovery results of our proposal using K-means with isotropy constraint.

constructed Negative Sample Shapley Field, aligning well with DBSCAN's mechanism. Equipped with both our proposed isotropy constraint and DBSCAN, we successfully avoid the mis-discovery, as exemplified in Figure 1(b), and consequently, contribute to uncovering medically meaningful cohorts.

It is essential to clarify that the primary contribution of our proposal lies in exploring negative samples for cohort discovery through a Shapley-based exploration of interrelationships between these samples. Consequently, after constructing the Negative Sample Shapley Field, the choice of the subsequent spatial clustering method is somewhat orthogonal to our primary contribution. Therefore, we remain open to the possibility of replacing DBSCAN with more innovative and better-performing spatial clustering alternatives, should they become available, to further enhance the cohort discovery results for the benefit of clinicians.

Choice of data valuation methods. The central concept of our proposed approach revolves around incorporating data valuation into spatial clustering analysis among negative samples, through a Shapley-based exploration of interrelationships between the samples. We opt to utilize data Shapley values for this purpose, given that data Shapley values are recognized as a prominent and well-established equitable data valuation technique in recent years, offering distinct advantages over alternative methods examined in related work (Ghorbani & Zou, 2019; Rozemberczki et al., 2022).

In our investigation, we have also compared data Shapley values with other data valuation methods, such as the influence function discussed in Section 4.4. However, we find that the influence function-based data valuation deletes informative data samples from all negative samples rather than retaining them, leading to degraded performance. Additionally, influence functions have also been identified to have robustness issues in prior research (Ghorbani et al., 2019). Further experimental results and corresponding analyses are detailed in Appendix G.4.

F.2. Limitations and Failure Modes

As outlined in Section 2, our approach comprises two primary phases: (i) leveraging data Shapley values for data validation and constructing the Negative Sample Shapley Field for cohort discovery, and (ii) conducting cohort discovery via manifold learning and density-based clustering. From this perspective, two potential failure modes emerge: unsuitable data valuation methods, and improper representation learning or clustering models. Besides, another possible failure mode of our approach lies in the limited capability of predictive models. We next elaborate on these three failure modes in detail.

Unsuitable data valuation methods. Achieving the goal of exploring interrelationships among negative samples via data valuation necessitates accurate and suitable methods. If an unsuitable method, such as the influence function discussed in

Section 4.4 (with corresponding results and analyses detailed in Appendix G.4), is adopted—resulting in the deletion rather than retention of informative data samples and exhibiting robustness issues—it could compromise the integrity of the data valuation results, potentially leading to degraded performance and reliability of our approach.

Improper representation learning or clustering models. Subsequent to the construction of the Negative Sample Shapley Field, effective representation learning and clustering are pivotal for cohort discovery. However, employing improper models for representation learning or clustering can significantly impede this process. For instance, utilizing models such as cPCA or our proposed model without isotropy constraint in Section 4.4 (details in Appendices G.2 and G.1, respectively), or opting for deep clustering models like DCN and DEKM in Section 4.4 (details in Appendix G.5), may result in the detection of only one large cohort and a few smaller cohorts, thereby failing to identify medically meaningful cohorts. These models follow a similar process of dimensionality reduction followed by clustering on embedded representations. Without the integration of our proposed isotropy constraint, which ensures uniform changes in data Shapley values across orientations, their efficacy is limited. Consequently, such improper representation learning or clustering models would adversely affect the quality of cohort discovery results in our approach.

Limited capability of predictive models. We present an innovative approach centered on data Shapley values to explore the interrelationships among negative samples. We posit that valuable cohorts should exhibit similar distributions characterized by high data Shapley values. Ideally, highly accurate predictive models could facilitate the computation of data Shapley values and subsequent cohort discovery within the constructed Negative Sample Shapley Field. However, in practical scenarios, particularly in challenging prediction tasks, learning such accurate models may prove difficult, impeding the computation of data Shapley values and the identification of cohorts.

F.3. Complexity Analysis

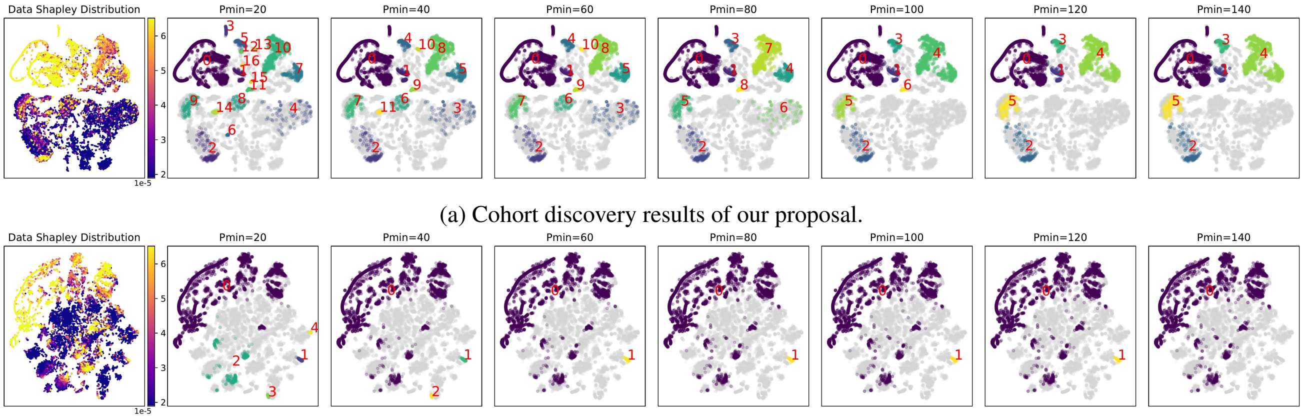
We analyze the complexity of our proposed cohort discovery approach in a step-wise manner as outlined below. Here, f represents the feature dimension, N is the total number of samples, and N^- is the number of negative samples. Since N and N^- are on the same order, we use N consistently in the following complexity analysis for simplicity.

Step 1. Negative Sample Shapley Field Construction. We employ Monte Carlo permutation sampling to calculate the data Shapley values for negative samples. Each Monte Carlo permutation involves using the LR model to compute the data Shapley value for each selected negative sample, with AUC serving as the evaluation metric. The computational complexity of LR is $O(Nf)$, while that of AUC calculation is $O(N \log N)$. As suggested by Ghorbani & Zou (2019), the convergence of Monte Carlo permutation sampling is generally reached with a sampling number on the order of N, and in our experiments, we run over $5N$ permutations. Further, considering the retraining of the LR model per negative sample, the complexity of Step 1 is $O(N^3(f + \log N))$.

Step 2. Manifold Learning with Structure Preservation and Isotropy Constraint. Our SDAE consists of K DAEs, where the input dimension of the k-th DAE's encoder is n_k and the output dimension is m_k. The overall complexity of SDAE is $O(N \sum_{k=0}^{K-1}(n_k m_k))$, which could be simplified to $O(N)$, considering that K, n_k and m_k are constants. Next, for the isotropy constraint, we calculate the distance between each pair of samples within each batch, resulting in $\binom{|\mathcal{B}|}{2} = |\mathcal{B}|(|\mathcal{B}| - 1)/2$ distance calculations. Since the complexity of computing the distance between each pair of samples is $O(f)$, the complexity of imposing the isotropy constraint for each batch is $O((|\mathcal{B}|(|\mathcal{B}| - 1)/2)f) = O(|\mathcal{B}|^2 f)$. With a total of $N/|\mathcal{B}|$ batches, the overall complexity of imposing the isotropy constraint is $O(N|\mathcal{B}|f)$. Combining the computation in both the SDAE and the isotropy constraint, the complexity of Step 2 is $O(N|\mathcal{B}|f)$.

Step 3. Cohort Discovery Among High Data Shapley Value Negative Samples. In this step, we mainly conduct DBSCAN on all the negative samples. The average complexity of this process is $O(N \log N)$ by employing an accelerating indexing structure, while the worst-case complexity is $O(N^2)$.

In summary, our proposed cohort discovery approach, encompassing all three aforementioned steps, demonstrates an overall complexity of $O(N^3 \log N)$, with the major computational overhead occurring in Step 1. During this step, we calculate the data Shapley values for negative samples using Monte Carlo permutation sampling in $O(N^2)$ time and subsequently compute the AUC metric of LR for each permutation per negative sample in $O(N \log N)$ time.



(a) Cohort discovery results of our proposal.

(b) Cohort discovery results of our proposal without isotropy constraint.

Figure 12: Comparison between our proposal with vs. without isotropy constraint in cohort discovery.

G. Supplementary Experimental Results

G.1. Experimental Results on Effectiveness of Isotropy Constraint

The comprehensive results of the ablation study, comparing the cohort discovery results of our proposal with and without the isotropy constraint, are presented in Figure 12, encompassing various P_{min} settings. According to the comparison results, we observe that the results of our proposal (with the isotropy constraint) remain relatively stable across varied P_{min} settings ($P_{min} = 100$ for Figure 3(c) as described in Appendix E.3). Notably, upon removing the isotropy constraint, effective cohort discovery through DBSCAN is impeded, irrespective of the chosen P_{min} values.

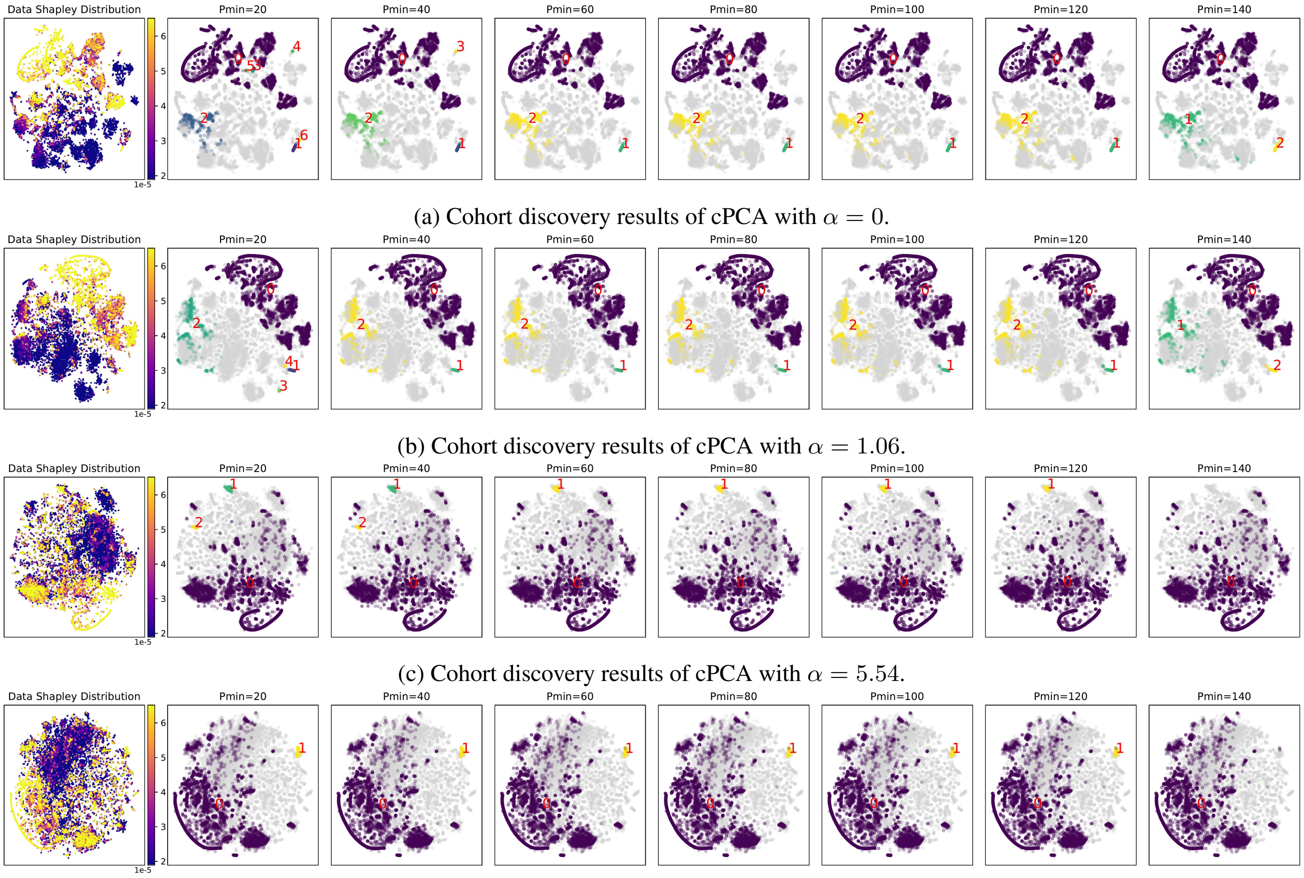
This is because DBSCAN, as a spatial clustering method, necessitates appropriate handling of spatial information within the Negative Sample Shapley Field to achieve meaningful cohort discovery. Therefore, the introduced isotropy constraint plays a pivotal role by ensuring uniform changes in data Shapley values across orientations, thereby rendering our proposal more amenable to subsequent spatial clustering. Consequently, we mitigate the risk of mis-discovery, as exemplified in Figure 1(b), and ensure robust cohort discovery outcomes resilient to variations in spatial clustering algorithm parameters, ultimately contributing to unveiling medically meaningful cohorts.

G.2. Comparison with cPCA

cPCA, short for contrastive principal component analysis, represents a generalization of the standard PCA. By utilizing a background dataset to eliminate common patterns, cPCA's objective is to unveil the unique patterns within the target dataset relative to the background dataset (Abid et al., 2017; 2018). In this regard, cPCA shares a conceptual similarity with our proposal. Therefore, we proceed to conduct a comparative analysis between our approach and cPCA for cohort discovery.

Considering our focus on identifying cohorts within negative samples, we construct the background dataset using positive samples, while the negative samples constitute the target dataset. Following the projection of the data through cPCA, we retain 64 contrastive principal components, a dimension that is consistent with the output of SDAE in our approach. Subsequently, we employ DBSCAN on the projected data resulting from cPCA to achieve cohort discovery. In essence, cPCA can be regarded as an embedding technique, serving as a counterpart to the combination of the first two components in our proposed approach: Negative Sample Shapley Field Construction, and Manifold Learning with Structure Preservation and Isotropy Constraint.

Comprehensive experimental results of cPCA are depicted in Figure 13, presenting cohort discovery outcomes for four distinct α values: 0, 1.06, 5.54, and 74.44 (automatically determined by cPCA), across different P_{min} settings. Comparing these cPCA results with our proposal's results illustrated in Figure 12(a), it is observed that our proposal demonstrates a clear superiority over cPCA in cohort discovery across different α values as well as varying P_{min} settings.



(a) Cohort discovery results of cPCA with $\alpha = 0$.

(b) Cohort discovery results of cPCA with $\alpha = 1.06$.

(c) Cohort discovery results of cPCA with $\alpha = 5.54$.

(d) Cohort discovery results of cPCA with $\alpha = 74.44$.

Figure 13: Cohort discovery results of cPCA with four different α settings.

G.3. Comparison with Positive-Unlabelled (PU) Learning Methods

While our primary focus centers on effective cohort discovery among negative samples rather than rectifying the asymmetry between positive and negative examples for enhanced performance, our approach holds relevant implications for the latter. Therefore, we compare our approach against three PU learning methods: Classic Elkanoto (Elkan & Noto, 2008), Weighted Elkanoto (Elkan & Noto, 2008), and Bagging-based PU-learning (Mordelet & Vert, 2014). Specifically, we train these three baseline methods, treating negative samples as unlabeled data, and evaluate their performance on the testing data.

The experimental results of these baseline methods in terms of AUC (mean ± std) from five repeats are presented in Figure 14. Among the benchmarked baselines, Bagging-based PU-learning outperforms the other two methods and also surpasses the performance of the "all d_i^-" setting, where all negative/unlabeled samples are included in the training. This validates the effectiveness of Bagging-based PU-learning, achieved through its bootstrap aggregating techniques. On the other hand, both Classic Elkanoto and Weighted Elkanoto fail to achieve satisfactory performance. They merely marginally outperform the "all d_i^-" setting when employing LR and AdaBoost. This observation suggests that the "selected completely at random" assumption inherent in these two baselines may not hold in our hospital-acquired AKI prediction utilizing real-world EMR data.

In contrast to these baselines, the "d_i^- with $s_i > 0$" setting of our proposal, which filters out the negative samples with negative data Shapley values, consistently achieves substantially higher AUC values across different classifiers. This firmly establishes the superiority of our approach in identifying negative samples in real-world medical data, which further underscores the validity of our constructed Negative Sample Shapley Field, thus providing a robust foundation for subsequent cohort discovery.

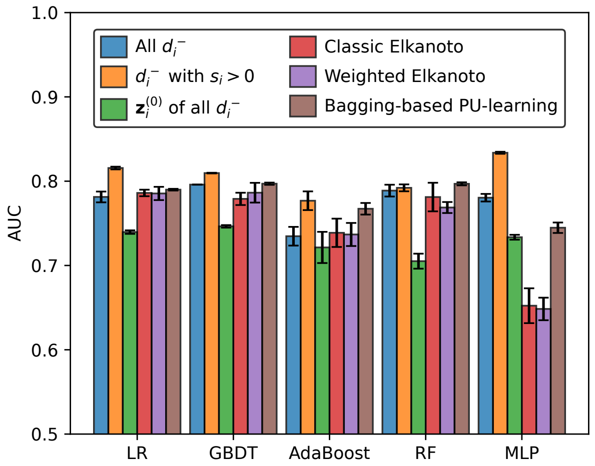


Figure 14: AKI prediction performance of widely adopted classifiers for three different settings of our proposal and three PU learning baselines.

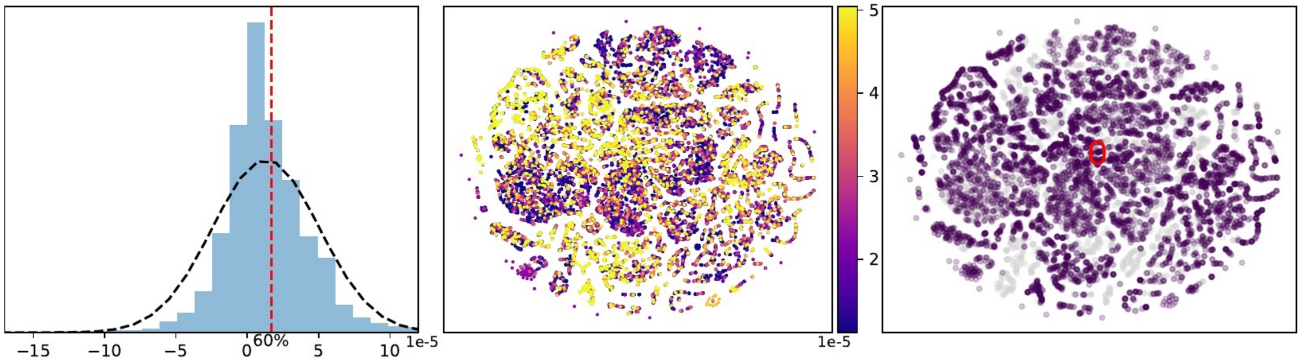
G.4. Comparison with Influence Function-based Data Valuation

As an alternative data valuation technique, influence functions (Weisberg & Cook, 1982) assess how the prediction model changes when the weight of a single sample is slightly altered. In this experiment, we compare our proposed data valuation technique, specifically data Shapley values for negative samples, against influence functions for negative samples. Specifically, in line with standard practice for comparing with influence functions, we employ the standard leave-one-out (LOO) method (Ghorbani & Zou, 2019; Rozemberczki et al., 2022) to calculate the influence function value for each negative sample.

The cohort discovery results of the influence function-based data valuation for AKI prediction are presented in Figure 15. Figure 15(a) displays the histogram of influence function values among all negative samples. This distribution can be fitted by a single Gaussian distribution with a mean of zero, which is different from the distribution of data Shapley values presented in Figure 3(a). To focus our analysis on the most influential negative samples, we set a threshold of 60% and exclude the lower 60% of negative samples based on their influence function values. Figure 15(b) illustrates the distribution of all negative samples in terms of their influence function values in the manifold space. From this figure, we cannot discern significant hot zones (i.e., with high influence function values), as samples with different temperatures are completely mixed together. This suggests that if influence functions are used for data valuation, they do not exhibit a clear trend of spatial proximity similarity in the feature vector space, which may make it challenging to effectively benefit spatial clustering algorithms for further cohort discovery. This concern is further experimentally validated, as shown in Figure 15(c). We implement the same subsequent clustering steps as in Figure 3(c) in the negative sample space based on influence functions (i.e., selecting 40% of high influence function values as input for DBSCAN). From Figure 15(c), it can be seen that DBSCAN fails to generate meaningful clustering results and only produces a single cluster. These cohort discovery results indicate that influence function-based data valuation does not reveal meaningful medical cohorts for AKI prediction.

Next, we compare the AKI prediction performance achieved by a new setting, "d_i^- with $IF > 0$," which includes only the negative samples with positive influence function values, with three different settings of our proposal. The experimental results in terms of AUC (mean ± std) from five repeats are shown in Figure 5. It is observed that influence functions underperform data Shapley values ("d_i^- with $IF > 0$" vs. "d_i^- with $s_i > 0$") and generally achieve lower AUC than using all negative samples (the setting "All d_i^-"). This indicates that the setting "d_i^- with $IF > 0$" deletes informative data samples from all negative samples rather than retaining them. Consequently, this set of experiments confirms that, compared to influence functions, data Shapley values are a more suitable and effective data valuation measure for the discovery of the relevant medical cohorts in AKI.

It is worth noting that data Shapley values and influence functions have also been compared as data valuation techniques in prior work (Ghorbani & Zou, 2019), where data Shapley values exhibit a significant performance advantage over influence functions. This aligns with our findings from the evaluation results, confirming the superiority of data Shapley values over influence functions. Moreover, influence functions, calculated as the performance difference of the prediction model with



(a) Influence function value histogram among all negative samples

(b) Influence function value distribution among all negative samples in the manifold space

(c) Discovered cohorts among high influence function value negative samples

Figure 15: Cohort discovery of influence function-based data valuation for AKI prediction.

and without a specific negative sample, do not satisfy equitability conditions. This limitation of influence functions arises from their inability to account for a sample's complex interactions with other samples (Ghorbani & Zou, 2019). However, this limitation deviates from our research focus on cohort discovery. Specifically, if there is a particular cohort, meaning there is a cluster of samples with a sufficient number of samples in it, and they are close to each other in the feature space and in terms of their data valuation, then the influence function value of any single sample in that cohort should be relatively low. Hence, influence functions may be more suited for identifying rare or isolated medical cases, which, while interesting, falls outside the scope of our research. Additionally, influence functions have also been identified to have robustness issues in prior work (Ghorbani et al., 2019).

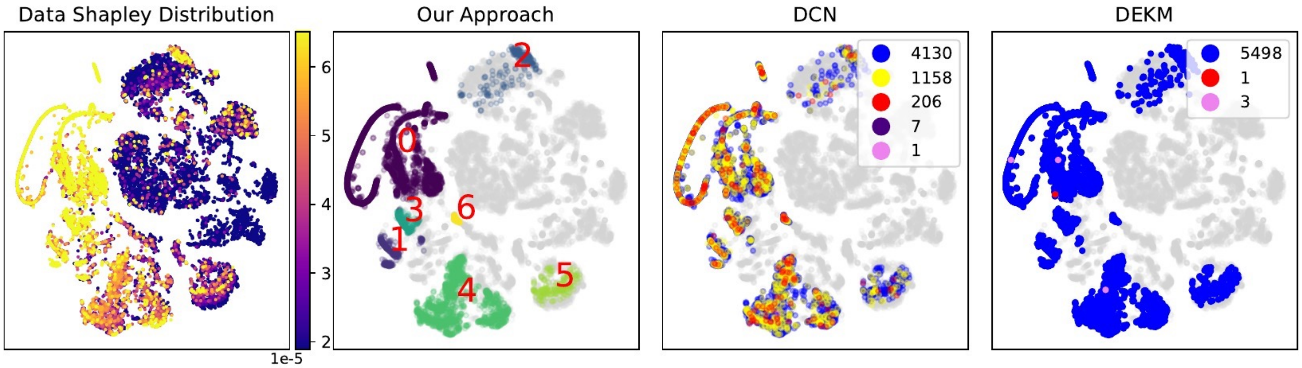
G.5. Comparison with Deep Clustering Baselines

In recent years, the concept of deep clustering, which involves simultaneously optimizing representation learning and clustering, has gained increasing attention. We further compare our proposed cohort discovery approach with two deep clustering baselines: the deep clustering network (DCN) (Yang et al., 2017) and the deep embedded K-means clustering (DEKM) (Guo et al., 2021). Specifically, DCN achieves joint dimensionality reduction by training a deep neural network alongside K-means clustering, while DEKM transforms the embedding space further to a new space in order to reveal cluster-structure information.

The comparison of cohort discovery results among high data Shapley value negative samples for our approach, DCN, and DEKM is shown in Figure 16. In terms of results, neither of the baseline methods can identify significant cohorts as effectively as our proposed approach does. Instead, they classify the majority of points into one large cluster with a few smaller cohorts (indicated in the legend by the number of samples per cohort). Notably, these results align with those obtained using cPCA in Figure 13 and our proposal without the isotropy constraint in Figure 12(b) in the ablation study. A shared characteristic of these methods is their two-step process: (i) dimensionality reduction of raw data using methods such as AE, SAE, SDAE, or cPCA, and (ii) clustering on the embedded representations using K-means or DBSCAN. This indicates that the subtle differences resulting from specific implementations of dimensionality reduction and clustering methods do not lead to significant differences in cohort discovery in the application scenario.

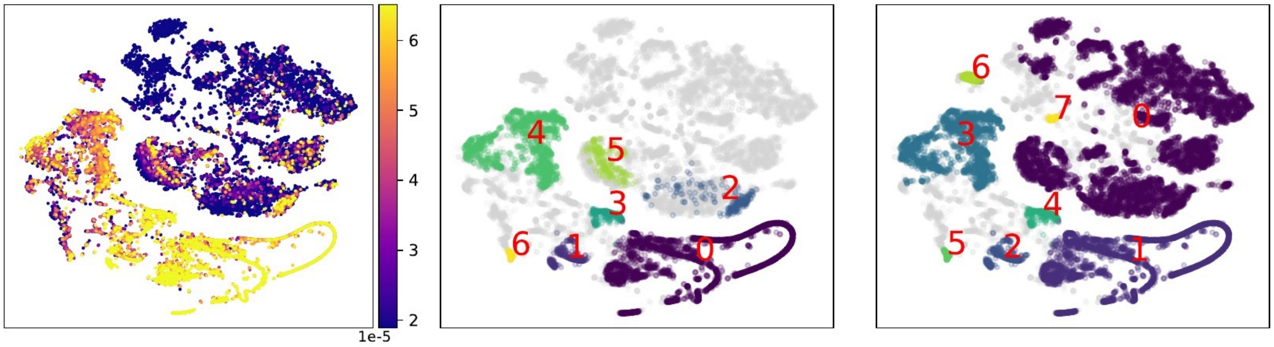
The key difference between them and our proposed approach lies in the absence of our proposed isotropy constraint, which ensures uniform changes in data Shapley values across orientations, facilitating the identification of hot zones in the field. These results substantiate the innovative contribution we have made in this paper by incorporating data valuation into medical cohort discovery. This highlights the importance of our approach, as traditional methods solely based on feature similarity struggle to effectively differentiate between different cohorts in complex and high-dimensional medical data. This Shapley-based exploration of interrelationships between samples is our primary innovation for real-world clinical practice.

In summary, these comparison results with DCN and DEKM, along with the previous results of cPCA and our proposal without the isotropy constraint, further validate the effectiveness of our proposed cohort discovery approach, underscoring its medical value.



(a) Data Shapley value histogram among all negative samples
(b) Our approach's discovered cohorts among high data Shapley value negative samples
(c) DCN's discovered cohorts among high data Shapley value negative samples
(d) DEKM's discovered cohorts among high data Shapley value negative samples

Figure 16: Comparison of cohort discovery results between our approach, DCN and DEKM.



(a) Data Shapley value distribution among all negative samples in the manifold space
(b) Discovered cohorts among high data Shapley value negative samples
(c) Discovered cohorts among all negative samples

Figure 17: Comparison of cohort discovery results between clustering high data Shapley value negative samples and clustering all negative samples.

G.6. Comparison with Clustering All Negative Samples

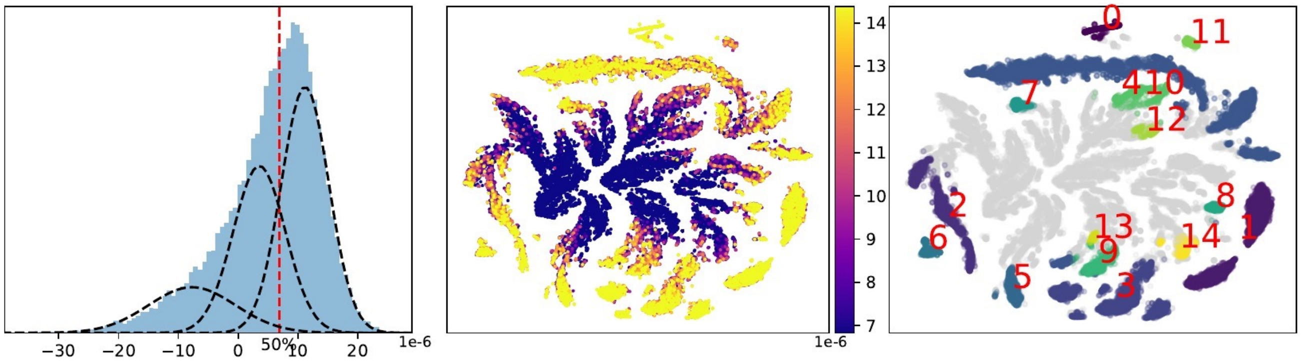
We conduct a comparison between our approach's cohort discovery results among high data Shapley value negative samples and results among all negative samples. The experiment results are presented in Figure 17, with Figure 17(a) illustrating the data Shapley value distribution among all negative samples in the manifold space, and Figure 17(b) and (c) serving as counterparts for comparison. In Figure 17(b), grey points represent samples either possessing low data Shapley values or being labeled as noise by DBSCAN. Conversely, in Figure 17(c), grey points exclusively correspond to samples labeled as noise by DBSCAN.

Our approach, even after incorporating all negative samples for clustering, remains capable of identifying representative cohorts primarily composed of high data Shapley value negative samples, with only a few merging with the large cluster of low data Shapley value negative samples.

It is essential to highlight that cohorts with high positive data Shapley values demonstrate a clear ability to facilitate the model to identify positive samples, indicating their fundamental value in medical research, and hence, we only focus on them in this paper. In contrast, the medical significance of cohorts without high positive data Shapley values (negative or near zero) becomes very unclear. They can either hinder the prediction task by potentially being erroneous or noisy data points, as indicated by negative data Shapley values, or they may represent complex and healthy samples of limited relevance, with data Shapley values close to zero.

Table 3: Key statistics of the MIMIC-III dataset for in-hospital mortality prediction.

Statistics	MIMIC-III Dataset
# of admissions	51826
# of positive samples	4280
# of negative samples	47546
# of lab tests	428
Input Window	48 hours



(a) Data Shapley value histogram among all negative samples

(b) Data Shapley value distribution among all negative samples in the manifold space

(c) Discovered cohorts among high data Shapley value negative samples

Figure 18: Cohort discovery of our proposal for in-hospital mortality prediction on the MIMIC-III dataset.

G.7. Evaluation on the MIMIC-III Public Benchmark Dataset

We evaluate our proposed cohort discovery approach on the widely recognized MIMIC-III dataset (Johnson et al., 2016). This dataset is esteemed as a benchmark in healthcare analytics and comprises EMR data for dozens of thousands of patients who are admitted to intensive care units (ICU) between 2001 and 2012. Our focus centers on predicting in-hospital mortality on the MIMIC-III dataset, leveraging laboratory test data as input. We define each patient admission as an individual sample if the duration of the admission exceeds 48 hours. Subsequently, we label each admission by assessing whether the patient passes away during their stay at the hospital. Utilizing a set of 428 laboratory tests within a 48-hour "Input Window," we predict the likelihood of mortality for each admission. Prior to using these test values as input features, we apply min-max standardization, followed by averaging the results. The key statistics pertaining to the MIMIC-III dataset for in-hospital mortality prediction are summarized in Table 3.

To validate the broad applicability of our proposed approach, we conduct the same cohort discovery analysis on the MIMIC-III dataset as described, with the results shown in Figure 18. By fitting the data to a Gaussian mixture model, we unveil three distinct components within the data Shapley value histogram, as depicted in Figure 18(a). To focus our investigation on the third component that holds significant relevance to the prediction task, we set a 50% threshold to exclude the lower 50% negative samples based on their data Shapley values while retaining the remaining 50% for further analysis. Figure 18(b) displays the data Shapley value distribution of all negative samples in the manifold space. Subsequently, we perform DBSCAN on the extracted 50% of negative samples exhibiting high data Shapley values, successfully discerning fifteen distinct cohorts, as demonstrated in the t-SNE plots in Figure 18(c). These fifteen cohorts effectively decompose the third component of Figure 18(a) into respective Gaussian distributions, as confirmed in Figure 19, validating consistent data Shapley values within each identified cohort. Compared to the results found in the AKI prediction, we can see that in terms of ICU mortality prediction, the negative sample cohorts are more distinct and prominent in the constructed Negative Sample Shapley Field. This aligns with expectations, as the negative samples in mortality (i.e., cases where patients are successfully resuscitated and discharged alive) among patients on the brink of death admitted to ICU often stem from causes that are more multi-sourced and more explicit, compared to those in a single-specialty nephrology department. This substantiates the robustness and capability of our cohort discovery approach in decomposing samples with high data Shapley values into medically relevant and distinct cohorts.

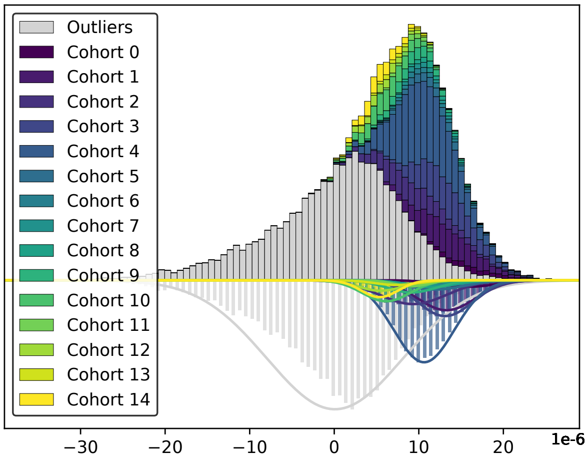


Figure 19: Data Shapley value histogram of the samples within our discovered cohorts on the MIMIC-III dataset.

The outcomes of this cohort discovery evaluation hold substantial promise in advancing our comprehension of ICU patients, and they are poised to contribute significantly to patient care and survival, particularly when combined with systematic and cross-sectoral analytical research. Our experimental findings underscore the efficacy of our proposed approach on this well-established public benchmark dataset, demonstrating its potential for broader application in the analysis of other medical datasets, thus highlighting its versatility and robustness.

H. Clinical Impact and Implications

H.1. Clinical Value of the Identified Cohorts in AKI Prediction

The identified cohorts among negative samples hold significant medical value as they unveil medically meaningful patterns, as exemplified in Section 4.3 of our paper. These patterns shed light on the investigated medical problem, in this case, AKI prediction. Further, the discovered cohorts can reveal potential future positives, pathological correlations, or similar conditions. This reciprocal relationship between negative and positive samples can contribute to defining positive samples in theoretical medical research. Such insights can significantly enhance clinicians' comprehension of the disease and the distribution of visiting patients during practical clinical consultations.

Let us illustrate with a more specific example. In this paper, we propose a cohort discovery approach on AKI negative sample cases admitted to the nephrology department, resulting in the identification of several cohorts (discussed in Section 4.3). Among them, our further investigation leads us to classify Cohort 2 as an "inflammatory cohort."

We believe this cohort reveals a clinical insight: patients with infections are often treated with antibiotics; some develop sepsis which in turn increases the metabolic burden on the kidneys. In some instances, the antibiotics used may carry nephrotoxic potential and contribute to drug-induced AKI, leading to the patients being triaged to nephrology for diagnosis and treatment.

However, even in cases of severe infection, physicians adhere to prompt resuscitation and early antimicrobial therapy, meaning that, in most cases, this does not lead to further renal damage (AKI). Nephrotoxic antimicrobials are administered with vigilance and frequent drug level monitoring with appropriate dose adjustments could reduce kidney risks. From a data perspective, these cases form a cohort with distinctive infection characteristics and antibiotic usage, and kidney injury may be subclinical and not be severe sufficiently to manifest with raised serum creatinine levels, representing a cohort among negative samples.

The clinical insights derived from the analysis of this cohort are of great medical value. For instance, nephrologists, when assessing patients, consider the toxicity of antibiotics administered recently. If a patient's profile matches these characteristics, a more conservative approach and further observation may be preferred. Such an approach may limit the extent of subclinical kidney injury and may reduce cumulative kidney damage over time in the case of repeated organ insults from acute illnesses.

Traditionally, clinical treatment strategies based on cohort insights rely on human experience and considering the complexity of clinical treatment, these experiences often tend to be one-sided. In the era of data science, we can use learning techniques to help us gain a deeper understanding of diseases and their treatments from new perspectives.

H.2. Clinical Impact Beyond AKI Prediction

Our proposed cohort discovery approach has the potential to yield significant medical impact across diverse medical applications beyond AKI prediction.

For instance, when a patient exhibits symptoms of respiratory tract infection, outpatient physicians typically prioritize measuring the patient's body temperature and inquiring about any recent travel history or epidemiological contacts. This essentially constitutes leveraging an identified negative sample cohort for efficient medical delivery, specifically for the diagnosis of high-risk infectious viral diseases such as COVID-19. This is crucial because, in this scenario with limited medical resources, it is impractical to triage all coughing patients into isolation wards.

In this routine example, the "patient without fever" can be considered to exhibit a high data Shapley value in the context of the COVID-19 task, representing a high-temperature zone within the constructed Negative Sample Shapley Field. Although our proposed approach stems from clinical experience and the examination of relevant medical data, we believe that this innovative approach offers valuable insights for other application domains (e.g., to discover typical classification and exclusion strategies in complex problems). Its recognition aids in deepening the understanding of negative samples, particularly in discerning the typical patterns of negative samples specific to certain tasks.