Tab-semiSL: Tabular Data-Driven Semi-Supervised Learning to Identify Factors Associated with Immune-Related Adverse Events

Ruhao Liu[1] , **Suixue Wang**[1] , **Hang Yu**[2] , **Peng Li**[3 *] and **Qingchen Zhang**[4 *]
[1]School of Information and Communication Engineering, Hainan University, Haikou, China
[2]School of Engineering, University of Warwick, Coventry, CV4 7AL, UK
[3]School of Computer Science and Technology, Dalian University of Technology Key Laboratory of Social Computing and Cognitive Intelligence, Dalian University of Technology
[4]School of Computer Science and Technology, Hainan University, Haikou, China
{lrh, wangsuixue, zhangqingchen}@hainanu.edu.cn, Hang.Yu.1@warwick.ac.uk, pli@dlut.edu.cn

Abstract

Immune Checkpoint Inhibitors (ICIs) have become a major therapeutic strategy in cancer treatment. However, widespread ICIs use can cause mild-to-severe immune-related Adverse Events (irAEs). Identifying factors associated with irAEs is beneficial for assessing the risk of irAEs occurrence during ICIs treatment. Nevertheless, the lack of sufficient irAEs-related clinical tabular data has slowed progress in this field. Thus, we proposed Tabular data-driven semi-Supervised Learning (Tab-semiSL), a novel Semi-Supervised Learning (SSL) method specifically designed for tabular data. Tab-semiSL can be divided into the pre-training phase and SSL phase. Specifically, in the pre-training, we used the collected dataset of 4,817 cancer patients without the irAEs label to train an auto-encoder. In the SSL phase, the predictor was fine-tuned using the collected dataset of 237 irAEs labeled cancer patients who treated with ICIs. We assigned class-specific thresholds and used samples as pseudo-labels only when their maximum confidence exceeded the threshold, with the thresholds dynamically adjusted during training to improve reliability. We tested Tab-semiSL against three traditional machine learning approaches and five semi-supervised or self-supervised methods using two public datasets plus the collected clinical data, achieving outperformed results across most metrics. SHapley Additive exPlanations (SHAP) analysis revealed ten key factors linked to irAEs. Code and Supplementary Material are available at https://github.com/RuhaoLiu/Tab-semiSL

1 Introduction

The high incidence and mortality rates of cancer threaten people all over the world. Although current treatment modalities include surgery, chemotherapy, radiotherapy, targeted therapy, poor prognosis remains a challenge [Haefner *et al.*, 2015; Smyth *et al.*, 2020]. While various therapies are currently available in clinical practice, immunotherapy, particularly Immune Checkpoint Inhibitors (ICIs), is regarded as one of the most promising approaches [Lee *et al.*, 2016]. ICIs restore anti-tumor T-cell activity by blocking inhibitory pathways such as Programmed cell Death protein 1 (PD-1)/Programmed cell Death Ligand 1 (PD-L1) and Cytotoxic T Lymphocyte-associated Antigen-4 (CTLA-4), thereby exerting therapeutic effects [Postow *et al.*, 2018]. However, studies have shown that the use of ICIs can introduce autoimmune inflammation in multiple organs, collectively referred to as immune-related Adverse Events (irAEs) [Ramos-Casals *et al.*, 2020]. These irAEs significantly affect patient prognosis and survival rates. Therefore, there is an urgent need to establish a reliable and clinically applicable predictive method for irAEs identification. Current predictive methods can be classified into two main categories: statistical analysis methods which usually applied meta-analysis, logistic regression, etc. [Jing *et al.*, 2021; Fujisawa *et al.*, 2017; Wang *et al.*, 2024], and machine learning approaches, e.g., Support Vector Machines (SVM), Random Forests (RF), eXtreme Gradient Boosting (XGBoost), etc. These machine learning models are often combined with interpretability algorithms for feature importance ranking and biomarker identification [Gong *et al.*, 2023; Zhou *et al.*, 2022]. However, these methods are based on retrospective hospital collected data, which is often small and prone to overfitting. Although some researchers have utilized the public datasets FDA Adverse Event Reporting System (FAERS) to identify irAEs biomarkers [Bomze *et al.*, 2019; Jing *et al.*, 2020]. However, FAERS is a spontaneous reporting system with a large amount of missing data [Sakaeda *et al.*, 2013], where critical clinical information such as cancer severity, laboratory and radiological findings is unavailable [Raschi *et al.*, 2019]. Recently, [Wang *et al.*, 2023] achieved significant improvements using SSL with a small amount of labeled CT data and a large amount of unlabeled data. The success of this study indicates that SSL has potential in medicine.

While SSL has achieved notable success with imaging data, adapting it to tabular data (e.g., lab values, gene expression) in irAEs research faces challenges. Most SSL methods

*Corresponding author

involve the process of data perturbation, such as image rotation [Sohn *et al.*, 2020] or two images and their convex combination(s) [Zhang *et al.*, 2017] and these perturbation methods are not suitable for application to tabular data [Yoon *et al.*, 2020]. VIME [Yoon *et al.*, 2020] is the first SSL method specifically designed for tabular data. These SSL methods applied to tabular data mainly have two problems: 1) The data perturbation is not effective enough. 2) Lack of sample selection mechanisms, where all unlabeled samples are used throughout training regardless of prediction confidence, allowing noisy and low-confidence unlabeled samples to degrade model performance.

In addition, several studies have extended self-supervised learning to tabular data (SubTab [Ucar *et al.*, 2021] and SCARF [Bahri *et al.*, 2021]). Unlike SSL, self-supervised learning first leverages unlabeled data for pre-training, followed by fine-tuning exclusively on labeled data. This approach, however, has some limitations, e.g., the absence of explicit supervision limits the model's ability to obtain task-specific representations that are optimal for downstream tasks. When only a limited number of labeled data is available for downstream task fine-tuning, the resulting improvements in model performance tend to be marginal [Li *et al.*, 2021].

In this study, we retrospectively collected the clinical laboratory tabular data of 237 cancer patients treated with ICIs and 4,817 unlabeled cancer patients without the irAEs label, as illustrated in Figure 1. We introduced Tab-semiSL, a novel SSL method specifically designed for tabular data. Tab-semiSL can be classified into two stages: In the pre-training stage, to enhance the effectiveness of perturbation for unlabeled samples, we used Variational Auto-Encoder (VAE) [Kingma *et al.*, 2013] combined with marginal distribution sampling to perturb unlabeled samples, then using original unlabeled samples and perturbed samples jointly trained the auto-encoder with masked reconstruction and contrastive learning. In the SSL stage, labeled samples undergo standard supervised learning. For unlabeled samples, we enforce the results of the perturbed samples to be consistent with the pseudo-labels, only the class with the maximum predicted probability was selected as the true labels [Lee and others, 2013]. Considering that the reliability of pseudo-labels generated for unlabeled samples fluctuates during the SSL training process, we set a specific threshold for each class and only regarded samples whose maximum confidence belongs to a class and is higher than the threshold of class-specific as pseudo-labels. Then, we proposed a self-adaptive dynamic threshold update mechanism to update the threshold according to different stages of SSL training. Once the model was obtained, we used Area Under the Curve (AUC), accuracy, recall and f1 score metrics for evaluation and SHapley Additive exPlanations (SHAP) values [Lundberg and Lee, 2017] for interpretability analysis and feature ranking, identifying factors associated with irAEs. Our contributions can be summarized as follows

- We designed Tab-semiSL, a novel semi-supervised learning method specifically for clinical tabular data processing, which was benchmarked against conventional classification models, SSL and self-supervised learning methods, almost achieving the best performance.
- We introduce a novel perturbation framework for tabular data that combines a VAE with marginal distribution sampling. Furthermore, masked reconstruction and contrastive learning are jointly incorporated in the pre-training stage to improve representation quality.
- We set a specific threshold for each class and proposed a self-adaptive and dynamic threshold update mechanism to update the threshold. Only unlabeled samples whose maximum confidence belongs to a class and is higher than the threshold of class-specific are regarded as pseudo-labels during different stages of SSL training.
- We collected and validated on real-world ICIs-treated patient data, and identified interpretable relevant factors of irAEs.

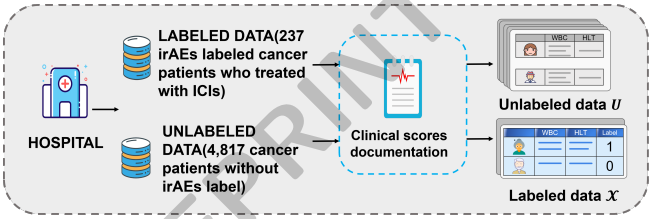

Figure 1: The process of both labeled and unlabeled clinical data collection

2 Related Work

SSL aims to enhance model performance by using a small amount of labeled samples and a large amount of unlabeled samples. Compared with self-supervised learning, SSL combines labeled samples and unlabeled samples during the training process, which can effectively capture the relevant information of both [Yang *et al.*, 2022]. SSL can be roughly divided into three methods: The first is consistency regularization [Samuli and Timo, 2017; Tarvainen and Valpola, 2017; Xie *et al.*, 2020]. The second is self-training [Tai *et al.*, 2021; Rizve *et al.*, 2021]. The third combines both approaches. Representative methods include FixMatch [Sohn *et al.*, 2020], FlexMatch [Zhang *et al.*, 2021], and FreeMatch [Wang *et al.*, 2022]. Most mainstream methods rely on strong and weak perturbations to the same sample: pseudo-labels are generated from the weakly perturbed version, and the strongly perturbed version is used to compute the loss based on those pseudo-labels. Initially, the majority of SSL research was concentrated in the domain of image data. VIME is the first method to apply SSL to tabular data. VIME constructs 'pretext tasks' to train the unlabeled samples, simultaneously generating perturbed samples by sampling from empirical marginal distributions. Then, consistency regularization is used to encourage the output consistency of different perturbed samples.

3 Methodology

3.1 Preliminaries

We defined an SSL scenario for tabular datasets with both labeled and unlabeled samples.Let $\mathcal{X} = \{(x_i, y_i)\}_{i=1}^{N}$ rep-

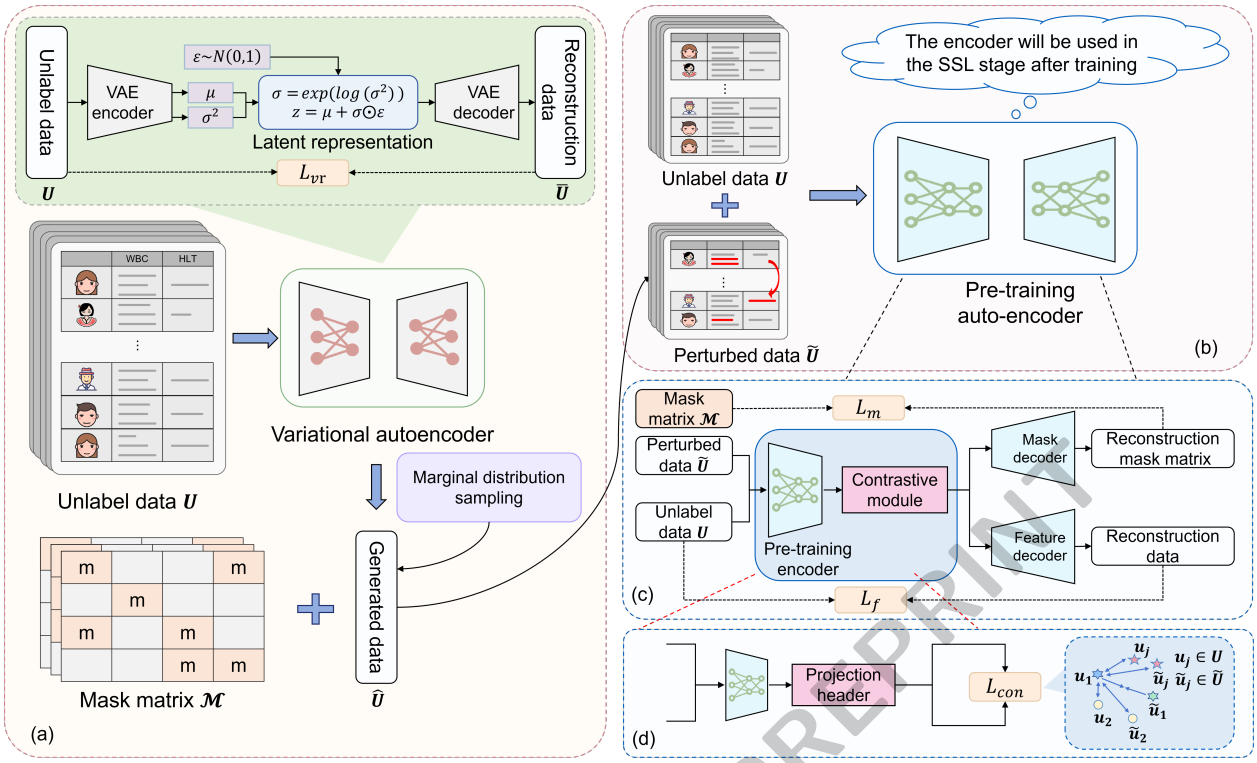

Figure 2: The pre-training phase of Tab-semiSL consists of training a VAE model and a pre-training auto-encoder model. (a) First, the VAE model is trained using unlabeled data, and then data generation is carried out. (b) (c) During the auto-encoder pre-training process, the model simultaneously processes two types of inputs: perturbed data and the original unlabeled data. (d) Contrastive learning module.

resents a label dataset of size N, where $x_i \in \mathbb{R}^K$ and $y_i \in \{1, \cdots, C\}$ denote labeled training samples with feature dimension of K and labels with C categories, respectively. $U = \{u_j\}_{j=N+1}^{N+M}$ denotes an unlabeled dataset of size M, where $u_j \in \mathbb{R}^K$. Generally, $N << M$. Let $\mathcal{Y} = \{y_1, \cdots, y_N\}$, the task objective of SSL is to train mapping relationships $\mathcal{X} \rightarrow \mathcal{Y}$ using both $\mathcal{D}_{lab}$ and $\mathcal{D}_{unlab}$ simultaneously.

3.2 Pre-Training Stage of Tab-SemiSL

Samples Generation using VAE

In the pre-training phase (see Figure 2), to achieve data perturbation, we first train a VAE using unlabeled samples, and then samples were generated by probabilistically modeling the latent space and sampling from it. Specifically, the encoder of VAE ψ_ϕ encodes the raw samples U into posterior distributions $q_\phi(z|U)$ of the latent representation z and utilizes Kullback-Leibler(KL) divergence to construct the loss function $L_{KL} = KL\left[q_\phi(z \mid U) \| p(z)\right]$ to constrain the posterior distribution, encouraging it to approximate a standard normal distribution, here ϕ denotes the encoder parameters of VAE, $p(z) = \mathcal{N}(0, I)$. Latent representations $z = \mu(U) + \sigma(U) \odot \varepsilon$ are sampled via the reparameterization trick, here $\mu(U)$ and $\sigma(U)$ represent the mean and standard deviation of U, respectively, $\varepsilon \sim \mathcal{N}(0, I)$, $\odot$ denotes element-wise product. Moreover, the VAE also encourages the samples generated by the decoder to be similar to the original samples, with the loss function formulated as:

$$L_{vr} = -E_{q_\phi(z|U)}\left[p_\theta(\bar{U} \mid z)\right], \tag{1}$$

where $p_\theta(\bar{U})$ represents the output distribution of the decoder, represents the reconstructed samples, θ denotes the VAE decoder parameter. Therefore, the loss function of VAE can be expressed as:

$$L_{vae} = L_{KL} + L_{vr}. \tag{2}$$

Thus, the generated samples was represented as $\hat{U} = \{\hat{u}_j\}_{j=N+1}^{N+M}$, where $\hat{u}_j$ can be expressed as:

$$\widehat{u}_j = \varphi_\theta\left(\mu_j + \sigma_j \odot \varepsilon_j\right), \varepsilon \sim \mathcal{N}(0, I), \tag{3}$$

where $\varphi_\theta(\cdot)$ represents the decoder of VAE with sigmoid activation, μ_j, $\log\, \sigma_j^2 = \psi_\phi(u_j)$. See Supplementary Material Section 2 for more details.

Samples Perturbation

Samples perturbation were achieved by stochastically replacing feature values of generated samples $\hat{U}$ through sampling from their marginal distributions. More precisely, we generated a mask matrix $\mathcal{M} \in \{0, 1\}^{M \times K}$, here $\mathcal{M}_{j,k} \sim Bernoulli(p_m)$, $p_m \in [0, 1]$ denotes a hyperparameter for setting the masked rate. If $\mathcal{M}_{j,k} = 1$ indicates that the value is retained, and $\mathcal{M}_{j,k} = 0$ indicates that the value is masked. Then for the perturbed samples $\tilde{U} = \{\tilde{u}_j\}_{j=N+1}^{N+M}$, the $j-th$ row and the $d-th$ column can be expressed as:

$$\widetilde{u}_{j,d} = \mathcal{M}_{j,d} \odot \widetilde{u}_{j,d} + (1 - \mathcal{M}_{j,d}) \odot s_{j,d}, \tag{4}$$

here $s \sim \{\hat{u}_{t,d} \mid t \in \{N+1, \cdots, M\} \backslash \{j\}\}$.

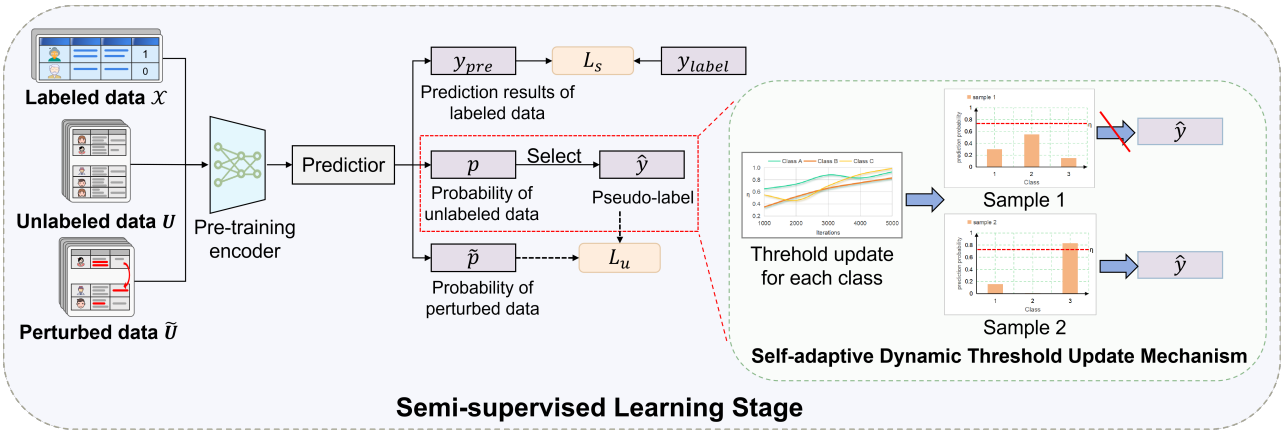

Figure 3: Semi-supervised learning pipeline of Tab-semiSL. The labeled data are used for supervised learning, while the unlabeled data undergo pseudo-label selection under the self-adaptive dynamic threshold update mechanism for unsupervised learning.

Training the Model of Pre-Training Auto-Encoder

The objective is to obtain a pre-training encoder for feature representation during the pre-training stage. We define the pre-training encoder and decoder as $f_\delta(\cdot)$ and $f_\Delta(\cdot)$, respectively. The pre-training auto-encoder used the original unlabeled samples U and their perturbed samples $\tilde{U}$ as inputs, integrating the collaborative training strategy of masked reconstruction and contrastive learning, where masked reconstruction needs to complete two 'pretext tasks' - mask matrix reconstruction and feature reconstruction. The loss in the pre-training stage is given by the sum of mask matrix reconstruction loss L_m, feature reconstruction loss L_f and contrastive loss L_{con}:

$$L_{pre} = L_f + \alpha L_m + \beta L_{con}, \tag{5}$$

where L_f and L_m both computed via mean squared error and binary cross-entropy, respectively, α and β are hyperparameters. When the feature is categorical, the mean squared error is replaced by the cross entropy loss. Further details can be found in Supplementary Material Section 3.

For an unlabeled sample $u_j \in \mathbb{R}^K$, we treat its perturbed sample $\tilde{u}_j$ as a positive pair and all other samples u_t and $\tilde{u}_t$, $t \neq j$ as negative pairs to construct the InfoNCE contrastive loss [Oord *et al.*, 2018]. The objective is to encourage the latent representations of positive pairs to be close to each other, while pushing those of negative pairs farther apart. Thus, contrastive loss L_{con} is defined as follows:

$$L_{\text{con}} = -\frac{1}{M}\sum_{j=N+1}^{N+M}\log\frac{\exp\left(\text{sim}\left(z_j,\tilde{z}_j\right)/\tau\right)}{\sum_{t=N+1}^{N+M}\exp\left(\text{sim}\left(z_j,z_t\right)/\tau\right)}, \tag{6}$$

where $z_j = g(f_\delta(u_j))$, $\tilde{z}_j = g(f_\delta(\tilde{u}_j))$, $z_t = g(f_\delta(u_t))$, $\text{sim}\left(z_j,\tilde{z}_j\right) = \frac{z_j^T\tilde{z}_j}{\|z_j\|_2\cdot\|\tilde{z}_j\|_2}$, τ denotes temperature parameter, $g(\cdot)$ is a projection head of two layers of neural networks, and activation functions are applied after all network layers.

3.3 SSL Stage of Tab-SemiSL

Training Predictor Model

In the SSL stage (see Figure 3), labeled samples $\mathcal{X}$, unlabeled samples U and unlabeled perturbed samples $\tilde{U}$ will be encoded by the encoder $f_\delta(\cdot)$ which retained from the pre-training stage and then input into a predictor $\mathcal{P}(\cdot)$ composed of two neural networks for training, where labeled samples undergo standard supervised learning and the unlabeled samples are trained via unsupervised learning. The objective loss in the semi-supervised stage is:

$$L_{semi} = L_s + \gamma L_u, \tag{7}$$

where γ is the hyperparameter for balancing supervised loss and unsupervised loss. The supervised loss L_s for the multi-class case is given as follows:

$$L_s = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\cdot\log\left(\mathcal{P}_c\left(\hat{z}_i\right)\right), \tag{8}$$

where $\hat{z}_i = f_\delta(x_i)$, N represents the number of label samples, and C represents the number of categories. When the task is binary classification, the binary cross entropy loss will be adopted or the mean square error loss will be used when it is a regression task.

Self-Adaptive Dynamic Threshold Update Mechanism

The objective of unsupervised loss is to keep the pseudo-labels of unlabeled samples U output consistent with the output of its perturbed samples $\tilde{U}$. However, in the early stage of training, the poor performance of the model led to low-quality pseudo-labels generated from the low-confidence unlabeled samples. We set a specific threshold for each class to select samples with the maximum confidence higher than threshold and have proposed a self-adaptive dynamic threshold update mechanism to update the specific threshold for each class. Unlike FreeMatch, which estimates thresholds from the global prediction distribution with class fairness adjustment, our method computes class-specific thresholds based on the confidence statistics of progressively selected high-confidence samples. In addition, we introduce an iteration-dependent sample selection strategy and EMA-based threshold smoothing to stabilize the pseudo-label generation process. Our method for dynamically calculating the threshold is as follows: Based on the number of iterations, selected

the high-confidence samples, and simultaneously calculate the average value of the class corresponding to the maximum confidence of selected samples as the threshold for this class (see Supplementary Material Algorithm 1). Specifically, for unlabeled samples $U = \{u_j\}_{j=N+1}^{N+M}$, its output confidence is denoted as $p_j = \mathcal{P}(f_\delta(u_j))$ and $p_j = \{p_{j,1}, \cdots, p_{j,C}\}$, $p_{j,c} \geq 0$, $\forall c \in \{1, \cdots, C\}$, $\sum_{c=1}^{C} p_{j,c} = 1$. The sample selected at the $t-th$ iteration can be represented as:

$$D_s = \mathcal{F}\left(\text{sort}\left(\{u_{N+1}, \cdots, u_{N+M}\}\right), t/T\right), \tag{9}$$

where T represents the total number of iterations sort denotes ordering the elements in the set in descending order based on the maximum predicted confidence of each sample, such that $p_{N+1}^{\max} \geq \cdots \geq p_{N+M}^{\max}$, $p_{N+1}^{\max} = \max\{p_{N+1,1}, \cdots, p_{N+1,C}\}$ represents the maximum predicted confidence for sample u_{N+1}. The role of $\mathcal{F}(\cdot)$ is to select the top-t/T samples from the sorted set. Then, calculate the average value based on the class corresponding to the maximum confidence value of each sample in D_s, which is regarded as the threshold of each class. Meanwhile, the threshold for each iteration is maintained by a dynamically updated Exponential Moving Average (EMA). Let the initial thresholds all be $\eta_0 = \infty$, and the threshold for the $k-th$ class in the $t-th$ iteration ($t > 0$) can be expressed as:

$$\eta_t(k) = \frac{\sum_{s=N+1}^{N+|D_s|} \max_{c \in \{1,\cdots,C\}} (p_{s,c}) \cdot \mathbb{1}\left(\arg\max_c(p_{s,c}) = k\right)}{\sum_{s=N+1}^{N+|D_s|} \mathbb{1}\left(\arg\max_c(p_{s,c}) = k\right)}. \tag{10}$$

When $\eta_{t-1}(k) \neq \infty$, we employ EMA to maintain the threshold, which can be expressed as:

$$\eta_t(k) = \lambda \eta_{t-1}(k) + (1-\lambda)\eta_t(k), \tag{11}$$

where $\mathbb{1}(\cdot)$ denotes Indication function, $p_s = \mathcal{P}(f_\delta(u_s))$, $u_s \in D_s$, λ denotes the momentum decay of EMA, and $\lambda = \lambda_0 + (0.99 - \lambda_0)\frac{t}{T}$, $\lambda_0 \in (0, 1)$. Therefore, the unsupervised loss in $t-th$ iterations can be expressed as:

$$L_u = \frac{1}{M} \sum_{j=N+1}^{N+M} \mathbb{1}(\max_c(p_{j,c}) > \eta_t(\hat{y}_j)) \cdot \mathcal{H}(\hat{y}_j, \tilde{p}_j), \tag{12}$$

where $\mathcal{H}$ denotes cross entropy loss, $\hat{y}_j = \arg\max_c(p_{j,c})$ is a pseudo-label, $\tilde{p}_j = \mathcal{P}(f_\delta(\tilde{u}_j))$.

4 Experiments and Results

4.1 Experimental Settings

Datasets

We used two publicly available datasets to evaluate the effectiveness of our method: the Adult Income dataset for binary classification, and the MNIST dataset for multi-class classification. For the public datasets, we select a small portion of the data as labeled samples, while discarding the labels of the remaining samples to treat them as unlabeled samples. Meanwhile, to determine factors associated with irAEs, this study retrospectively collected clinical laboratory test data from 237 cancer patients who had received ICIs treatment and 4,817 cancer patients without irAEs labels from partner hospitals, including 87 characteristics. For cancer patients who had received ICIs treatment (labeled samples), we collect the most recent examination values of the patients before they use ICIs. According to the medical records, the occurrence of irAEs in patients was assessed. Patients (N =58 or 24%) were identified as having experienced irAEs, whereas patients (N=179 or 76%) did not experience irAEs. For cancer patients without irAEs labels (unlabeled samples), we collect the latest examination results of the patients. The advantage of adding a large amount of unlabeled samples lies in alleviating the model overfitting and poor generalization caused by the limited number of labeled samples. Detailed clinical laboratory test indicators can be found in Supplementary Material Section 4. For the collected samples, missing values are filled using the KNN imputer from the scikit-learn package. To address label imbalance, we apply the Synthetic Minority Oversampling Technique (SMOTE) [Chawla *et al.*, 2002] to the labeled data. All datasets are then split into training, validation, and test sets in a 7:1:2 ratio.

Implementation Details

Tab-semiSL consists of the following modules: 1) VAE module. 2) Pre-training auto-encoder module. 3) Predictor module. Our approach was developed in Python utilizing the PyTorch framework, with computations on NVIDIA RTX 4080 GPU and AMD Ryzen 9 7945HX. The epochs for training the VAE module and the pre-training auto-encoder module are 50 rounds and 30 rounds respectively. In the SSL stage, the predictor module was trained for a maximum of 5,000 iterations, with an initial warm-up phase of 50 iterations using labeled data. All learning rates in the pre-training stage are $lr = 1 \times 10^{-3}$. the learning rate of the predictor module trained using the Adult income dataset is $lr = 1 \times 10^{-5}$, and the learning rates trained employing both the MNIST dataset and the private clinical dataset are $lr = 1 \times 10^{-3}$.

Comparative Methods

The contrastive methods used in this study can be divided into two categories: 1) Supervised learning-based methods, including XGBoost [Chen and Guestrin, 2016], CatBoost [Prokhorenkova *et al.*, 2018], and Lightgbm [Ke *et al.*, 2017]. For these methods, we only used labeled samples for training. 2) Semi-supervised and self-supervised learning-based methods, including VIME, SubTab, SCARF, FlexMatch and FreeMatch. For the FlexMatch and FreeMatch methods, we conducted experiments using the same data perturbation strategy adopted in this study. For the self-supervised methods, we first performed pre-training using unlabeled samples, followed by fine-tuning only with labeled samples. All baseline methods were implemented in strict accordance with the experimental settings described in their original papers or by directly employing their official open-source implementations.

Quantitative Results

We evaluate different methods on MNIST and Adult income datasets, reporting the mean ± standard deviation over 10 runs with different random seeds (see Table 1). 'Label Amount'

Method	MNIST dataset				Adult income dataset			
	500	**1000**	**3000**	**6000**	**32**	**64**	**128**	**256**
Xgboost	82.71% (±0.64%)	87.63% (±0.44%)	92.64% (±0.24%)	94.62% (±0.19%)	76.89% (±4.99%)	80.94% (±3.10%)	82.98% (±2.60%)	85.42% (±1.82%)
Catboost	86.96% (±0.79%)	90.20% (±0.39%)	93.48% (±0.19%)	94.76% (±0.16%)	78.90% (±4.35%)	84.61% (±1.41%)	86.36% (±1.31%)	**88.11%** (±**0.74%**)
Lightgbm	82.66% (±1.40%)	87.45% (±0.56%)	92.45% (±0.28%)	94.35% (±0.24%)	50.00% (±0.00%)	78.89% (±3.21%)	83.84% (±1.45%)	85.16% (±1.07%)
VIME	89.26% (±0.78%)	91.57% (±0.49%)	94.14% (±0.38%)	95.01% (±0.23%)	77.15% (±2.68%)	78.99% (±4.83%)	82.60% (±1.61%)	84.15% (±0.73%)
SubTab	92.08% (±0.50%)	93.88% (±0.29%)	95.39% (±0.27%)	96.68% (±0.19%)	79.10% (±4.03%)	83.45% (±2.69%)	84.18% (±1.35%)	85.94% (±0.51%)
SCARF	90.69% (±0.91%)	92.90% (±0.51%)	95.28% (±0.22%)	96.29% (±0.27%)	72.11% (±5.39%)	76.64% (±4.57%)	80.48% (±0.99%)	81.41% (±0.90%)
FlexMatch	91.17% (±0.83%)	93.01% (±0.37%)	94.98% (±0.16%)	96.04% (±0.29%)	78.17% (±4.23%)	83.99% (±4.12%)	84.01% (±2.26%)	85.64% (±1.18%)
FreeMatch	91.83% (±0.66%)	93.44% (±0.47%)	95.02% (±0.32%)	**96.73%** (±**0.18%**)	79.93% (±3.28%)	84.07% (±2.76%)	85.12% (±1.14%)	87.86% (±0.79%)
Tab-semiSL(ours)	**93.86%** (±**0.56%**)	**94.57%** (±**0.19%**)	**95.80%** (±**0.29%**)	96.52% (±0.21%)	**81.07%** (±**3.47%**)	**85.33%** (±**1.01%**)	**86.82%** (±**0.19%**)	87.13% (±0.12%)

Table 1: Comparison of different methods on the MNIST and Adult income datasets. For MNIST, the scores denote accuracy, while for Adult income dataset, the scores denote AUC. Bold indicates the best results, and underlining indicates the second-best results.

refers to the number of labeled samples. Accuracy is used as the metric for MNIST, and AUC for Adult Income. On the MNIST dataset, our method almost achieves the best scores under different numbers of labels. For the Adult Income dataset, Catboost performs best when the number of labeled samples exceeds 256. This phenomenon may be attributed to the relatively simple classification task of this dataset, which enables the model to quickly reach performance saturation even with a limited number of labeled samples. However, as the number of labeled samples decreases, the performance gap between our model and other baseline methods gradually widens, highlighting the effectiveness of our SSL method.

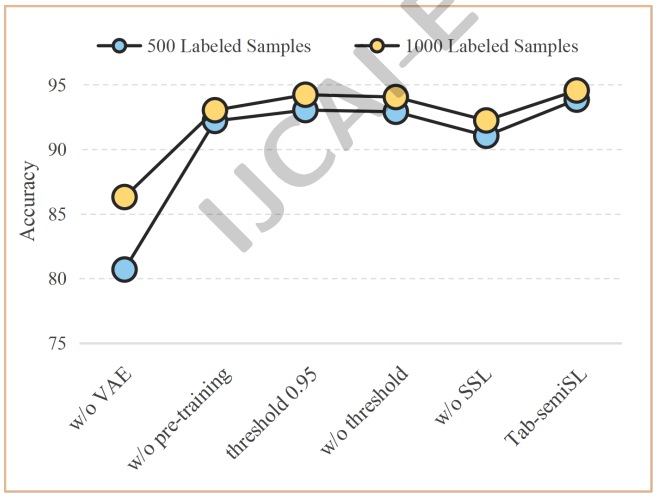

Figure 4: Ablation experiments on Tab-semiSL

Method	AUC	Accuracy	Recall	F1 score
Xgboost	76.92%	72.91%	61.54%	55.17%
Catboost	78.24%	75.00%	69.23%	62.50%
Lightgbm	64.07%	68.75%	53.85%	48.27%
VIME	84.84%	81.25%	69.23%	64.00%
SubTab	74.73%	72.91%	61.54%	58.06%
SCARF	78.24%	75.00%	53.85%	53.85%
FlexMatch	74.73%	68.75%	61.54%	58.06%
FreeMatch	84.84%	75.00%	76.92%	64.00%
Tab-semiSL	**87.69%**	**81.25%**	**84.62%**	**66.67%**

Table 2: The results of the effects of clinical laboratory test dataset on different methods.

4.2 Ablation Study

To evaluate the contribution of each module to the overall performance, we conducted ablation experiments on the MNIST dataset under the same experimental conditions. The key components of Tab-semiSL include the pre-training stage and the SSL stage. We have set up several variants of Tab-semiSL.

Ablation Experiments in the Pre-Training Stage

1) We replaced the VAE-generated samples with samples drawn from a Gaussian distribution, denoted as w/o VAE . 2) We removed pre-training stage, and the raw data are used directly without feature representation learning, denoted as w/o pre-training.

Ablation Experiments in the SSL Stage

1) To verify the effectiveness of our proposed self-adaptive dynamic threshold update mechanism, we conducted experiments using different unlabeled sample selection strategies.

Specifically, we employed two approaches: using a fixed value (0.95) as the experimental threshold or not setting a threshold at all, denoted as threshold 0.95 or w/o threshold. 2) We removed the SSL stage and directly fed the feature representations learned during pre-training into a multilayer perceptron to obtain the final predictions, denoted as w/o SSL.

We conducted ablation experiments with 500 and 1000 labeled samples, and the results are shown in Figure 4.

4.3 Parameter Sensitivity Analysis

We conduct a parameter sensitivity analysis using the MNIST dataset, where 6,000 labeled samples are selected. The purpose is to analysis how different parameter settings influence the model's performance. The hyperparameter ranges are respectively $p_m \in \{0.1, \ldots, 0.9\}$, $\alpha \in \{0.1, 0.5, 0.7, 1, 2, 5\}$, $\beta \in \{0.1, 0.5, 0.7, 1, 2, 5\}$ and $\gamma \in \{0.1, 0.5, 0.7, 1, 2, 5\}$. The analysis results are presented in Supplementary Material Figure 2. Ultimately, $p_m = 0.5$, $\alpha = 2.0$, $\beta = 0.1$ and $\gamma = 0.1$.

4.4 Determine the Factors Associated with irAEs and Analysis

Determine the Factors

We classified the collected clinical laboratory test data using Tab-semiSL and other contrastive methods respectively. Tab-semiSL still achieved the best results (see Table 2). We then used SHAP to rank the feature importance of the pre-training encoder-predictor combined model after training (see Figure 5). SHAP is a game-theoretic interpretability algorithm that determines the marginal contribution of each feature by computing its shapley value [Shapley and others, 1953] when added to a combination of other features. This value fairly attributes the contribution of each feature to the model output. We ranked the features according to shapley values and visualized the top-10 features contributing most to the model output in Supplementary Material Section 5. Ultimately, the 10 important factors associated with irAEs that we identified are: white blood cell count (WBC), the percentage of lymphocytes in the total number of WBC (Lymph), blood platelet count (PLT), basophils as a percentage of total WBC (Baso), free thyroxine (FT4), red blood cell count (RBC), hemoglobin (Hb), potassium ion (K), alpha fetoprotein (AFP), squamous cell carcinoma antigen (SCC).

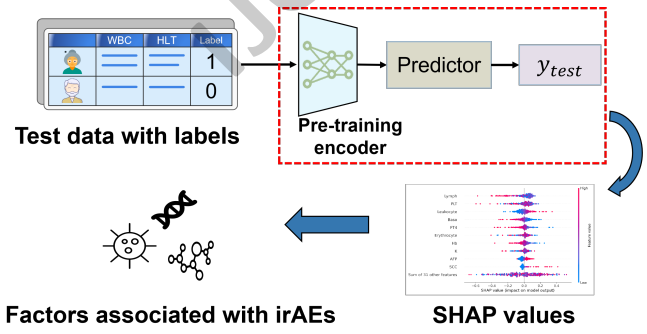

Figure 5: Identify factors associated with irAEs.

Analysis

Lymphocytes, as a subset of leukocytes, mainly include T cells, B cells, and Natural Killer (NK) cells. They play a crucial role in adaptive immunity and are also involved in the systemic inflammatory response [LaRosa and Orange, 2008; Buonacera *et al.*, 2022]. Platelets have been confirmed to be closely associated with inflammation and immune responses and can interact with various cells involved in innate and adaptive immunity. Through the release of chemokines, platelets guide lymphocytes, neutrophils, and monocytes to migrate to sites of inflammation [Sonmez and Sonmez, 2017; Semple *et al.*, 2011]. Additionally, intercellular communication exists between platelets and different types of leukocytes, influencing inflammatory and immune responses [Totani and Evangelista, 2010]. Basophils are considered to induce allergic symptoms by releasing soluble pro-inflammatory mediators and have subsequently been found to be involved in inflammatory and immune responses [Falcone *et al.*, 2011; Harvima *et al.*, 2014]. These lines of evidence suggest that the identified factors associated with irAEs have the potential to serve as predictive biomarkers for irAEs.

5 Conclusion and Future Work

In conclusion, we proposed a novel semi-supervised learning framework entitled 'Tab-semiSL' for clinical tabular data, and demonstrate its ability to predict irAEs by jointly leveraging 237 ICIs-treated patients with labeled irAEs and 4,817 unlabeled oncology cases. The results of Tab-semiSL outperforms conventional supervised, self-supervised, and semi-supervised methods for almost all metrics. Future work will focus on expanding the labeled dataset with additional real-world clinical samples and integrating multimodal inputs, including imaging, genomics, to enhance robustness and uncover further predictive factors for irAEs.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62462022, Grant No. 62572157).

References

[Bahri *et al.*, 2021] Dara Bahri, Heinrich Jiang, Yi Tay, and Donald Metzler. Scarf: Self-supervised contrastive learning using random feature corruption. *ArXiv Preprint ArXiv:2106.15147*, 2021.

[Bomze *et al.*, 2019] David Bomze, Omar Hasan Ali, Andrew Bate, and Lukas Flatz. Association between immune-related adverse events during anti–pd-1 therapy and tumor mutational burden. *JAMA Oncology*, 5(11):1633–1635, 2019.

[Buonacera *et al.*, 2022] Agata Buonacera, Benedetta Stancanelli, Michele Colaci, and Lorenzo Malatino. Neutrophil to lymphocyte ratio: an emerging marker of the relationships between the immune system and diseases. *International Journal of Molecular Sciences*, 23(7):3636, 2022.

[Chawla *et al.*, 2002] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[Chen and Guestrin, 2016] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[Falcone *et al.*, 2011] FH Falcone, EF Knol, and BF Gibbs. The role of basophils in the pathogenesis of allergic disease. *Clinical & Experimental Allergy*, 41(7):939–947, 2011.

[Fujisawa *et al.*, 2017] Yasuhiro Fujisawa, Koji Yoshino, Atsushi Otsuka, Takeru Funakoshi, Taku Fujimura, Yuki Yamamoto, Hiroo Hata, Masahiko Gosho, Ryota Tanaka, Kei Yamaguchi, et al. Fluctuations in routine blood count might signal severe immune-related adverse events in melanoma patients treated with nivolumab. *Journal of Dermatological Science*, 88(2):225–231, 2017.

[Gong *et al.*, 2023] Li Gong, Jun Gong, Xin Sun, Lin Yu, Bin Liao, Xia Chen, and Yong-sheng Li. Identification and prediction of immune checkpoint inhibitors-related pneumonitis by machine learning. *Frontiers in Immunology*, 14:1138489, 2023.

[Haefner *et al.*, 2015] Matthias F Haefner, Kristin Lang, David Krug, Stefan A Koerber, Lorenz Uhlmann, Meinhard Kieser, Juergen Debus, and Florian Sterzing. Prognostic factors, patterns of recurrence and toxicity for patients with esophageal cancer undergoing definitive radiotherapy or chemo-radiotherapy. *Journal of Radiation Research*, 56(4):742–749, 2015.

[Harvima *et al.*, 2014] Ilkka T Harvima, Francesca Levi-Schaffer, Petr Draber, Sheli Friedman, Iva Polakovicova, Bernhard F Gibbs, Ulrich Blank, Gunnar Nilsson, and Marcus Maurer. Molecular targets on mast cells and basophils for novel therapies. *Journal of Allergy and Clinical Immunology*, 134(3):530–544, 2014.

[Jing *et al.*, 2020] Ying Jing, Jin Liu, Youqiong Ye, Lei Pan, Hui Deng, Yushu Wang, Yang Yang, Lixia Diao, Steven H Lin, Gordon B Mills, et al. Multi-omics prediction of immune-related adverse events during checkpoint immunotherapy. *Nature Communications*, 11(1):4946, 2020.

[Jing *et al.*, 2021] Ying Jing, Yongchang Zhang, Jing Wang, Kunyan Li, Xue Chen, Jianfu Heng, Qian Gao, Youqiong Ye, Zhao Zhang, Yaoming Liu, et al. Association between sex and immune-related adverse events during immune checkpoint inhibitor therapy. *JNCI: Journal of the National Cancer Institute*, 113(10):1396–1404, 2021.

[Ke *et al.*, 2017] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 2017.

[Kingma *et al.*, 2013] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

[LaRosa and Orange, 2008] David F LaRosa and Jordan S Orange. 1. lymphocytes. *Journal of Allergy and Clinical Immunology*, 121(2):S364–S369, 2008.

[Lee and others, 2013] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, page 896. Atlanta, 2013.

[Lee *et al.*, 2016] Lucy Lee, Manish Gupta, and Srikumar Sahasranaman. Immune checkpoint inhibitors: An introduction to the next-generation cancer immunotherapy. *The Journal of Clinical Pharmacology*, 56(2):157–169, 2016.

[Li *et al.*, 2021] Junnan Li, Caiming Xiong, and Steven CH Hoi. Comatch: Semi-supervised learning with contrastive graph regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9475–9484, 2021.

[Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

[Oord *et al.*, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv Preprint ArXiv:1807.03748*, 2018.

[Postow *et al.*, 2018] Michael A Postow, Robert Sidlow, and Matthew D Hellmann. Immune-related adverse events associated with immune checkpoint blockade. *New England Journal of Medicine*, 378(2):158–168, 2018.

[Prokhorenkova *et al.*, 2018] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31, 2018.

[Ramos-Casals *et al.*, 2020] Manuel Ramos-Casals, Julie R Brahmer, Margaret K Callahan, Alejandra Flores-Chávez, Niamh Keegan, Munther A Khamashta, Olivier Lambotte, Xavier Mariette, Aleix Prat, and Maria E Suárez-Almazor. Immune-related adverse events of checkpoint inhibitors. *Nature Reviews Disease Primers*, 6(1):38, 2020.

[Raschi *et al.*, 2019] Emanuel Raschi, Alessandra Mazzarella, Ippazio Cosimo Antonazzo, Nicolò Bendinelli, Emanuele Forcesi, Marco Tuccori, Ugo Moretti, Elisabetta Poluzzi, and Fabrizio De Ponti. Toxicities with immune checkpoint inhibitors: emerging priorities from disproportionality analysis of the fda adverse event reporting system. *Targeted Oncology*, 14(2):205–221, 2019.

[Rizve *et al.*, 2021] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *ArXiv Preprint ArXiv:2101.06329*, 2021.

[Sakaeda *et al.*, 2013] Toshiyuki Sakaeda, Akiko Tamon, Kaori Kadoyama, and Yasushi Okuno. Data mining of the public version of the fda adverse event reporting system. *International Journal of Medical Sciences*, 10(7):796, 2013.

[Samuli and Timo, 2017] Laine Samuli and Aila Timo. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations (ICLR)*, volume 4, page 6, 2017.

[Semple *et al.*, 2011] John W Semple, Joseph E Italiano, and John Freedman. Platelets and the immune continuum. *Nature Reviews Immunology*, 11(4):264–274, 2011.

[Shapley and others, 1953] Lloyd S Shapley et al. A value for n-person games. 1953.

[Smyth *et al.*, 2020] Elizabeth C Smyth, Magnus Nilsson, Heike I Grabsch, Nicole CT van Grieken, and Florian Lordick. Gastric cancer. *The Lancet*, 396(10251):635–648, 2020.

[Sohn *et al.*, 2020] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33:596–608, 2020.

[Sonmez and Sonmez, 2017] Ozge Sonmez and Mehmet Sonmez. Role of platelets in immune system and inflammation. *Porto Biomedical Journal*, 2(6):311–314, 2017.

[Tai *et al.*, 2021] Kai Sheng Tai, Peter D Bailis, and Gregory Valiant. Sinkhorn label allocation: Semi-supervised classification via annealed self-training. In *International Conference on Machine Learning*, pages 10065–10075. PMLR, 2021.

[Tarvainen and Valpola, 2017] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30, 2017.

[Totani and Evangelista, 2010] Licia Totani and Virgilio Evangelista. Platelet–leukocyte interactions in cardiovascular disease and beyond. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 30(12):2357–2361, 2010.

[Ucar *et al.*, 2021] Talip Ucar, Ehsan Hajiramezanali, and Lindsay Edwards. Subtab: Subsetting features of tabular data for self-supervised representation learning. *Advances in Neural Information Processing Systems*, 34:18853–18865, 2021.

[Wang *et al.*, 2022] Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, et al. Freematch: Self-adaptive thresholding for semi-supervised learning. *ArXiv Preprint ArXiv:2205.07246*, 2022.

[Wang *et al.*, 2023] Xi Wang, Yuming Jiang, Hao Chen, Taojun Zhang, Zhen Han, Chuanli Chen, Qingyu Yuan, Wenjun Xiong, Wei Wang, Guoxin Li, et al. Cancer immunotherapy response prediction from multi-modal clinical and image data using semi-supervised deep learning. *Radiotherapy and Oncology*, 186:109793, 2023.

[Wang *et al.*, 2024] Jingting Wang, Yan Ma, Haishan Lin, Jing Wang, and Bangwei Cao. Predictive biomarkers for immune-related adverse events in cancer patients treated with immune-checkpoint inhibitors. *BMC Immunology*, 25(1):8, 2024.

[Xie *et al.*, 2020] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268, 2020.

[Yang *et al.*, 2022] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):8934–8954, 2022.

[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. *Advances in Neural Information Processing Systems*, 33:11033–11043, 2020.

[Zhang *et al.*, 2017] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *ArXiv Preprint ArXiv:1710.09412*, 2017.

[Zhang *et al.*, 2021] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021.

[Zhou *et al.*, 2022] Jian-Guo Zhou, Ada Hang-Heng Wong, Haitao Wang, Fangya Tan, Xiaofei Chen, Su-Han Jin, Si-Si He, Gang Shen, Yun-Jia Wang, Benjamin Frey, et al. Elucidation of the application of blood test biomarkers to predict immune-related adverse events in atezolizumab-treated nsclc patients using machine learning methods. *Frontiers in Immunology*, 13:862752, 2022.