SemanticMask: A Contrastive View Design for Anomaly Detection in Tabular Data

Shuting Tao[1,2], **Tongtian Zhu**[1], **Hongwei Wang**[1,2*] and **Xiangming Meng**[1,2*]
[1]College of Computer Science and Technology, Zhejiang University
[2]The Zhejiang University-University of Illinois Urbana-Champaign Institute, Zhejiang University
{shuting.17, hongweiwang, xiangmingmeng}@intl.zju.edu.cn, raiden@zju.edu.cn

Abstract

Contrastive learning based on data augmentation techniques has recently achieved substantial advancement in learning a representation well-suited for anomaly detection in image domain. However, due to the lack of spatial structure, designing effective data augmentation methods for tabular data remains challenging. Conventional techniques, such as random mask, disregard the inter-feature correlations and fail to accurately represent the data. To address this issue, we propose a novel augmentation technique called SemanticMask which leverages the semantic information from column names to generate better augmented views. SemanticMask aims to ensure that the shared information between views contains sufficient information for anomaly detection without redundancy. We analyze the relationship between shared information and anomaly detection performance and empirically demonstrate that good views for tabular anomaly detection tasks are feature-dependent. Our experiment results validate the superiority of SemanticMask over the state-of-the-art anomaly detection methods and existing augmentation techniques for tabular data. In further evaluations of the multi-class novelty detection task, SemanticMask also significantly outperforms the baseline.

1 Introduction

Anomaly detection has extensive applications across various domains, such as medical diagnosis, credit card fraud, and industrial equipment maintenance [Chandola *et al.*, 2009; Ruff *et al.*, 2021]. The goal of anomaly detection is to identify whether a test sample falls into the same distribution as the training data, which can be viewed as a one-class classification problem [Schölkopf *et al.*, 1999]. However, most anomalies in real-world applications are typically new and unknown, thus posing great challenges for supervised anomaly detection. By contrast, without the need to access to supervision of anomalies, self-supervised anomaly detection (SSAD) becomes much more preferred [Sehwag *et al.*, 2021]. Specifically, contrastive learning (CL) based-SSAD utilizing data augmentation strategies, e.g., ContrastiveCrop [Peng *et al.*, 2022], colorization [Tian *et al.*, 2020a], and CutPaste [Li *et al.*, 2021], has recently made significant progress for anomaly detection in the image domain [Cho *et al.*, 2021; Tack *et al.*, 2020].

Unfortunately, for data other than images, such as tabular data, despite its widespread usage, little consideration has been paid on what data augmentation technique can generate a useful representation for anomaly detection. It is also hard to design these transformations manually, because current skills heavily depend on the prior knowledge of the spatial structure of images. Existing augmentation skills for tabular data, such as mixup [Verma *et al.*, 2021], internal contrastive learning (ICL) [Shenkar and Wolf, 2022] and random corrupt [Yoon *et al.*, 2020; Bahri *et al.*, 2021], primarily rely on randomness and overlook the semantic information of features, potentially degrading the quality of data representation.

In CL, selecting informative augmented views is crucial in producing a meaningful representation that helps downstream anomaly detection. The shared information between different augmented views controls the information that the representation extracts [Tian *et al.*, 2020b]. Therefore, if anomalies are attributed to specific features, it is essential to include these features in the shared information for a responsive representation. In conventional anomaly detection tasks, the specific features with abnormal values that lead to anomalies remain unidentified. As a result, it is very important to balance the amount of shared information between views so that we can generate representations that are not only robust to insignificant variations but also still preserving the task-relevant information. One natural question arises: ***How can we design an effective data augmentation method to craft better contrastive views for anomaly detection in tabular data?***

It is noteworthy that for real-word tabular data, the column names usually carry valuable semantic information [Wang and Sun, 2022]. This implies that features with semantically similar column names are typically correlated, exhibiting similar patterns of abnormal values when anomalies occur. For instance, anomalies that indicate the presence of diabetes in the Pima dataset [Rayana, 2016] may exhibit anomalous values in semantically related features, such as "insulin" and "glucose". To leverage such semantic connection, we propose a semantic-aware masking approach (dubbed as Se-

*Corresponding Authors.

manticMask) for CL-based anomaly detection. Specifically, SemanticMask uses a language model and clustering algorithm to divide features into k clusters, grouping those with semantically similar column names together in the same cluster. In designing contrastive views, as anomalous features are unknown, it is crucial that the shared information includes full feature coverage across the entire semantic spectrum, i.e., the shared features between augmented views include at least one feature from each cluster. SemanticMask achieves this by performing a random and equitable division of these clusters into two distinct subsets. Each of the two augmented views then selects different subsets and applies a predetermined masking proportion to each cluster within its respective selected subset, while leaving unselected clusters unaffected. Consequently, SemanticMask lowers the likelihood of information loss of each cluster, while also preventing redundancy, thereby achieving balance in the amount of shared information between views. In contrast, traditional data augmentation techniques, such as random mask, overlook this semantic connection. They randomly select features as shared information, thus the shared features are more likely to exclusively come from partial clusters, resulting in an insufficient representation that fails to capture the crucial anomalous features. The main contributions are summarized as follows:

- We propose SemanticMask, a method that incorporates the semantic information of column names to create effective augmented views for anomaly detection in the tabular domain. Additionally, we introduce two extended variations: one incorporates a mask estimation module to address situations where categorical features naturally contain zero values that cause ambiguity; and the other integrates a one-sentence prompt describing the anomaly detection task as prior knowledge, assisting in the selection of shared features.
- We analyze the relationship between the amount of shared information and downstream anomaly detection performance in various settings of SemanticMask's mask probability. We also empirically demonstrate that good views for CL-based anomaly detection are feature-dependent.
- SemanticMask surpasses state-of-the-art anomaly detection methods and augmentation techniques for tabular data. In addition, we also extend SemanticMask to multi-class novelty detection and further demonstrate its effectiveness and versatility.

2 Related Works

Anomaly detection. Studies on unsupervised anomaly detection can be broadly classified into four categories [Sehwag *et al.*, 2021; Yang *et al.*, 2021]: (1) density-based methods [Eskin, 2000; Zhai *et al.*, 2016; Li *et al.*, 2020], which detect anomalies by assessing data points in low-density regions compared to nearby high-density areas.; (2) reconstruction-based methods [Pidhorskyi *et al.*, 2018; Yan *et al.*, 2021; Nguyen *et al.*, 2019], which identify anomalies by comparing the reconstruction cost of a reconstruction model trained on the normal data; (3) classification-based methods [Ruff *et al.*, 2018; Wang and Cherian, 2019; Reiss *et al.*, 2021], which separate space containing normal data from all other regions; (4) self-supervised methods [Sehwag *et al.*, 2021; Cho *et al.*, 2021; Tack *et al.*, 2020; Li *et al.*, 2020; Golan and El-Yaniv, 2018], which utilize the strong representation learned from self-supervision [Sehwag *et al.*, 2021; Sohn *et al.*, 2021] or design surrogate tasks to help distinguish anomalies from normal samples [Tack *et al.*, 2020; Shenkar and Wolf, 2022]. Our study belongs to category (4) the self-supervised approach, as we use CL to learn representations. Studies by Sehwag et al. [Sehwag *et al.*, 2021] and Cho et al. [Cho *et al.*, 2021] have shown that CL results in significant improvements in detection performance, as it enhances the quality of the learned representations.

Data augmentation for tabular data in CL. Data augmentation is essential in CL tasks that produce effective representations [Chen *et al.*, 2020]. Most existing techniques apply only to images because their way of defining similarity rely heavily on spatial relationships [Tian *et al.*, 2020a; Chen *et al.*, 2020; Oord *et al.*, 2018; He *et al.*, 2020]. Due to the lack of a "common" correlation structure, designing data augmentation techniques for tabular data can be challenging [Yoon *et al.*, 2020]. There are a few existing methods. Mixup [Verma *et al.*, 2021] creates positive pairs by mixing data samples. Scarf [Bahri *et al.*, 2021] forms augmented views by corrupting a random subset of features. ICL [Shenkar and Wolf, 2022] considers a subset of features from a sample, along with its remaining features, as positive pairs to train the network. However, these methods primarily depend on randomness and thus ignore the semantic information inherent in features, making it difficult to guarantee the quality of the augmented views.

Good views for CL. In CL, shared information between augmented views controls the information that the representation learns and thus significantly impacts the representation quality [Oord *et al.*, 2018]. Tian et al. [Tian *et al.*, 2020a] have demonstrated that there exists a sweet spot of mutual information between views that leads to the best performance in the downstream task. Views with either too high or too low mutual information result in inferior representations. Building on this, Tian et al. [Tian *et al.*, 2020b] suggest that reducing the shared information between views and retaining the task-relevant information exclusively can enhance downstream performance. These previous studies primarily investigate the relationship between shared information and representation quality when the downstream task is classification for image data. This paper delves into the relationship when the downstream task is anomaly detection for tabular data.

3 Problem Formulation

Given a table T with column description (e.g., column names) composed of n rows and d columns, where each row is a training sample $\boldsymbol{x}$ and each column represents a feature $\boldsymbol{f}$, we can obtain the training dataset $\boldsymbol{X}_{train} \in \mathbb{R}^{n \times d}$ and the set of column names $C = \{c_1, \ldots, c_d\}$. Column names are usually in the form of short phrases. The underlying distribution of the training data is denoted as $\mathbb{P}^{in}_{\mathcal{X}}$. Anomalies not belonging to $\mathbb{P}^{in}_{\mathcal{X}}$ denote semantic anomalies coming from novel classes,

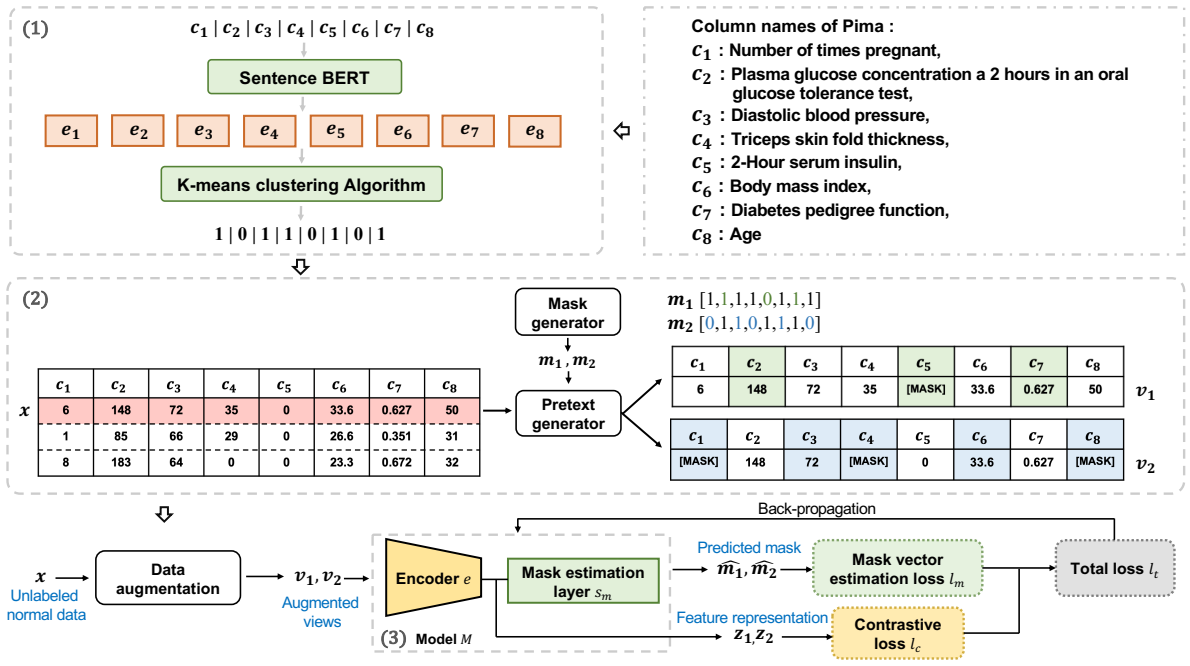


Figure 1: Block diagram of the SemanticMask framework for representation learning on tabular data. (1) SBERT and k-means clustering algorithms are employed to cluster features based on the column names C. Clusters from different subsets are distinguished by the shades of blue and green. (2) The mask generator utilizes the partition information to generate binary mask vectors $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$, which are then applied to the original data $\boldsymbol{x}$, yielding the augmented views $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. (3) A contrastive encoder e and a mask estimation layer s_m together form the complete model M. The augmented views are first processed by the encoder e to obtain representations $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, with contrastive loss l_c calculated from them. These representations are then fed into s_m to obtain the predicted mask, and we calculate the reconstruction loss l_m using the original mask and the predicted mask. The total loss l_t, a weighted sum of l_c and l_m, is used to update the entire model M.

characterized by abnormal feature values [Yang *et al.*, 2021]. The task is to determine whether a new data sample belongs to the data distribution $\mathbb{P}_{\mathcal{X}}^{in}$ or not. Our detectors aim to address this task, with access to only $\boldsymbol{X}_{train}$ and C.

4 Proposed Method: SemanticMask

In this section, we introduce an SSAD method for tabular data that incorporates a semantic-aware data augmentation technique called SemanticMask. The motivation behind this work is to utilize the semantic cues inherent in column names to extract crucial information on the correlation between features. This information can enhance the comprehension of the tabular data [Suhara *et al.*, 2022], and ultimately improve downstream anomaly detection tasks.

4.1 SemanticMask: A Semantic-Aware Masking Scheme for CL-Based Anomaly Detection

The framework of SemanticMask includes three steps. In the first step, we employ Sentence-BERT (SBERT) [Reimers and Gurevych, 2019] to acquire sentence-level embeddings for column names. SBERT is a modification of the pre-trained BERT model that generates fixed-size, semantically meaningful vectors for input sentences. "Semantically meaningful" here means that semantically similar sentences are close in the vector space. Then we partition features into k clusters, designated as g_1 through g_k, based on the embeddings of the column names. This partitioning is performed using the k-means clustering method, due to its effectiveness and low computational complexity [Hartigan *et al.*, 1979]. The value of hyperparameter k adapts to the feature dimensionality. As the dimensionality increases, k also increases. In Figure 1, we take the Pima diabetes dataset [Rayana, 2016] as an example. This dataset employs eight features to determine whether a female patient will develop diabetes, with k set to 2. After k-means clustering, features 2, 5, and 7 have been grouped into one cluster, while the remaining features have been grouped into the other cluster.

In the second step, we propose a data augmentation module that generates positive pairs for CL. Traditional tabular data augmentation techniques, such as RandomMask, indiscriminately mask features with equal probabilities, disregarding the inherent correlations between features and the impact of shared information on representation. These techniques often leads to insufficient representation. To tackle this issue, we design a semantic-aware augmentation technique called SemanticMask, which effectively adjusts the shared information between views to an appropriate level. In the process of generating augmented views, k clusters are evenly divided into two disjoint subsets, labeled G_1 and G_2, with each subset incorporating features from $\lfloor k/2 \rfloor$ clusters. Each of the two augmented views selects a different subset for masking. When generating one of the augmented view $\boldsymbol{v}_i$ of the sample $\boldsymbol{x}$, a subset of features denoted as G_i is initially selected. Subsequently, a binary vector $\boldsymbol{m}_i = [m_{i,1}, \ldots, m_{i,d}]^\top \in \{0,1\}^d$

is produced by a mask vector generator. Specifically, for the j-th feature f_j of $\boldsymbol{x}$, if it belongs to clusters in G_i, then $m_{i,j}$ follows a Bernoulli distribution with a probability p_m of being zero and $1-p_m$ of being one. For features not belonging to G_i, $m_{i,j}$ is set to one. The generation of $\boldsymbol{v}_i$ is given by $\boldsymbol{v}_i = \boldsymbol{m}_i \odot \boldsymbol{x}$. $\odot$ here denotes element-wise multiplication. In the example shown in Figure 1, where there are only two clusters ($k=2$), each subset G_i contains a single cluster g_i.

In the third step, we pass the augmented view $\boldsymbol{v}_i$ through an encoder network e to get representation $\boldsymbol{z}_i$. The InfoNCE contrastive loss l_c is computed, encouraging representations of the same instance to be close and representations of different instances to be far apart [Chen *et al.*, 2020]. Additionally, for tabular datasets that contain categorical features, it is common for many cells to have a value of zero. In order to differentiate between cells that have been masked to zero and those that are naturally zero, we add an optional mask estimation layer s_m to predict which features have been masked in each augmented view. Specifically, this module utilizes the encoded representation of each view $\boldsymbol{z}_i$ as input, and outputs a vector $\hat{\boldsymbol{m}}_i$ that predict which features of $\boldsymbol{x}$ have been masked. The total objective loss l_t of encoder is:

$$l_t = l_c + \lambda \cdot l_m, \tag{1}$$

where the first loss function l_c corresponds to the contrastive loss, and the second loss function l_m, representing the reconstruction loss of the mask vector $\boldsymbol{m}_i$, is defined as follows:

$$l_m(\boldsymbol{m}_i, \hat{\boldsymbol{m}}_i) = \frac{1}{d}\left[\sum_{j=1}^{d}\left(m_{i,j} - (e \circ s_m)_j(\boldsymbol{v}_i)\right)^2\right]. \tag{2}$$

In equation (2), $e \circ s_m$ represents the entire model M in Figure 1, and $\boldsymbol{v}_i$ is the input to M. The expression $(e \circ s_m)_j(\boldsymbol{v}_i)$ is the output of M, representing the value in the j-th position of the predicted mask. The trade-off between these two losses is determined by the hyperparameter λ.

Once the training of the encoder e is completed, we use Mahalanobis distance [Sehwag *et al.*, 2021] to calculate the anomaly score $s_{\boldsymbol{x}_t}$ for the new sample $\boldsymbol{x}_t$:

$$s_{\boldsymbol{x}_t} = (\boldsymbol{z}_{\boldsymbol{x}_t} - \boldsymbol{\mu}_{train})^T \boldsymbol{\Sigma}_{train}^{-1}(\boldsymbol{z}_{\boldsymbol{x}_t} - \boldsymbol{\mu}_{train}), \tag{3}$$

where $\boldsymbol{z}_{\boldsymbol{x}_t}$ is the representation of $\boldsymbol{x}_t$ encoded by e, with $\boldsymbol{\mu}_{train}$ and $\boldsymbol{\Sigma}_{train}$ being the sample mean and covariance of the training data's representation.

4.2 Incorporation of SemanticMask with One-Sentence Task Description

Based on SemanticMask, we propose an extension that leverages a single sentence s as prior knowledge to facilitate the early identification of important features for anomaly detection. The sentence s is structured using a prompt template, "***The task is to detect {description}.***", with a specific description relevant to the dataset. For instance, the sentence "***The task is to detect the presence of heart disease.***" is used to describe the Heart disease dataset [Derrac *et al.*, 2015]. We input s along with the column names $C = \{c_1, \ldots, c_d\}$ into the SBERT model. We then apply k-means clustering to divide $\{s, c_1, \ldots, c_d\}$ into k clusters, with one cluster containing s, and the remaining clusters without s. The features whose column names belong to the cluster containing s corresponds to the feature set g_s. Aligned with SemanticMask, we proceed to partition clusters into two subsets.

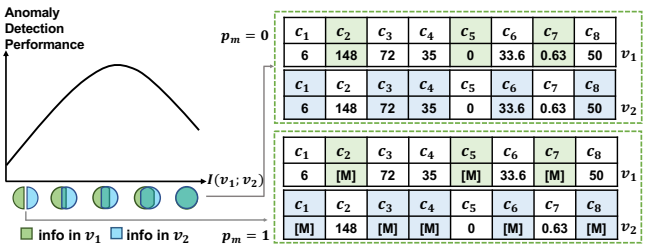
In the phase of generating augmented views, a mask vector is constructed. Specifically, a mask probability of $p_m - \varepsilon$ is applied to the cluster g_s, while other $k-1$ clusters adopt a mask probability of $p_m + \varepsilon$, where $\varepsilon \in (0, p_m)$ serves as a hyperparameter. By doing so, the shared information contains more features that are semantically relevant and closely related to the task description s. Incorporating the relationship between features and the task description into the SemanticMask framework can assist in feature selection, leading to improved performance in anomaly detection.

5 What Makes Good Augmented Views for Anomaly Detection

This section investigates the impact of view selection on shared information and downstream anomaly detection performance in two ways: (1) we combine detailed analysis and empirical evidence to show that the optimal performance for anomaly detection is achieved when the amount of shared information between views is at an appropriate level; (2) we demonstrate that good views of anomaly detection depend on the features that cause the anomalies.

5.1 View Selection Influences Shared Information and Detection Performance

After getting augmented views $\boldsymbol{v}_1, \boldsymbol{v}_2$ of the sample $\boldsymbol{x}$ by SemanticMask, we trace out how the amount of shared information between views affects the downstream anomaly detection performance. Previous studies have shown that for computer vision classification tasks, optimal performance can be achieved by a set of views that retains relevant information while eliminating all irrelevant information [Tian *et al.*, 2020b]. However, in anomaly detection, views of training data lack anomaly information. We do not know which features are anomalous, so we want to preserve the shared information as comprehensively as possible, in order to include task-relevant information while avoiding excessive redundancy. In the SemanticMask approach, the amount of shared information is controlled by adjusting the cluster-wise



c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
6	148	72	35	0	33.6	0.63	50
c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
6	148	72	35	0	33.6	0.63	50

c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
6	[M]	72	35	[M]	33.6	[M]	50
c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
[M]	148	[M]	[M]	0	[M]	0.63	[M]

Figure 2: As the cluster-wise mask probability p_m decreases, the information present in each view and the shared information between views increases, gradually incorporating the feature information relevant to the downstream anomaly detection until it is over-included, thereby affecting the quality of the learned representation.

mask probability p_m. In the following, we present an analysis that suggests that the optimal performance for anomaly detection is achieved when the shared information between views is moderate, neither too high ($p_m = 0$) nor too low ($p_m = 1$).

SemanticMask partitions the feature set F of sample $\boldsymbol{x}$ into two disjoint subsets, G_1 and G_2, based on the semantic information of the column names, where $G_1 \cup G_2 = F$. Assuming the augmented view $\boldsymbol{v}_1$ selects G_1 for masking, SemanticMask applies a mask probability $p_m \in [0, 1]$ to each cluster within G_1 to generate the masked version $G_1^{'}$, resulting in $\boldsymbol{v}_1 = G_1^{'} \cup G_2$. Similarly, $\boldsymbol{v}_2 = G_1 \cup G_2^{'}$. Adjusting the masking ratio p_m from 0 to 1 controls the quantity of information in each view and the shared information between views. When $p_m = 0$, both views are identical, $\boldsymbol{v}_1 = \boldsymbol{v}_2 = G_1 \cup G_2 = F$. Conversely, when $p_m = 1$, each view comprises distinct unmasked feature subsets: $\boldsymbol{v}_1 = G_2, \boldsymbol{v}_2 = G_1$.

The feature decoupling principle [Wen and Li, 2021] provides a theoretical explanation for the underperformance when $p_m = 0$. Without data augmentation techniques, deep neural networks tend to learn representations that accentuate dense noise, which may overshadow the more semantically aligned sparse features. When $p_m = 1$, $\boldsymbol{v}_1 = G_2, \boldsymbol{v}_2 = G_1$. Our analysis follows the concept of minimal sufficient representation $\boldsymbol{z}_{min}$ in CL, as defined in [Wang *et al.*, 2022], which includes the shared information between views and disregards any non-shared information. In the context of CL applied to the downstream anomaly detection task T, the minimal sufficient representation of $\boldsymbol{v}_1$ is denoted as $\boldsymbol{z_1}^{min}$. In the ideal case, we assume that there is no interrelation among different clusters; thus, G_1 and G_2 are semantically non-intersecting and mutually independent. Then $\boldsymbol{z_1}^{min}$ is found to lack sufficient task-relevant information, since $I(\boldsymbol{z_1}^{min}; T) = I(\boldsymbol{v_1}; T) - I(\boldsymbol{v_1}; T|\boldsymbol{v}_2) = I(G_2; T) - I(G_2; T|G_1) = 0$, leading to a suboptimal performance. A conceptual illustration of the above analysis is shown in Figure 2. The optimal point, located at the curve's peak, corresponds to the optimal $0 < p_m < 1$ that balances the shared information between views to an optimal level, covering the required feature information for anomaly detection without redundancy.

Empirically, we evaluate the anomaly detection performance by adjusting p_m on two public tabular datasets: the South African Heart dataset [Derrac *et al.*, 2015] and the Heart disease dataset [Derrac *et al.*, 2015]. The results, shown in Figure 5(a), reveal a reverse-U curve for both datasets, with the optimal point around $0 < p_m = 0.4 < 1$. This finding is consistent with our analysis illustrated in Figure 2.

5.2 Good Views Are Feature-Dependent

When anomalies are caused by certain anomalous features, we empirically validate that achieving a favorable detection performance relies on ensuring that the shared information between views includes relevant information about these features. Two anomaly detection tasks are conducted on the Pima diabetes dataset [Rayana, 2016].

In the first task, we generate a synthetic dataset by artificially modifying the values of the feature f_2, corresponding to the column named "Plasma glucose concentration a 2 hours in an oral glucose tolerance test". Normally, this feature ranges from 0 to 197 with a mean of 179. We select a subset

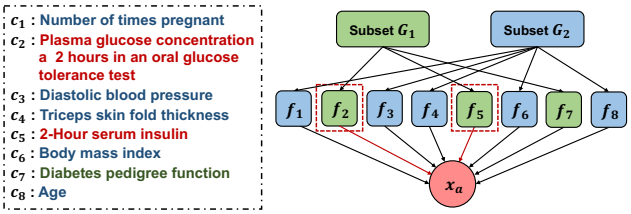


Figure 3: Synthetic anomalies induced by features f_2 and f_5, corresponding to c_2: "Plasma glucose concentration a 2 hours in an oral glucose tolerance test" and c_5: "2-Hour serum insulin".

Task 1					Task 2				
shared feature	f_2	f_1	G_1	G_2	shared feature	f_2	f_5	G_1	G_2
Acc(%)	90.3	69.0	91.2	54.4	Acc(%)	93.1	94.1	93.4	54.9

Table 1: We investigate the impact of the shared information between views on the representation quality by evaluating two synthetic anomalous datasets.

of normal samples, replace their f_2 values with values generated from a Gaussian distribution with a mean of 190 and a standard deviation of 5, and labeled this subset as anomalous. We then combine this subset with a portion of normal data to create test datasets for anomaly detection during the testing phase. Four experiments are designed to manipulate shared information between views in different forms: one including solely the feature f_2 itself, another involving the feature f_1 with the column name "Number of times pregnant" which is semantically unrelated to the column name of f_2, one comprising the subset G_1 that contains f_2, and the last encompassing the subset G_2 that does not include f_2. Results in Table 1 indicate a notable influence of the shared information between augmented views $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ on the anomaly detection accuracy (Acc) of the model. Specifically, it is crucial to ensure that the anomalous feature is included in the shared information, rather than other features. For instance, when $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ only share features like f_1 or the subset G_2 that are semantically unrelated to f_2, the CL approach can hardly capture information of f_2, leading to a performance decline.

In the second task, shown in Figure 3, we concurrently manipulate two features with semantically correlated column names, f_2 and f_5, to artificially induce anomalies in the dataset. Under normal conditions, the range of values for f_5 whose column name is "2-Hour serum insulin", lies between 0 and 744, with a mean of 68.8. Based on the anomalous samples selected from the first task, we modify the value of f_5 to values generated from a Gaussian distribution with a mean of 500 and a standard deviation of 5. Four scenarios of shared information between v_1 and v_2 are exhibited in this task. In Scenario 1, f_2 is shared; in Scenario 2, f_5 is shared; in Scenario 3, features of the subset G_1 that contains both f_2 and f_5 are shared; in Scenario 4, features of the subset G_2 are shared. According to Table 1, the accuracy in Scenario 1$\approx$ Scenario 2$\approx$ Scenario 3$>$ Scenario 4, which suggests that good performance is achieved when the shared information includes at least a portion of the relevant information associated with the anomalous features. Performance declines when all anoma-

Methods	Saheart	Pima	Heart	Wbc	Arrhythmia	Drug	Mammographic	Bupa	Seismic	Average
OC-SVM [Schölkopf *et al.*, 2001]	68.9	64.9	78.4	95.4	77.4	59.9	80.2	59.3	59.3	71.5
LOF [Breunig *et al.*, 2000]	61.9	66.7	80.1	95.0	76.3	54.1	44.3	62.5	61.7	67.0
COPOD [Li *et al.*, 2020]	59.1	64.0	68.3	**95.5**	/	54.8	22.4	57.7	/	60.3
ECOD [Li *et al.*, 2022]	60.7	59.7	57.8	89.2	/	55.8	46.7	53.2	/	60.4
FB [Lazarevic and Kumar, 2005]	63.3±1.0	66.7±0.3	76.7±3.7	94.7±0.2	76.0±0.3	54.5±0.9	68.1±6.2	62.5±0.5	62.6±1.6	69.4
DeepSVDD [Ruff *et al.*, 2018]	51.4±0.6	52.1±0.8	73.7±3.2	87.1±3.1	71.8±1.6	51.0±4.8	56.1±1.3	58.8±0.9	57.9±10.9	62.2
ICL [Shenkar and Wolf, 2022]	65.7±1.2	66.7±1.7	80.3±1.4	94.1±3.3	77.3±0.5	56.0±2.0	76.9±0.7	58.5±1.4	58.3±1.5	70.4
OptIForest[Xiang *et al.*, 2023]	69.6±1.2	**68.4±1.1**	78.9±1.3	95.2±0.5	**78.0±0.3**	61.7±3.4	78.8±0.8	61.4±0.7	63.4±1.0	72.8
Gaussian [Verma *et al.*, 2021]	64.6±3.2	65.4±1.3	72.9±2.1	91.7±0.9	76.0±2.7	63.9±3.8	74.8±1.7	60.7±1.8	**69.7±0.7**	71.1
Crop [Xie *et al.*, 2022]	65.8±1.9	64.5±1.7	77.4±4.6	93.1±2.6	73.9±1.8	**66.8±3.2**	74.5±2.1	61.7±3.2	63.5±2.5	71.2
Mixup [Verma *et al.*, 2021]	67.0±2.0	65.5±1.3	72.8±2.2	92.6±4.8	75.7±3.5	63.4±1.2	77.1±3.6	60.1±2.2	68.4±1.1	71.4
Vime [Yoon *et al.*, 2020]	68.5±2.7	64.6±1.9	75.5±4.4	93.3±0.8	76.3±1.6	64.2±4.6	77.8±4.8	62.2±2.6	66.4±1.6	72.1
Random Mask [Xie *et al.*, 2022]	67.4±2.8	67.0±1.7	74.1±3.3	93.1±2.4	76.8±2.1	63.8±3.2	78.5±0.8	62.3±1.7	67.6±1.3	72.3
SemanticMask (ours)	**70.4±1.0**	67.5±1.1	**81.4±1.5**	95.0±0.7	77.6±1.0	64.7±2.0	**80.2±1.9**	**62.6±1.3**	**69.7±1.1**	**74.3**
SemanticMask+position (ours)	**70.5±2.1**	**68.0±1.4**	**80.7±1.7**	**96.3±0.9**	**78.1±1.2**	**66.5±1.2**	**80.4±1.5**	**64.0±2.0**	**69.7±0.6**	**74.9**
SemanticMask+description (ours)	**70.9±1.7**	**68.5±1.9**	**81.8±2.5**	**96.3±1.0**	**78.4±1.0**	**66.0±2.0**	**81.7±3.0**	**63.3±2.3**	69.6±0.4	**75.2**

Table 2: Comparison of AUCROC with state-of-the-art unsupervised anomaly detection methods and existing augmentation techniques.

lous feature information is absent. The outcomes explain the necessity of SemanticMask in including information from all clusters within shared information between views when the anomalous features are unknown, thus preventing the omission of cluster-specific information.

6 Experiment

In our experiments, we assess SemanticMask and its variants across various tabular datasets to validate their effectiveness. For anomaly detection, we compare the proposed method with the state-of-the-art anomaly detection methods and other existing augmentation techniques for CL on tabular data. Additionally, we expand the scope of SemanticMask to accommodate multi-class novelty detection by integrating the labels of training data into our framework. The source code and appendix are available on GitHub at *https://github.com/TST826/SemanticMask*.

6.1 Experiment Setup

Datasets. Aligned with the definition of anomaly detection, we refer to the class observed during training as "normal" data, treating samples from other classes as "anomalies". We conduct experiments on nine datasets with column names sourced from the Outlier Detection DataSets (ODDS) [Rayana, 2016], the KEEL datasets [Derrac *et al.*, 2015] and the UCI datasets [Markelle *et al.*, 2013]. We train our method on a random selected 50% subset of the normal data. The validation set, consisting of 25% normal data, is used to determine the threshold. The methods are then tested on the remaining normal data and all anomalous samples.

Hyperparameter setups and implementation details. For SemanticMask and its variants, λ is set to 0.5, p_m is selected from the set $\{0.4, 0.5, 0.6\}$. For SemanticMask+description, ϵ is set to 0.1. We set k of k-means proportionally to the feature dimension d. For $d < 18$, $k = 2$. For $18 \leq d < 100$, $k = 3$. For complex datasets such as Arrhythmia [Rayana, 2016], where $d \geq 100$, $k = \lfloor d/100 \rfloor + 3$, features are partitioned into k clusters, forming two disjoint subsets with $\lfloor k/2 \rfloor$ clusters each. Contrastive loss uses a constant temperature τ of 0.01. The threshold for identifying anomalies is determined by the 85th quantiles of the Mahalanobis distance in the validation set. The encoder is a multilayer perceptron consisting of two hidden layers with 128 and 64 hidden units, along with the ReLU activation layer. The encoder is trained using the Adam optimizer with a learning rate of 0.001 and default values for other hyperparameters. Each experiment is repeated 5 times, and the mean and standard deviation (σ) of the results are reported.

Evaluation protocol and baselines. We evaluate the performance of all anomaly detection methods using two commonly used metrics [Han *et al.*, 2022]: Area Under the Receiver Operating Characteristic Curve (AUCROC) and Area Under Precision-Recall Curve (AUCPR). We employ the following unsupervised anomaly detection methods for comparison: One-Class Support Vector Machines (OCSVM) [Schölkopf *et al.*, 2001], Local Outlier Factor (LOF) [Breunig *et al.*, 2000], COPOD [Li *et al.*, 2020], ECOD [Li *et al.*, 2022], Feature Bagging (FB) [Lazarevic and Kumar, 2005], DeepSVDD [Ruff *et al.*, 2018], internal contrastive learning (ICL) [Shenkar and Wolf, 2022] and OptIForest [Xiang *et al.*, 2023]. We also evaluate five baseline data augmentation techniques. For Gaussian Noise [Verma *et al.*, 2021], the mean is set to 0 and the standard deviation is selected from the set $\{0.001, 0.005, 0.01\}$. For Random Crop [Xie *et al.*, 2022], the crop proportion is selected from the set $\{0.6, 0.75, 0.9\}$. For Mixup [Verma *et al.*, 2021], the linear mixing ratio is selected from the set $\{0.5, 0.7, 0.9\}$. For Random Mask [Xie *et al.*, 2022] and Vime [Yoon *et al.*, 2020], the mask proportion is selected from the set $\{0.2, 0.3, 0.4\}$.

6.2 Performance Analysis

Results. In table 2, "SemanticMask" here denotes the framework proposed without the mask estimation layer. "SemanticMask+position" refers to the framework with the mask estimation layer. "SemanticMask+description" denotes the SemanticMask framework enhanced with a sentence-based description incorporating prior knowledge of anomaly detection tasks. The results demonstrate that SemanticMask and its variants outperform state-of-the-art unsupervised baseline methods and common data augmentation techniques for tab-

	AUCROC (%)						AUCPR (%)					
Datasets	Gaussian	Crop	Mixup	Vime	Mask	SemanticMask (ours)	Gaussian	Crop	Mixup	Vime	Mask	SemanticMask (ours)
Wine	86.1±3.9	85.6±4.9	88.6±4.4	87.5±4.3	89.4±1.2	**91.4±2.2**	87.3±4.6	88.3±5.3	90.3±4.4	87.8±5.7	91.8±1.7	**93.6±1.4**
Dermatology	91.5±5.2	81.8±4.9	83.0±5.7	89.2±4.6	86.8±2.8	**94.5±4.1**	94.3±3.2	88.9±5.2	91.1±2.4	93.8±2.5	93.4±1.1	**96.8±2.5**
Ecoli	75.9±5.0	78.5±1.9	78.6±4.5	79.2±5.2	78.2±7.9	**82.0±4.9**	58.2±5.7	64.0±3.5	67.1±6.2	68.7±6.0	68.5±9.7	**71.9±8.0**
Vehicle	79.7±6.6	79.8±3.0	77.1±5.0	77.9±4.4	73.3±2.6	**82.4±5.9**	82.7±4.9	78.5±4.4	77.3±5.0	79.1±3.3	75.1±3.2	**83.5±4.9**
Pageblocks	81.0±2.7	86.9±0.7	82.6±2.1	84.1±2.8	84.1±1.4	**87.1±1.2**	51.2±3.6	54.9±1.2	51.8±3.2	53.2±3.0	54.9±3.9	**55.1±4.0**
Yeast	79.5±1.7	79.6±1.7	78.5±3.2	79.0±2.0	78.7±2.4	**81.9±0.9**	72.8±3.6	71.1±3.8	69.8±4.3	71.0±2.9	72.1±4.2	**77.0±1.2**
RedWine	59.1±1.5	61.6±2.4	60.5±2.1	60.8±2.8	60.7±2.8	**63.0±3.5**	23.8±1.3	**27.8±2.2**	24.0±0.8	25.6±1.3	25.3±1.0	26.3±2.8
HayesRoth	89.7±3.8	93.2±2.3	87.4±3.1	92.4±2.9	90.7±5.1	**94.2±2.9**	89.2±4.6	92.5±3.2	86.6±3.5	91.2±4.9	89.4±5.8	**94.0±3.0**
Cleveland	77.9±3.6	75.4±2.6	77.2±2.2	75.0±1.9	76.3±2.2	**79.0±2.4**	83.9±3.1	79.6±2.6	79.5±3.3	81.0±1.3	81.6±2.4	**85.7±2.0**
Glass	81.0±1.5	80.4±1.4	80.0±2.9	81.1±1.3	81.0±1.2	**82.5±0.9**	87.9±1.5	87.4±1.9	87.3±2.6	87.1±1.1	87.8±0.9	**88.3±1.0**

Table 3: Comparison of AUCROC and AUCPR with existing augmentation techniques in multi-class novelty detection.

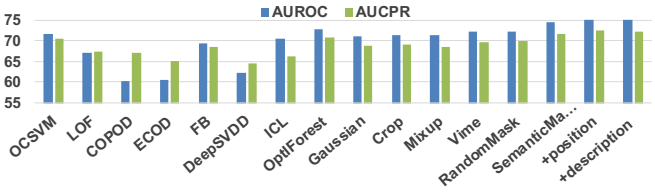


Figure 4: Average anomaly detection performance of all models on AUCROC and AUCPR.

ular data. Notably,"SemanticMask+description" achieves an average AUCROC of 0.752, which is 2.4% higher than the top baseline, OptiForest, and 2.9% higher than the leading existing augmentation technique, random mask. This improvement over the random mask is attributed to SemanticMask's effective design and utilization of semantic information to create views. Figure 4 displays the average values of AUCROC and AUCPR across all datasets, highlighting the superiority of SemanticMask and its variants in both metrics.

Hyperparameter sensitivity analysis. In addition to the analysis performed in Section 5.1 and Figure 5(a) for the hyperparameter p_m, we examine two critical hyperparameters: k and λ. k denotes the number of clusters, and λ controls the balance between the mask estimation loss and the contrastive loss. Experiments are conducted on both the Drug dataset ($d = 12$) and the more intricate Arrhythmia dataset ($d = 274$). Specifically, we explore k values of $\{2, 3, 4, 5, 6, 7, 8\}$ and λ values of $\{0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$. Figure 5(b) indicates that the choice of k impacts the results, with higher dimensions (such as Arrhythmia) benefiting from a larger k to maintain intra-cluster feature correlations. Conversely, in lower dimensions (such as Drug), an excessively large k can lead to sparse clusters, disrupting correlations and resulting in poor performance. Figure 5(c) reveals that SemanticMask is relatively insensitive to λ variations when $0.5 \leq \lambda \leq 0.9$.

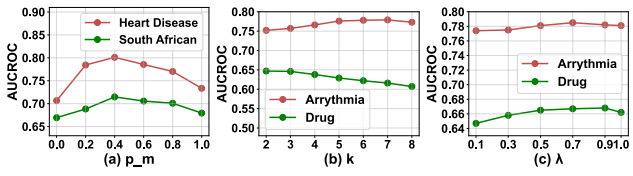


Figure 5: Sensitivity analysis of the hyperparameters p_m, k, λ.

6.3 Extension to Multi-Class Novelty Detection

Setup. The proposed SemanticMask can also be applied to tasks beyond standard anomaly detection, such as multi-class novelty. Unlike anomaly detection with single-class normal samples, multi-class novelty detection aims to distinguish anomalies from normal samples in multi-class scenarios, each characterized by discriminative label information [Yang *et al.*, 2021]. We evaluated our methods on ten multi-class datasets with column names, where specific classes are designated as normal and the rest as anomalies. In multi-class novelty detection, we incorporate label information to further improve the learned representations by using supervised contrastive learning (SCL) loss function [Khosla *et al.*, 2020]. We assessed SemanticMask's performance with the aforementioned baseline augmentation techniques.

Results. SemanticMask demonstrates a significant advantage over other data augmentation techniques, achieving average AUCROC and AUCPR scores of 0.838 and 0.772, respectively. These scores represent increases of 3.2% and 3.4% over the top baseline, Vime. Please see Table 3 for details.

7 Conclusion

In this study, we propose SemanticMask, a data augmentation technique, specifically designed to generate more effective contrastive views for anomaly detection tasks in tabular data. SemanticMask incorporates semantic information from column names, and considers the correlation between features to adjust the shared information between views and learn an effective representation. We establish that to achieve a good representation for anomaly detection, the shared information between views should be appropriately balanced, avoiding both insufficiency and redundancy. Additionally, our empirical results reveal that good views for anomaly detection are feature-dependent. Extensive experimental results demonstrate the superior performance of SemanticMask compared to state-of-the-art anomaly detection methods and prevalent augmentation techniques for tabular data. Note that this study focuses exclusively on clear column names. In the presence of unclear or typo-laden column names, the performance of SemanticMask might deteriorate. To remedy the lack of semantic clarity, large language models like ChatGPT-4 [OpenAI, 2023] can be used for automatic standardization and column name interpretation, a task we leave as future work.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No.62276230 and No.62306277), Zhejiang Provincial Natural Science Foundation of China (LDT23F02023F02).

References

[Bahri *et al.*, 2021] Dara Bahri, Heinrich Jiang, Yi Tay, and Donald Metzler. Scarf: Self-supervised contrastive learning using random feature corruption. In *Proceedings of the International Conference on Learning Representations*, 2021.

[Breunig *et al.*, 2000] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

[Chandola *et al.*, 2009] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607, 2020.

[Cho *et al.*, 2021] Hyunsoo Cho, Jinseok Seol, and Sang-goo Lee. Masked contrastive learning for anomaly detection. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 1434–1441, 2021.

[Derrac *et al.*, 2015] J Derrac, S Garcia, L Sanchez, and F Herrera. Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic & Soft Computing*, 17, 2015.

[Eskin, 2000] Eleazar Eskin. Anomaly detection over noisy data using learned probability distributions. In *International Conference on Machine Learning*, pages 255–262, 2000.

[Golan and El-Yaniv, 2018] Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. In *Advances in Neural Information Processing Systems*, volume 31, pages 9781–9791, 2018.

[Han *et al.*, 2022] Songqiao Han, Xiyang Hu, Hailiang Huang, Minqi Jiang, and Yue Zhao. Adbench: Anomaly detection benchmark. In *Advances in Neural Information Processing Systems*, volume 35, pages 32142–32159, 2022.

[Hartigan *et al.*, 1979] John A Hartigan, Manchek A Wong, et al. A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9726–9735, 2020.

[Khosla *et al.*, 2020] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673, 2020.

[Lazarevic and Kumar, 2005] Aleksandar Lazarevic and Vipin Kumar. Feature bagging for outlier detection. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 157–166, 2005.

[Li *et al.*, 2020] Zheng Li, Yue Zhao, Nicola Botta, Cezar Ionescu, and Xiyang Hu. Copod: copula-based outlier detection. In *2020 IEEE International Conference on Data Mining*, pages 1118–1123, 2020.

[Li *et al.*, 2021] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9664–9674, 2021.

[Li *et al.*, 2022] Zheng Li, Yue Zhao, Xiyang Hu, Nicola Botta, Cezar Ionescu, and George Chen. Ecod: Unsupervised outlier detection using empirical cumulative distribution functions. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–13, 2022.

[Markelle *et al.*, 2013] Kelly Markelle, Longjohn Rachel, and Nottingham Kolby. The uci machine learning repository. https://archive.ics.uci.edu, 2013. Accessed: 2023-6-28.

[Nguyen *et al.*, 2019] Duc Tam Nguyen, Zhongyu Lou, Michael Klar, and Thomas Brox. Anomaly detection with multiple-hypotheses predictions. In *International Conference on Machine Learning*, volume 35, pages 4800–4809, 2019.

[Oord *et al.*, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018.

[OpenAI, 2023] OpenAI. Gpt-4 technical report, 2023.

[Peng *et al.*, 2022] Xiangyu Peng, Kai Wang, Zheng Zhu, Mang Wang, and Yang You. Crafting better contrastive views for siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16010–16019, 2022.

[Pidhorskyi *et al.*, 2018] Stanislav Pidhorskyi, Ranya Almohsen, and Gianfranco Doretto. Generative probabilistic novelty detection with adversarial autoencoders. In *Advances in Neural Information Processing Systems*, volume 31, pages 6823–6834, 2018.

[Rayana, 2016] Shebuti Rayana. Odds library. http://odds.cs.stonybrook.edu, 2016. Accessed: 2023-3-28.

[Reimers and Gurevych, 2019] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the Conference*

on Empirical Methods in Natural Language Processing, pages 3980–3990, 2019.

[Reiss *et al.*, 2021] Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. Panda: Adapting pretrained features for anomaly detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2806–2814, 2021.

[Ruff *et al.*, 2018] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International Conference on Machine Learning*, pages 4390–4399, 2018.

[Ruff *et al.*, 2021] Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021.

[Schölkopf *et al.*, 1999] Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. In *Advances in Neural Information Processing Systems*, volume 12, pages 582–588, 1999.

[Schölkopf *et al.*, 2001] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

[Sehwag *et al.*, 2021] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. In *Proceedings of the International Conference on Learning Representations*, 2021.

[Shenkar and Wolf, 2022] Tom Shenkar and Lior Wolf. Anomaly detection for tabular data with internal contrastive learning. In *Proceedings of the International Conference on Learning Representations*, 2022.

[Sohn *et al.*, 2021] Kihyuk Sohn, Chun-Liang Li, Jinsung Yoon, Minho Jin, and Tomas Pfister. Learning and evaluating representations for deep one-class classification. In *Proceedings of the International Conference on Learning Representations*, 2021.

[Suhara *et al.*, 2022] Yoshihiko Suhara, Jinfeng Li, Yuliang Li, Dan Zhang, Çağatay Demiralp, Chen Chen, and Wang-Chiew Tan. Annotating columns with pre-trained language models. In *Proceedings of the 2022 ACM SIGMOD international conference on Management of Data*, pages 1493–1503, 2022.

[Tack *et al.*, 2020] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *Advances in Neural Information Processing Systems*, volume 33, pages 11839–11852, 2020.

[Tian *et al.*, 2020a] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision*, pages 776–794, 2020.

[Tian *et al.*, 2020b] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems*, volume 33, pages 6827–6839, 2020.

[Verma *et al.*, 2021] Vikas Verma, Thang Luong, Kenji Kawaguchi, Hieu Pham, and Quoc Le. Towards domain-agnostic contrastive learning. In *International Conference on Machine Learning*, pages 10530–10541, 2021.

[Wang and Cherian, 2019] Jue Wang and Anoop Cherian. Gods: Generalized one-class discriminative subspaces for anomaly detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8200–8210, 2019.

[Wang and Sun, 2022] Zifeng Wang and Jimeng Sun. Transtab: Learning transferable tabular transformers across tables. In *Advances in Neural Information Processing Systems*, volume 35, pages 2902–2915, 2022.

[Wang *et al.*, 2022] Haoqing Wang, Xun Guo, Zhi-Hong Deng, and Yan Lu. Rethinking minimal sufficient representation in contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16020–16029, 2022.

[Wen and Li, 2021] Zixin Wen and Yuanzhi Li. Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*, pages 11112–11122, 2021.

[Xiang *et al.*, 2023] Haolong Xiang, Xuyun Zhang, Hongsheng Hu, Lianyong Qi, Wanchun Dou, Mark Dras, Amin Beheshti, and Xiaolong Xu. Optiforest: Optimal isolation forest for anomaly detection. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 2379–2387, 2023.

[Xie *et al.*, 2022] Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Jiandong Zhang, Bolin Ding, and Bin Cui. Contrastive learning for sequential recommendation. In *International Conference on Data Engineering*, pages 1259–1273, 2022.

[Yan *et al.*, 2021] Xudong Yan, Huaidong Zhang, Xuemiao Xu, Xiaowei Hu, and Pheng-Ann Heng. Learning semantic context from normal samples for unsupervised anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3110–3118, 2021.

[Yang *et al.*, 2021] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey, 2021.

[Yoon *et al.*, 2020] Jinsung Yoon, Yao Zhang, James Jordon, and Mihaela van der Schaar. Vime: Extending the success of self-and semi-supervised learning to tabular domain. In *Advances in Neural Information Processing Systems*, volume 33, pages 11033–11043, 2020.

[Zhai *et al.*, 2016] Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang. Deep structured energy based models for anomaly detection. In *International Conference on Machine Learning*, pages 1100–1109, 2016.