MVSTAB: BRIDGING STATISTICAL AND SEMANTIC SUBSPACES FOR TABULAR REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised representation learning on tabular data predominantly relies on heuristic feature perturbations or ad-hoc column partitioning. However, these paradigms frequently suffer from a fundamental semantic-statistical gap: they optimize solely for empirical correlations while remaining blind to domain-level semantic structures. To bridge this divide, we propose MVSTab, a multi-view tabular representation learning framework that synergizes empirical statistical patterns with high-level semantic priors. Departing from heuristic masking, MVSTab introduces a dual-driven partitioning strategy that combines language model representations with correlation clustering to disentangle features into complementary semantic and statistical subspaces. To prevent representations from exploiting superficial numerical shortcuts, we design a semantic-aware hard negative generation mechanism. By selectively perturbing intra-group dependencies while preserving single-feature marginal plausibility, this mechanism compels the model to resolve nuanced logical contradictions rather than trivially detectable statistical outliers. Optimized via a tri-objective contrastive loss and aggregated through latent cross-attention, MVSTab produces dense, fixed-dimensional, and logically consistent embeddings. Extensive evaluations across diverse benchmarks show that MVSTab consistently outperforms competitive tree-based and deep learning baselines, especially in label-scarce regimes. Crucially, downstream analyses reveal that MVSTab uniquely excels at uncovering stealthy, logic-driven anomalies that evade conventional density- and reconstruction-based approaches.

1 INTRODUCTION

Deep neural networks excel in domains with clear structural priors like spatial grids in images Nguyen et al. (2025) or sequential orders in text Gardazi et al. (2025). In comparison, tabular data, as the most common data format in finance, healthcare, and recommendation systems, exhibits a high degree of heterogeneity, lacks structural locality, and shows an irregular distribution. Therefore, deep tabular models often underperform even tree-based models Gorishniy et al. (2021) Borisov et al. (2022) Grinsztajn et al. (2022). To bridge this gap, researchers focus on tabular representation learning. The goal is to extract dense, logically consistent embeddings for diverse downstream applications.

To achieve robust representation learning without expensive manual labels, self-supervised learning (SSL) has become the preferred paradigm Wang et al. (2025). Existing SSL paradigms include contrastive learning, reconstruction, and mask prediction Rubachev et al. (2022) Rabbani et al. (2025). These methods guide models to discover consistent patterns among heterogeneous attributes. Frameworks like TabNet Arik & Pfister (2021) and SCARF Bahri et al. (2021) create two views of a sample through random feature corruption for contrastive learning. Models like SubTab Ucar et al. (2021), SAINT Somepalli et al. (2021), and Stunt Nam et al. (2023) use data augmentation or feature subset partitioning. All these frameworks address two key bottlenecks. First, how to handle heterogeneous features. Solutions range from treating the entire row as one vector to splitting it into localized, multi-view subspaces. Second, how to generate SSL signals. Models build proxy tasks through data corruption, masking, or negative sampling to define decision boundaries.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Despite significant progress, existing frameworks face two limitations. First, despite each column in tabular data encapsulating a distinct physical or logical concept, current subspace partitioning strategies rely on random splitting or purely empirical correlations Peel et al. (2022). They treat these columns merely as independent numerical values, overlooking the physical significance underlying the features. This results in a fundamental disconnect between empirical data statistics and human domain knowledge. For example, in a medical application, systolic blood pressure and age may show low statistical correlation in a noisy data batch. A data-driven partitioner might split them apart. However, these two variables have important joint diagnostic significance.

Second, this semantic blindness severely compromises the model's ability to learn robust decision boundaries. Existing SSL signal generation methods struggle to construct hard negative samples that require deep reasoning. Current approaches typically construct these signals by random dropout Dorkenwald et al. (2022), Mixup or global feature swapping, thereby generating contrastive views. However, these perturbation methods overlook the semantic information inherent in the features. Global or random swapping disrupts the joint distribution throughout the feature space, resulting in the generated pseudo anomalies having obvious contradictions. For example, the patient's gender does not match the gender-specific medical indicators. These abnormal samples significantly deviate from the real data manifold. Therefore, networks exploit these numerical discrepancies as lazy statistical shortcuts. By merely memorizing shallow boundary deviations to minimize the loss, they fail to acquire deep reasoning capabilities. Ultimately, models remain oblivious to structurally stealthy anomalies, samples whose individual dimensions appear statistically normal, yet their combinatorial logic is inherently contradictory..

To address these issues, we propose the **M**ulti-**V**iew **S**elf-supervised **Tab**ular Representation Model (**MVSTab**), which learns tabular representations through a dual-driven framework that integrates empirical statistical evidence with human-level semantic priors. Rather than relying solely on observed correlations or predefined domain rules, MVSTab harnesses their complementary strengths. While statistical patterns capture dataset-specific regularities, semantic relationships offer principled guidance when empirical signals are noisy or incomplete. This dual perspective enables the model to uncover meaningful feature interactions instead of treating columns as isolated numerical inputs.

Specifically, MVSTab preserves the holistic context of each record by retaining global numerical and categorical views, while constructing two distinct types of local views: statistical views that cluster features based on empirical correlations, and semantic views that group conceptually related features via language-model representations of column metadata. We optimize these views using three complementary contrastive objectives. Within-view contrastive learning enhances representation robustness against perturbations; cross-view contrastive learning aligns data-driven dependencies with semantic priors without enforcing rigid view equivalence; and hard-negative contrastive learning tightens decision boundaries by penalizing violations of local feature dependencies. To construct informative hard negatives, we perturb values within semantically related feature groups, preserving marginal plausibility while disrupting joint compatibility. This compels the model to capture joint feature consistency rather than relying on trivially detectable, out-of-distribution corruptions. Finally, a latent cross-attention mechanism adaptively pools information across these variable-length views into fixed-dimensional representations for downstream prediction.

The contributions of this research are summarized as follows:

- We propose MVSTab, a dual-driven multi-view tabular representation learning framework that bridges empirical data distributions with human-level conceptual priors. By coupling data-driven statistical correlations with LLM-guided semantic knowledge, MVSTab constructs complementary global and localized subspace views, effectively alleviating feature redundancy.

- We formulate multi-view representation learning as a multi-objective contrastive optimization problem with logic-corrupted hard negatives, governed by three specialized objectives: Within-View Contrastive (WVC), Cross-View Contrastive (CVC), and Hard Negative Contrastive (HNC) loss. Crucially, by perturbing feature values strictly within LLM-defined semantic groups, we synthesize semantically hard negative samples that corrupt local joint dependencies while preserving marginal invariants, jointly enforcing perturbation robustness, cross-view semantic consensus, and fine-grained counterfactual discrimination.

- We achieve dimension-agnostic latent fusion and superior generalization across diverse tabular benchmarks by introducing a latent cross-attention mechanism with learnable queries to compress variable numbers of feature views into a fixed-dimensional unified representation, with extensive experiments demonstrating that MVSTab consistently outperforms state-of-the-art deep and tree-based baselines while exhibiting exceptional data efficiency.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 details the architecture, key components, and theoretical foundations of MVSTab. Section 4 presents the experimental setup and evaluates MVSTab across multiple benchmark datasets. Finally, Section 5 concludes the paper.

2 Related Works

Deep Models for Tabular Data. Without the spatial or sequential structure of images and text, tabular data remains challenging for representation learning, and gradient-boosted decision trees often outperform generic deep networks Speiser et al. (2019); Pasaribu et al. (2024); Shwartz-Ziv & Armon (2022). Recent deep tabular models mainly improve the modeling of feature importance and feature interactions. TabNet Arik & Pfister (2021) employs sequential attentive feature selection to identify informative attributes, while TabTransformer Huang et al. (2020) applies self-attention to contextualize categorical features. FT-Transformer Gorishniy et al. (2021) further extends Transformer-based modeling to heterogeneous tabular features, and numerical feature embedding methods show that the representation of continuous attributes has a substantial impact on downstream performance Gorishniy et al. (2022). Other approaches, such as GANDALF Joseph (2022), combine gating mechanisms with tree-inspired structures to balance predictive capacity and interpretability. Despite these advances, most deep tabular models learn feature interactions directly from observed values and do not explicitly distinguish between empirical dependencies and semantically meaningful relationships. Moreover, global interaction modeling may overlook local dependencies among groups of related features.

Self-Supervised Tabular Representation Learning. Existing methods mainly use masking, reconstruction, perturbation, and multi-view consistency to learn from unlabeled data. VIME Yoon et al. (2020) combines mask estimation with feature reconstruction; SCARF Bahri et al. (2021) creates contrastive views through random feature corruption; and STab Hajiramezanali et al. (2022) uses stochastic regularization to improve robustness. While effective, these instance-level perturbations are generally independent of feature semantics. Other methods construct views from feature subsets: SubTab Ucar et al. (2021) uses subset partitioning and collaborative reconstruction, while related work disentangles feature subspaces for one-class anomaly detection Ye et al. (2025). SAINT Somepalli et al. (2021) combines row- and feature-level attention with contrastive learning; RaTab Eo et al. (2025) uses truncated singular value decomposition for dimensionality reduction; and TabDPT Ma et al. (2026) scales pretraining to large tabular corpora. Although these methods capture feature heterogeneity, random partitioning, generic augmentation, and global corruption may overlook local dependencies. Their negative samples may also exhibit marginal distribution shifts, allowing the model to exploit statistical shortcuts rather than learn feature compatibility.

Semantic and Schema Knowledge for Tabular Learning. Observed feature values provide only one source of information for tabular learning. Column names, descriptions, data types, and schema-level relations can provide additional knowledge about the meaning and possible interactions of features. Recent studies have therefore explored the use of textual information and language models for tabular understanding. GOGGLE Liu et al. (2023b) investigates the use of learned structural information for tabular generation, while retrieval-augmented and language-model-based approaches incorporate table descriptions or external knowledge to improve table understanding Sui et al. (2024); Liu et al. (2023a); Dong & Wang (2024). ConTextTab Spinaci et al. (2026) directly injects column-name text embeddings into an in-context tabular learner, conditioning predictions on natural-language column semantics. DiSPaT Vuong et al. (2026) further applies LLM-guided self-play with an f-divergence objective to synthesize anomalies for tabular anomaly detection.XTab Zhu et al. (2023) exploits cross-table pretraining to transfer knowledge across heterogeneous schemas, demonstrating the value of metadata and shared semantic information for tabular generalization. Nevertheless, existing methods typically treat semantic information merely as auxiliary input, textual augmentation, or a medium for cross-table transfer, overlooking its explicit integration with self-

supervised feature organization and contrastive signal generation. In contrast, MVSTab incorporates column semantics as a structural prior: metadata-derived representations serve to identify semantically related features, thereby complementing the dependencies inferred from observed data. Consequently, this design empowers the model to learn representations that harmonize dataset-specific statistics with high-level semantic relationships.

3 METHODOLOGY

3.1 PROBLEM FORMULATION

Let the input tabular dataset be $\mathcal{X} \in \mathbb{R}^{N \times M}$, where N denotes the sample size and M represents the total number of features. In this paper, our objective is to process $\mathcal{X}$ to extract a dense, logically consistent global representation h_{fused}. This tabular representation can subsequently be applied to various downstream tasks, such as unsupervised anomaly detection, as well as supervised classification, regression tasks, and others.

3.2 FRAMEWORK OVERVIEW

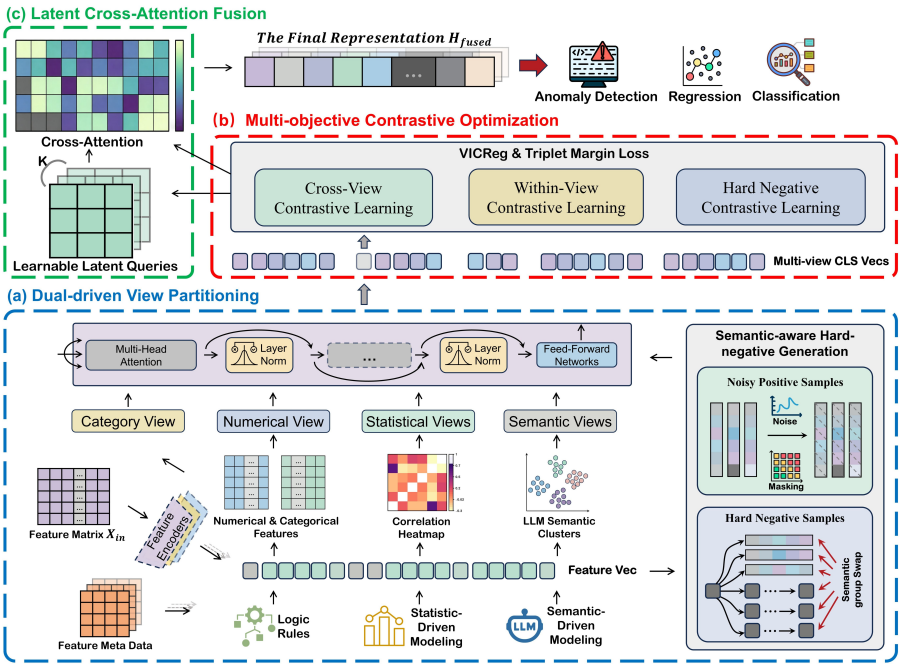

Figure 1: Overall architecture of MVSTab. (a) Raw features are mapped into tokens and partitioned into global views and local subspaces. (b) WVC, CVC, and HNC jointly shape the latent space. (c) K learnable latent queries use cross-attention to aggregate all view representations into a fixed-dimensional embedding.

To model tabular heterogeneity without incurring the $O(M^2)$ quadratic complexity of global attention, MVSTab decomposes representation learning into a multi-view hierarchical architecture, as illustrated in Figure 1:

Dual-Driven View Partitioning: Features are decoupled along two axes: a statistical subspace based on statistical evidence and a semantic subspace based on semantic priors. Global numerical and categorical views are simultaneously preserved to maintain macro-level contextual information. **Shared Multi-View Encoding:** All subspaces and global views are tokenized and fed into a shared Transformer backbone equipped with learnable CLS tokens. These CLS vectors serve to aggregate high-order interaction information. **Multi-objective contrastive optimization and latent space fusion:** The latent space is constructed using three complementary contrastive mechanisms. Finally, a latent cross-attention module aggregates all view representations into a fixed-dimensional

global embedding, h_{fused}, thereby decoupling the output dimensionality from the original number of features, M.

Figure 2 further illustrates the details of stage (b). During the multi-objective contrastive optimization stage, MVSTab uses the extracted CLS vectors to compute three complementary contrastive forces: Within-View Contrastive, Cross-View Contrastive, and Hard Negative Contrastive. The global numerical and categorical CLS vectors bypass this contrastive stage to prevent the destruction of macro-level information.

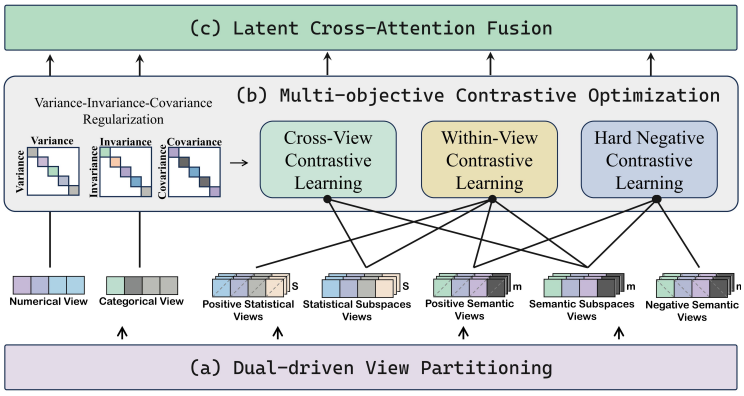

Figure 2: Details of the Multi-objective contrastive optimization phase.

3.3 Dual-driven View Partitioning and Shared Encoding

We separate continuous features $\mathcal{F}_{num}$ from discrete features $\mathcal{F}_{cat}$. To capture empirical distributions, we construct statistical subspaces $\mathcal{S}_{stat} = \{S_{stat}^{(1)}, S_{stat}^{(2)}, ..., S_{stat}^{(s)}\}$ by computing the Pearson correlation matrix $C \in \mathbb{R}^{M \times M}$ and performing agglomerative clustering (details in Appendix B).

Crucially, to capture realization-independent logic, we construct semantic subspaces $\mathcal{S}_{sem} = \{S_{sem}^{(1)}, ..., S_{sem}^{(m)}\}$. We generate semantic embeddings $E_i \in \mathbb{R}^{d_{llm}}$ by prompting a pre-trained LLM with feature metadata and applying sequence-length mean pooling

$$E_i = \frac{\sum_{l=1}^{L} w_l \cdot \text{LLM}(T_i)_l}{\max(\sum_{l=1}^{L} w_l, \varepsilon)}. \tag{1}$$

Applying K-Means to $\{E_1, ..., E_M\}$ yields S_{sem}. This orthogonal dual-partitioning ensures that features sharing physical definitions are grouped together, even if local noise diminishes their statistical correlation. All views are projected into d_{model}, appended with view-specific learnable [CLS] tokens, and processed by a shared Transformer encoder to yield representations h_v. Theoretical proofs regarding the schema-data independence of this dual partitioning are deferred to Appendix A.

3.4 Semantic-aware Hard-negative Generation

To synthesize hard negative samples that require rigorous logical reasoning, we introduce an intra-semantic swapping mechanism. Given a batch of data and a specific semantic subspace $S_{sem}^{(j)}$ containing N_j features, we apply a random permutation function π across the sample dimension for each feature $f \in S_{sem}^{(j)}$ with probability p_{swap}. We define the corrupted feature matrix $\tilde{X}_{hard}$ as:

$$\tilde{X}_{hard}[i,f] = \begin{cases} X[\pi(i),f], & \text{if } u \sim U(0,1) < p_{swap} \\ X[i,f], & \text{otherwise} \end{cases} \tag{2}$$

The encoder processes the corrupted data to obtain the hard negative representation h_{hard}. In contrast, we generate positive augmented representations h_{aug} by applying a standard dropout mask with probability p_{mask}. This mechanism preserves the marginal distribution $P(f)$ of individual features

completely unchanged. Shallow linear models cannot detect the anomaly. However, it severely disrupts the joint distribution $P(f_1, f_2)$ within the same semantic domain. This forces the deep learning architecture to understand logical contradictions.

3.5 Multi-objective Contrastive Optimization

To learn statistically stable, semantically consistent, and logically discriminative representations, our Multi-objective Contrastive Optimization module coordinates three distinct mechanisms: Within-View Contrastive (WVC), Cross-View Contrastive (CVC), and Hard Negative Contrastive (HNC) loss. Let θ denote the trainable parameters of the shared encoder and view-specific representation modules. We express the learning problem in its vector-valued form:

$$\min_{\theta} \mathcal{L}_{total} = [\mathcal{L}_{WVC}(\theta), \mathcal{L}_{CVC}(\theta), \mathcal{L}_{HNC}(\theta)] \tag{3}$$

To obtain a tractable single training objective while retaining the multi-objective interpretation, we employ a positive weighted scalarization:$\mathcal{L}_{total} = \alpha\mathcal{L}_{WVC} + \beta\mathcal{L}_{CVC} + \gamma\mathcal{L}_{HNC}$. The resulting scalar objective is optimized jointly with the shared encoder parameters. At training iteration t, the parameters are updated according to $\theta^{t+1} = \theta^t - \eta_t \nabla_\theta L_{total}(\theta^t)$.

Within-View Contrastive (WVC) Loss: Instead of batch-heavy InfoNCE, WVC adopts VICReg Bardes et al. (2022) to maintain perturbation invariance while preventing collapse via variance and covariance terms:

$$\mathcal{L}_{WVC} = \frac{1}{|V|} \sum_{v \in V} (\lambda s(h_v, h'_v) + \mu v(h_v) + \nu c(h_v) \tag{4}$$

where s, v, c denote the invariance, variance, and covariance regularization terms respectively, detailed in Appendix B.

Cross-View Contrastive (CVC) Loss: CVC aligns different modalities of the tabular data. The algorithm contrasts the representations of the statistical subspaces h_{stat} with the representations of the semantic subspaces h_{sem}. CVC applies VICReg to enforce alignment between these heterogeneous views:

$$\mathcal{L}_{CVC} = \frac{1}{k \times m} \sum_{i=1}^{k} \sum_{j=1}^{m} \text{VICReg}(h_{stat}^{(i)}, h_{sem}^{(j)}) \tag{5}$$

CVC narrows the gap between representations from different view categories. Its objective is to maximize the mutual information between these views.

Hard Negative Contrastive (HNC) Loss: HNC explicitly endows the latent space with anomaly detection capabilities. The algorithm contrasts the normal sample (Anchor: h_{sem}) with the masked normal sample (Positive: h_{aug}) and the logically swapped sample (Negative: h_{hard}). We use a Triplet Margin Loss:

$$\mathcal{L}_{HNC} = \frac{1}{m} \sum_{j=1}^{m} \max\left(0, ||h_{sem}^{(j)} - h_{aug}^{(j)}||_2 - ||h_{sem}^{(j)} - h_{hard}^{(j)}||_2 + \delta\right) \tag{6}$$

where δ is the margin boundary. HNC trains the model to recognize structural anomalies by pulling the anchor and positive closer while pushing realistic but flawed negatives away. This establishes exact boundaries of normal logic.

The update rule for the total loss in the gradient flow is $\nabla_\theta \mathcal{L}_{total} = \alpha\nabla_\theta \mathcal{L}_{WVC} + \beta\nabla_\theta \mathcal{L}_{CVC} + \gamma\nabla_\theta \mathcal{L}_{HNC}$. Within our framework, multi-view representations divide into global macro views (h_{num}, h_{cat}) and local grouped views (h_{stat}, h_{sem}). The self-supervised contrastive constraints apply exclusively to the local grouped views. Conversely, exempting global macro views from aggressive contrastive constraints preserves the integrity and purity of the original information.

3.6 Latent Cross-Attention Fusion

MVSTab employs a latent cross-attention mechanism to map the variable-length set of views into a fixed-dimensional representation. Let $H_{all} \in \mathbb{R}^{N \times V \times d_{model}}$ denote the representations of all global and subspace views, where V is the number of views. We initialize a set of $K_{latents}$ learnable latent

queries $Q_{latent} \in \mathbb{R}^{K_{latents} \times d_{model}}$. The number K is fixed across datasets. Each latent query attends to all view representations and extracts a different combination of view-level information. We then use multi-head cross-attention to update the latent queries and produce the fixed-dimensional h_{fused}.

$$\begin{aligned} &Q = Q_{latent}\mathbf{W}_Q, K = H_{all}\mathbf{W}_K, U = H_{all}\mathbf{W}_V \\ &H_{latent} = \text{Softmax}(\frac{QK^T}{\sqrt{d_{model}}})U \end{aligned} \tag{7}$$

We process the output through a feed-forward network and flatten it to generate the final global representation

$$h_{fused} = Flatten(FFN(H_{latent})) \in R^{N \times (K_{latents} \cdot d_{model})} \tag{8}$$

The latent queries learn to extract complementary information from different views. Each query can attend to multiple views with different weights. Thus, the queries can capture different combinations of view-level information. By recording the attention matrix $A \in \mathbb{R}^{K \times V}$ and averaging it across the K queries, one can obtain view-level importance scores. This offers a simple method to examine which views contribute the most to the final representation. Complexity analysis demonstrates MVSTab's linear scalability w.r.t the number of views, and details for downstream task adaptation are provided in Appendix B.

4 Experiments

4.1 Dataset and Baselines

We evaluate MVSTab on 12 public tabular datasets spanning anomaly detection, classification, and regression. We compare against 26 advanced baselines, including tree-based models, density-based anomaly detectors, and deep self-supervised frameworks. Comprehensive details on dataset statistics, baseline configurations, and evaluation metrics are meticulously documented in Appendix C.

In these experiments, we randomly split the data into a training set (80%) and a test set (20%). To control for initialization and split variance, each configuration was independently repeated under four random seeds (42, 2026, 7, 123). All values reported are the mean over these four runs. The parameter sensitivity analysis is presented in Appendix D.

For the unsupervised anomaly detection and the supervised classification task, we use Accuracy and Area Under the Receiver Operating Characteristic Curve (AUROC) as the evaluation metrics. For the supervised regression task, we used Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) as evaluation metrics.

4.2 Results and Analysis

Table 1 presents the performance of MVSTab and 13 baseline methods in the unsupervised anomaly detection task. The best results for each metric are highlighted in bold. All self-supervised baseline methods, as well as MVSTab, are combined with the ECOD method to calculate anomaly scores. The performance of MVSTab on classification and regression tasks is presented in Appendix D.

The results presented in Table 1 demonstrate that MVSTab exhibits competitive performance. It consistently tops the charts across all benchmarked datasets in accuracy. For the continuous ranking metric AUROC, the proposed method establishes the best results on the Heart and Marketing datasets with significant margins over contrastive baselines like SCARF and GOAD. However, on the Sleep, Bank, and Donors datasets, reconstruction-based or density-based methods such as VIME and HBOS occasionally yield higher AUROC scores. This phenomenon suggests that while our multi-view latent representation provides a highly separable global space for general anomaly thresholding, the continuous distribution ranking evaluated by AUROC might occasionally favor algorithms explicitly optimized for local density estimation or feature reconstruction. This depends on the intrinsic data manifold.

Table 1: The performance of MVSTab and baseline methods on unsupervised anomaly detection tasks.

Dataset	Sleep		Heart		Marketing		Bank		Donors	
Metrics	Accuracy	AUROC	Accuracy	AUROC	Accuracy	AUROC	Accuracy	AUROC	Accuracy	AUROC
KNN	0.760	0.404	0.633	0.634	0.657	0.483	0.886	0.680	0.878	0.872
Iforest	0.747	0.537	0.656	0.711	0.660	0.557	0.878	0.624	0.889	0.840
HBOS	0.773	0.566	0.683	0.689	0.622	0.504	0.887	**0.722**	0.927	0.780
DeepSVDD	0.787	0.597	0.733	0.707	0.681	0.694	0.879	0.617	0.859	0.680
COPOD	0.867	0.705	0.650	0.614	0.641	0.495	0.896	0.705	0.908	0.789
ECOD	0.853	0.835	0.683	0.735	0.651	0.522	0.890	0.688	0.927	0.865
REPEN	0.760	0.488	0.717	0.720	0.751	0.781	0.881	0.606	0.887	0.880
GOAD	0.693	0.645	0.767	0.783	0.747	0.693	0.889	0.584	0.870	0.817
ICL	0.707	0.623	0.783	0.767	0.724	0.667	0.876	0.594	0.892	0.772
SCARF	0.813	0.590	0.633	0.728	0.756	0.601	0.939	0.632	0.919	0.836
SubTab	0.853	0.847	0.700	0.760	0.753	0.581	0.949	0.696	0.892	0.718
VIME	0.873	**0.875**	0.667	0.783	0.776	0.612	0.944	0.681	0.903	**0.896**
SemanticMask	0.877	0.839	0.785	0.812	0.799	0.697	0.961	0.684	0.916	0.839
MVSTab	**0.881**	0.797	**0.866**	**0.815**	**0.808**	**0.789**	**0.978**	0.693	**0.941**	0.852

4.3 Ablation and Case Studies

Table 2 details the ablation configurations, results, and execution times for the Marketing, Salary, and LOS datasets across three types of tasks.

Table 2: The configuration and experimental results of the ablation study for MVSTab.

Modle	Ablation Settings	Marketing			Salary			LOS		
		Accuracy	AUROC	Time	Accuracy	AUROC	Time	MAE	RMSE	Time
M0	Full Model	0.808	0.789	42	0.882	0.899	160	0.424	0.636	431
M1	Fixed-Query Fusion	0.803	0.784	42	0.877	0.894	159	0.429	0.644	428
M2	No CVC-loss	0.799	0.781	40	0.873	0.890	153	0.434	0.658	415
M3	Simple Fixed-Dim Aggregation	0.797	0.776	41	0.869	0.885	154	0.439	0.667	418
M4	No WVC-loss	0.792	0.771	40	0.862	0.875	148	0.453	0.691	376
M5	No Global Views	0.785	0.760	39	0.856	0.870	146	0.461	0.704	395
M6	No HNC-loss	0.781	0.748	34	0.865	0.888	130	0.453	0.681	343

As shown in Table 2, all components contribute to the performance of MVSTab. Removing CVC (M2) or WVC (M4) leads to a significant decline in anomaly detection performance, confirming the necessity of semantic-statistical alignment and perturbation invariance. Crucially, removing HNC (M6) severely impairs anomaly detection effectiveness, directly validating its indispensable role in explicitly penalizing logical contradictions. The global view (M5) and cross-attention mechanisms (M1, M3) also consistently provide macro-level contextual information and task-specific view weighting.

Preprocessing costs are amortized across the entire training process. Computing the Pearson correlation matrix and hierarchical clustering for statistical partitioning takes approximately 2-3 seconds for datasets with $M < 100$ features. Semantic partitioning via LLM embeddings and clustering is performed once offline. This requires less than 5 seconds for typical feature counts. These one-time costs are negligible compared to the iterative training overhead.

Empirical validation of the Hard negative mechanism. We validate two core properties of Eq.2: (i) preserving individual feature marginals, and (ii) disrupting intra-group joint dependencies. First, a two-sample Kolmogorov-Smirnov test between pre- and post-swap columns yields $p > 0.05$ for 100% of the swapped features across all datasets, confirming that marginal distributions $P(f)$ remain strictly unchanged. Second, Table 3 reports the percentage drop in mutual information $I(f_1; f_2)$ after semantic-group swapping. All datasets exhibit a measurable decrease, verifying that our mechanism precisely disrupts the intended joint dependencies. Furthermore, as shown by the IsolationForest anomaly score shifts (IF-shift), our semantic-group swapping strategy more effectively generates hard negatives that challenge traditional detectors compared to global-shuffling strategies.

Detection of samples that have disrupted the marginal distribution. We evaluated model sensitivity using the Sleep and Heart datasets under two anomaly paradigms. Easy anomalies were constructed by forcing specific continuous features into physically extreme bounds, e.g., setting Heart Rate to 200 or 300, and Sleep Duration to 12 or 20. Or pushing them slightly beyond normal thresholds, e.g., Ejection Fraction of 80/99, Serum Creatinine of 15/50. This evaluates models' basic sensitivity to marginal distribution shifts. In contrast, hard anomalies keep all feature values within

Table 3: The validation of the negative sampling mechanism was conducted on 12 datasets, and the results represent the average of four runs with different random seeds.

Dataset	MI-drop (%)	IF-shift (Sem.)	IF-shift (Global)
Sleep	50.0	0.68	0.65
Heart	14.5	0.12	0.07
Marketing	16.3	0.06	0.07
Bank	51.0	0.18	0.02
Donors	68.9	0.32	0.25
Credit	23.0	0.10	0.04
DMFT	73.4	0.24	0.27
CMC	31.9	0.10	0.09
Salary	37.8	0.13	0.07
LOS	18.9	0.08	0.02
Penguin	27.2	0.27	0.29
DMC Heart	25.4	0.19	0.13

normal statistical ranges but disrupt the combinatorial logic by swapping feature values between two randomly selected samples within the same semantic group.

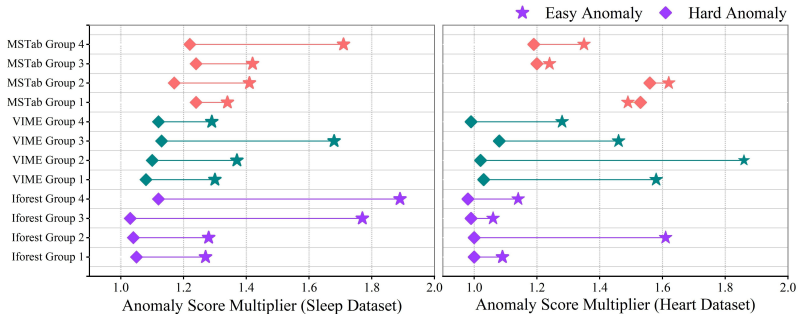

Figure 3: The average anomaly score multipliers of the three models, based on different types of anomalous samples.

When dealing with simple anomalies, all three models exhibit a multiplier greater than 1, indicating successful detection. However, when evaluating difficult anomalies, the score multipliers for IForest and VIME drop to near 1, suggesting that they misclassify samples with logical-level corruption as normal. In contrast, MVSTab maintains a robust detection multiplier even for complex anomalies.

4.4 Discussion

Our results validate that pure data-driven learning is insufficient for complex tabular manifolds, as conventional models easily overfit spurious statistical correlations. By incorporating LLM-derived semantics and enforcing cross-view alignment, MVSTab acts as a knowledge-driven regularizer that bridges the empirical-semantic gap. Furthermore, the anomaly detection case studies expose a critical vulnerability in existing unsupervised paradigms: models relying purely on marginal distribution deviations completely fail when presented with structurally stealthy, logic-based anomalies. By introducing semantic-aware hard negative contrastive learning, MVSTab shifts the tabular anomaly detection paradigm from mere outlier identification to rigorous logical inconsistency detection.

5 Conclusion

This paper presents MVSTab, a multi-view self-supervised framework for tabular representation learning. By coupling statistical correlations with LLM-guided conceptual priors through a multi-objective contrastive optimization space, MVSTab produces dense, logically consistent, and robust embeddings. Extensive evaluations demonstrate its superior generalization across diverse downstream tasks. While MVSTab's reliance on descriptive column metadata presents a limitation in heavily anonymized settings, it establishes a foundational stepping stone toward reasoning-capable systems for tabular data.

AI USE STATEMENT

In this work, we used generative AI tools to polish the research paper. We have reviewed all work assisted by AI. We take responsibility for the final content of this work, including any text, statements, or outputs generated using generative AI.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, we have provided sample source code. Full details regarding the datasets, experimental setup, and specific implementation are provided in the appendix.

REFERENCES

Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey E Hinton. Neural additive models: Interpretable machine learning with neural nets. *Advances in neural information processing systems*, 34:4699–4711, 2021.

Sercan Ö Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 6679–6687, 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.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Roger Barga, Valentine Fontama, Wee Hyong Tok, and Luis Cabrera-Cordon. *Predictive analytics with Microsoft Azure machine learning*. Springer, 2015.

Liron Bergman and Yedid Hoshen. Classification-based anomaly detection for general data. In *International Conference on Learning Representations (ICLR)*, 2020.

Bernd Bischl, Giuseppe Casalicchio, Matthias Feurer, Pieter Gijsbers, Frank Hutter, Michel Lang, Rafael G Mantovani, Jan N van Rijn, and Joaquin Vanschoren. Openml benchmarking suites. In *Proceedings of the Neural Information Processing Systems*, 2021.

Sandeep Chandra Bollepalli, Ashish Kumar Sahani, Naved Aslam, Bishav Mohan, Kanchan Kulkarni, Abhishek Goyal, Bhupinder Singh, Gurbhej Singh, Ankit Mittal, Rohit Tandon, et al. An optimized machine learning model accurately predicts in-hospital outcomes at admission to a cardiac unit. *Diagnostics*, 12(2):241, 2022.

Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE transactions on neural networks and learning systems*, 35(6):7499–7519, 2022.

Kuan-Yu Chen, Ping-Han Chiang, Hsin-Rung Chou, Ting-Wei Chen, and Tien-Hao Chang. Trompt: Towards a better deep neural network for tabular data. In *Proceedings of the 40 th International Conference on Machine Learning (ICML)*, 2023.

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*, pp. 785–794, 2016.

Yi Cheng, Renjun Hu, Haochao Ying, Xing Shi, Jian Wu, and Wei Lin. Arithmetic feature interaction is necessary for deep tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11516–11524, 2024.

Davide Chicco and Giuseppe Jurman. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC medical informatics and decision making*, 20(1):16, 2020.

David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 20(2):215–232, 1958.

Haoyu Dong and Zhiruo Wang. Large language models for tabular data: Progresses and future directions. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pp. 2997–3000, 2024.

Michael Dorkenwald, Fanyi Xiao, Biagio Brattoli, Joseph Tighe, and Davide Modolo. Scvrl: Shuffled contrastive video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4132–4141, 2022.

Moonjung Eo, Kyungeun Lee, Hye-Seung Cho, Dongmin Kim, Ye Seul Sim, and Woohyung Lim. Representation space augmentation for effective self-supervised learning on tabular data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11625–11633, 2025.

Nadia Mushtaq Gardazi, Ali Daud, Muhammad Kamran Malik, Amal Bukhari, Tariq Alsahfi, and Bader Alshemaimri. Bert applications in natural language processing: a review. *Artificial Intelligence Review*, 58(6):166, 2025.

Markus Goldstein and Andreas Dengel. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. *KI-2012: poster and demo track*, 1:59–63, 2012.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in neural information processing systems*, 34:18932–18943, 2021.

Yury Gorishniy, Ivan Rubachev, and Artem Babenko. On embeddings for numerical features in tabular deep learning. *Advances in Neural Information Processing Systems*, 35:24991–25004, 2022.

Kristen B Gorman, Tony D Williams, and William R Fraser. Ecological sexual dimorphism and environmental variability within a community of antarctic penguins (genus pygoscelis). *PloS one*, 9(3):e90081, 2014.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35: 507–520, 2022.

Ehsan Hajiramezanali, Nathaniel Lee Diamant, Gabriele Scalia, and Max W Shen. Stab: Self-supervised learning for tabular data. In *NeurIPS 2022 first table representation workshop*, 2022.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*, 2020.

Manu Joseph. Gandalf: Gated adaptive network for deep automated learning of features. *arXiv preprint arXiv:2207.08548*, 2022.

Arlind Kadra, Sebastian Pineda Arango, and Josif Grabocka. Interpretable mesomorphic networks for tabular data. *Advances in Neural Information Processing Systems*, 37:31759–31787, 2024.

Ron Kohavi. Census Income. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5GP7S.

Md Tahmid Rahman Laskar, Jimmy Xiangji Huang, Vladan Smetana, Chris Stewart, Kees Pouw, Aijun An, Stephen Chan, and Lei Liu. Extending isolation forest for anomaly detection in big data via k-means. *ACM Transactions on Cyber-Physical Systems (TCPS)*, 5(4):1–26, 2021.

Zheng Li, Yue Zhao, Nicola Botta, Cezar Ionescu, and Xiyang Hu. Copod: copula-based outlier detection. In *2020 IEEE international conference on data mining (ICDM)*, pp. 1118–1123. IEEE, 2020.

Zheng Li, Yue Zhao, Xiyang Hu, Nicola Botta, Cezar Ionescu, and George H Chen. Ecod: Unsupervised outlier detection using empirical cumulative distribution functions. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12181–12193, 2022.

Shang-Ching Liu, ShengKun Wang, Tsungyao Chang, Wenqi Lin, Chung-Wei Hsiung, Yi-Chen Hsieh, Yu-Ping Cheng, Sian-Hong Luo, and Jianwei Zhang. Jarvix: A llm no code platform for tabular data analysis and optimization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 622–630, 2023a.

Tennison Liu, Zhaozhi Qian, Jeroen Berrevoets, and Mihaela van der Schaar. Goggle: Generative modelling for tabular data by learning relational structure. In *The Eleventh International Conference on Learning Representations*, 2023b.

Wei-Yin Loh. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(1):14–23, 2011.

Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Alex Labach, Jesse Cresswell, Keyvan Golestan, Guangwei Yu, Anthony L Caterini, and Maks Volkovs. Tabdpt: Scaling tabular foundation models on real data. *Advances in Neural Information Processing Systems*, 38:172692–172722, 2026.

Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

Jaehyun Nam, Jihoon Tack, Kyungmin Lee, Hankook Lee, and Jinwoo Shin. Stunt: Few-shot tabular learning with self-generated tasks from unlabeled tables. In *International Conference on Learning Representations (ICLR)*, 2023.

Duy-Kien Nguyen, Mahmoud Assran, Unnat Jain, Martin R Oswald, Cees G Snoek, and Xinlei Chen. An image is worth more than 16x16 patches: Exploring transformers on individual pixels. In *International Conference on Learning Representations*, volume 2025, pp. 59753–59773, 2025.

Guansong Pang, Longbing Cao, Ling Chen, and Huan Liu. Learning representations of ultrahigh-dimensional data for random distance-based outlier detection. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2041–2050, 2018.

Guansong Pang, Chunhua Shen, and Anton van den Hengel. Deep anomaly detection with deviation networks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 353–362, 2019.

Jonindo Pasaribu, Novanto Yudistira, and Wayan Firdaus Mahmudy. Tabular data classification and regression: Xgboost or deep learning with retrieval-augmented generation. *IEEE Access*, 12: 191719–191732, 2024.

Leto Peel, Tiago P Peixoto, and Manlio De Domenico. Statistical inference links data and theory in network science. *Nature Communications*, 13(1):6794, 2022.

Shourav B Rabbani, Ivan V Medri, and Manar D Samad. Attention versus contrastive learning of tabular data: a data-centric benchmarking. *International Journal of Data Science and Analytics*, 20(4):3069–3091, 2025.

Jonas Ranstam and Jonathan A Cook. Lasso regression. *Journal of British Surgery*, 105(10):1348–1348, 2018.

Ivan Rubachev, Artem Alekberov, Yury Gorishniy, and Artem Babenko. Revisiting pretraining objectives for tabular deep learning. *arXiv preprint arXiv:2207.03208*, 2022.

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*, pp. 4393–4402. PMLR, 2018.

Ali Shehadeh, Odey Alshboul, Rabia Emhamed Al Mamlook, and Ola Hamedat. Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, lightgbm, and xgboost regression. *Automation in Construction*, 129: 103827, 2021.

Tom Shenkar and Lior Wolf. Anomaly detection for tabular data with internal contrastive learning. In *International conference on learning representations*, 2022.

Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information fusion*, 81:84–90, 2022.

Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C Bayan Bruss, and Tom Goldstein. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv preprint arXiv:2106.01342*, 2021.

Jaime Lynn Speiser, Michael E Miller, Janet Tooze, and Edward Ip. A comparison of random forest variable selection methods for classification prediction modeling. *Expert systems with applications*, 134:93–101, 2019.

Marco Spinaci, Marek Polewczyk, Maximilian Schambach, and Sam Thelin. Contexttab: A semantics-aware tabular in-context learner. *Advances in Neural Information Processing Systems*, 38:147005–147040, 2026.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pp. 645–654, 2024.

Shuting Tao, Tongtian Zhu, Hongwei Wang, and Xiangming Meng. Semanticmask: A contrastive view design for anomaly detection in tabular data. In *IJCAI*, pp. 2370–2378, 2024.

Hugo Thimonier, José Lucas de Melo Costa, Fabrice Popineau, Arpad Rimmel, and Bich-Liên Doan. T-jepa: Augmentation-free self-supervised learning for tabular data. In *International Conference on Learning Representations*, volume 2025, pp. 58743–58768, 2025.

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.

Hoang Tran Vuong, Linh Ngo Van, Dang Nguyen, Thin Nguyen, Phuoc Nguyen, Mehrtash Harandi, and Trung Le. f-divergence self-play for tabular anomaly detection via large language models. In *Proceedings of the International Conference on Machine Learning*, 2026.

Wei-Yao Wang, Wei-Wei Du, Derek Xu, Wei Wang, and Wen-Chih Peng. A survey on self-supervised learning for non-sequential tabular data. *Machine Learning*, 114(1):16, 2025.

Jianan Ye, Zhaorui Tan, Yijie Hu, Xi Yang, Guangliang Cheng, and Kaizhu Huang. Disentangling tabular data towards better one-class anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 13061–13068, 2025.

Shi Ying, Bingming Wang, Lu Wang, Qingshan Li, Yishi Zhao, Jianga Shang, Hao Huang, Guoli Cheng, Zhe Yang, and Jiangyi Geng. An improved knn-based efficient log anomaly detection method with automatically labeled samples. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(3):1–22, 2021.

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.

Bing Zhou and Weiwei Chen. Enhancing ft-transformer with a matérn-driven kolmogorov-arnold feature tokenizer for tabular data-based in-bed posture classification. *IEEE Access*, 2025.

Bingzhao Zhu, Xingjian Shi, Nick Erickson, Mu Li, George Karypis, and Mahsa Shoaran. Xtab: Cross-table pretraining for tabular transformers. In *International Conference on Machine Learning*, pp. 43181–43204. PMLR, 2023.

A Appendix A: Theoretical Justifications

A.1 Justification of Dual-Driven Partitioning

The word orthogonally used above characterizes the two partitioning criteria, not the resulting feature-membership sets, S_{stat} and S_{sem} are not required, and in practice are not expected, to be disjoint. A feature pair such as systolic/diastolic pressure is legitimately co-clustered under both criteria, since it is simultaneously correlated and semantically related. This overlap is evidence of agreement, not a flaw. What we formalize below is the weaker, provable claim that the two criteria are extracted from two informationally independent channels.

Let $\mathcal{X} \in \mathbb{R}^{N\times M}$ be the observed data matrix drawn from an underlying joint distribution $P(\mathcal{X})$, and let $T = (T_1, ..., T_M)$ denote the column metadata fixed by the schema designer at table-creation time.

Schema-Data Independence (SDI): The metadata-generation process satisfies $P(T \mid \mathcal{X}) = P(T)$, column names and descriptions are fixed by table design and do not depend on which rows are later

sampled. This holds for every static tabular schema, including all benchmarks used in this paper. Under SDI, S_{sem} is statistically independent of the data realization $\mathcal{X}$, i.e., $I(S_{sem};\mathcal{X}) = 0$, whereas S_{stat} is a (generically non-degenerate) statistic of $\mathcal{X}$, i.e., $I(S_{stat};\mathcal{X}) > 0$ except in degenerate cases. By SDI, $T \perp \mathcal{X}$. Since the LLM encoder ϕ and the clustering operator h_{sem} are deterministic functions of T alone, $S_{sem} = h_{sem}(\phi(T))$ is a deterministic function of a variable independent of $\mathcal{X}$, hence $S_{sem} \perp \mathcal{X}$. Conversely, h_{stat} is a non-constant function of $C(\mathcal{X})$, which is itself a non-constant measurable function of $\mathcal{X}$ whenever $P(\mathcal{X})$ is non-degenerate. Hence S_{stat} carries strictly positive information about $\mathcal{X}$ in general.

This directly formalizes the motivating example: when a noisy batch yields spuriously low correlation between two clinically linked variables, S_{stat}'s signal is weakened because it is a function of that particular realization of $\mathcal{X}$, while S_{sem} is entirely unaffected, since it carries zero information from, and hence zero sensitivity to, any specific realization of $\mathcal{X}$. Under this framing, the role of the CVC loss is not to force S_{stat} and S_{sem} toward set-disjointness or representational redundancy elimination, but to let the shared latent space reach a consensus geometry that reinforces agreement between two independent evidence sources while VICReg's variance and covariance terms preserve each view's distinctive information.

A.2 Verification of the orthogonality assumption between the statistical subspace and the semantic subspace.

The Schema-Data Independence proposition we proposed in the methodology section establishes the independence of the partitioning criteria. It does not set an upper limit on their consistent occurrence frequency in a given dataset. We verify empirically that this independence does not collapse into S_{stat} and S_sem being a relabeling of the same partition on real data. Using the exact partitioning pipeline, we compute the pairwise Jaccard coefficient and Adjusted Rand Index (ARI) between S_{stat} and S_{sem} on all twelve datasets used from Tables. Table 4 reports the results.

Table 4: Feature-pair overlap between the statistical subspace S_{stat} and the semantic subspace S_{sem} across all benchmark datasets. Low Jaccard/ARI values indicate that the two partitions are not a relabeling of one another.

Dataset	M	#Stat	#Sem	Jaccard	ARI
Sleep	13	4	4	0.167	0.079
Heart	12	11	4	0.000	-0.029
Marketing	16	13	5	0.125	0.175
Bank	62	43	20	0.059	0.088
Donors	10	6	3	0.200	0.226
Credit	20	16	6	0.056	0.070
DMFT	4	3	2	0.500	0.571
CMC	9	6	3	0.053	-0.095
Salary	14	12	4	0.000	-0.042
LOS	23	20	7	0.018	0.012
Penguin	10	5	3	0.105	-0.049
DMC Heart	52	38	17	0.008	-0.007

Across the eleven datasets with $M \geq 9$, Jaccard ranges from 0.000 to 0.200 and ARI ranges from -0.095 to 0.226. Five of the eleven ARI values fall below zero. This means no more consistency than a random partition of the same cluster sizes. This confirms that S_{stat} and S_{sem} behave as two largely independent partitions rather than one criterion rediscovering the other. This is consistent with the Proposition. DMFT is the sole outlier. With only $M = 4$ features, the partition space is so small that pair-counting overlap statistics are high-variance and unreliable regardless of the underlying independence property.

B Appendix B: Further Methodological Details

B.1 Dual-driven View Partitioning and Shared Encoding

Dual-driven View Partitioning MVSTab transforms flattened tabular features into structured multimodal representations via multi-view partitioning. This process captures both explicit distributions and implicit logic. It partitions the feature space $\mathcal{F} = \{f_1, f_2, ..., f_M\}$ into data-driven statistical and knowledge-driven semantic views. These views are extracted along two informationally independent axes rather than as disjoint feature-membership sets.

First, we perform a base partitioning to separate continuous and discrete features. We assign a feature f_i to the numeric set $\mathcal{F}_{num}$ if it contains continuous real values and its cardinality exceeds a predefined threshold τ_{card} (e.g., 10 unique values). Otherwise, we assign the feature to the categorical set $\mathcal{F}_{cat}$.

To construct the statistical subspaces, the DVP computes the Pearson correlation coefficient matrix $C \in \mathbb{R}^{M \times M}$. Each element $C_{i,j}$ represents the linear correlation between f_i and f_j. We derive a distance matrix D as:

$$D_{i,j} = 1 - |C_{i,j}| \tag{9}$$

The DVP performs hierarchical clustering on D. It automatically groups highly correlated features into s statistical subspaces, denoted as $\mathcal{S}_{stat} = \{S^{(1)}_{stat}, S^{(2)}_{stat}, ..., S^{(s)}_{stat}\}$.

We use a pre-trained LLM to construct semantic subspaces. For each feature f_i, we generate a text prompt T_i using the column name and metadata. When obtaining word vectors, the DVP module does not simply pass in the column name. Instead, it constructs a prompt based on metadata: "The tabular dataset feature column name is *col*, which means *metadata*." We believe that providing a complete sentence produces more accurate semantic representations. The LLM processes T_i to output hidden states. We aggregate these via mean pooling over the sequence length L to produce a semantic embedding $E_i \in \mathbb{R}^{d_{llm}}$:

$$E_i = \frac{\sum_{l=1}^{L} w_l \cdot \text{LLM}(T_i)_l}{\max(\sum_{l=1}^{L} w_l, \varepsilon)} \tag{10}$$

, where w_l is the indicator of the attention mask. We apply K-Means clustering to the embeddings $\{E_1, ..., E_M\}$ to minimize the within-cluster sum of squares:

$$\arg\min_{\mathcal{S}_{sem}} \sum_{j=1}^{m} \sum_{E_i \in S^{(j)}_{sem}} ||E_i - \mu_j||_2^2 \tag{11}$$

This yields m semantic subspaces $\mathcal{S}_{sem} = \{S^{(1)}_{sem}, ..., S^{(m)}_{sem}\}$.

This dual-driven approach ensures that features with high statistical correlation form one type of perspective. Features sharing fundamental physical definitions form another, even if local noise reduces their statistical correlation.

B.1.1 Shared Multi-View Encoding

A shared encoding mechanism maps heterogeneous multi-view tabular features into a unified latent space. This prevents parameter explosion. The model extracts features across diverse perspectives. Weight sharing controls the parameter count and encourages the use of CLS prompt tokens for attention routing.

For numeric features $x_{num} \in \mathcal{F}_{num}$, we apply a linear projection with ReLU activation to map the scalar to a vector $v_i \in \mathbb{R}^{d_{model}}$:

$$v_i = \text{ReLU}(\mathbf{W}_i x_{num} + b_i) \tag{12}$$

For categorical features $x_{cat} \in \mathcal{F}_{cat}$, we use an embedding lookup table:

$$v_j = \text{Emb}(x_{cat}) \in \mathbb{R}^{d_{model}} \tag{13}$$

To distinguish different views during shared encoding, we initialize globally learnable prompt tokens (CLS) for each view $v \in \{\text{num}, \text{cat}, S^{(1)}_{stat} ... S^{(k)}_{stat}, S^{(1)}_{sem} ... S^{(m)}_{sem}\}$. We formulate the input sequence for a

specific view Z_v by concatenating the CLS token with the corresponding feature tokens:

$$Z_v = [\text{CLS}_v, v_1, v_2, ..., v_{|v|}] \tag{14}$$

A shared Transformer encoder processes Z_v. We extract the output vector corresponding to the position of CLS_v as the view-specific representation h_v:

$$h_v = \text{Transformer}(Z_v)[0] \tag{15}$$

B.2 Within-View Contrastive (WVC) Loss

For a given view, we contrast the original representation h_v with its augmented counterpart h'_v. For a batch of representations Z and Z', we calculate the variance $v(Z)$, invariance $s(Z,Z')$, and covariance $c(Z)$ as follows:

$$s(Z,Z') = \frac{1}{B}\sum_{i=1}^{B} ||z_i - z'_i||_2^2 \tag{16}$$

$$v(Z) = \frac{1}{d_{model}}\sum_{j=1}^{d_{model}} \max\left(0, 1 - \sqrt{\text{Var}(Z_{\cdot,j}) + \varepsilon}\right) \tag{17}$$

$$c(Z) = \frac{1}{d_{model}}\sum_{i\neq j}\left(\frac{(Z-\bar{Z})^T(Z-\bar{Z})}{B-1}\right)_{i,j}^2 \tag{18}$$

$$\begin{aligned}\mathcal{L}_{WVC} = \frac{1}{|V|}\sum_{v\in V}\Big(&\lambda s(h_v, h'_v) + \mu\big(v(h_v) + v(h'_v)\big) \\ &+ \nu\big(c(h_v) + c(h'_v)\big)\Big)\end{aligned} \tag{19}$$

We introduce a small amount of noise to the raw data from all perspectives. Invariance pulls the resulting noisy vectors closer to their original, pre-noise counterparts. This enables the model to disregard superficial noise and grasp the underlying essence of the data. However, if the model were driven solely by this Invariance-based attractive force, it would tend to output an absolute zero for every sample. This results in perpetual zero error (a phenomenon known as Representation Collapse). Variance pushes apart the representations of distinct samples. It compels them to disperse uniformly throughout the latent space. Covariance enforces orthogonality across different dimensions. This prevents information redundancy.

B.3 Latent Cross-Attention Fusion

Let $H_{all} \in \mathbb{R}^{N\times V\times d_{model}}$ denote the representations of all global and subspace views, where V is the number of views. We initialize a set of $K_{latents}$ learnable latent queries $Q_{latent} \in \mathbb{R}^{K_{latents}\times d_{model}}$. The number K is fixed across datasets. Each latent query attends to all view representations and extracts a different combination of view-level information. We then use multi-head cross-attention to update the latent queries:

$$\begin{aligned}&Q = Q_{latent}\mathbf{W}_Q, K = H_{all}\mathbf{W}_K, U = H_{all}\mathbf{W}_V \\ &H_{latent} = \text{Softmax}(\tfrac{QK^T}{\sqrt{d_{model}}})U\end{aligned} \tag{20}$$

We process the output through a feed-forward network and flatten it to generate the final global representation

$$h_{fused} = Flatten(FFN(H_{latent})) \in R^{N\times(K_{latents}\cdot d_{model})} \tag{21}$$

B.4 Adaptation to Downstream Tasks

Following self-supervised pre-training, MVSTab yields a dense, logically consistent global representation h_{fused}. The latent queries mechanism fixes this representation dimension at $K_{latents}\times d_{model}$. This completely decouples it from the original feature count. This property ensures seamless compatibility across diverse downstream applications.

Unsupervised Anomaly Detection: Unsupervised anomaly detection eliminates requirements for external neural networks and labels. MVSTab representations integrate directly into existing machine learning or deep learning anomaly detection algorithms. They serve as inputs for density-based, isolation-based, or distribution-based methods. Migrating the operational space of these external algorithms into the MVSTab latent space enhances their sensitivity to underlying structural anomalies. This improves overall detection robustness.

Supervised Classification and Regression: Extracted global representations serve as high-level features for tabular machine learning algorithms like XGBoost and LightGBM. In gradient-free mode, extracting and decoupling training and test set representations from the computational graph allows utilizing these dense vectors directly as model inputs. MVSTab combines with deep networks by attaching a fully connected neural network under two distinct training paradigms. Linear probing strictly freezes the shared backbone parameters and optimizes only the MLP weights to evaluate the inherent discriminative power of the pre-trained representations. Fine-tuning jointly updates the entire architecture using a minimal learning rate. It optimizes classification tasks via Binary Cross-Entropy or standard Cross-Entropy loss and regression tasks via Mean Squared Error loss.

Additionally, MVSTab integrates with neural network architectures like TabNet or FT-Transformer through a residual enhancement strategy. Letting h_{SOTA} denote the internal representation generated by a state-of-the-art neural network, the MVSTab fused representation h_{fused} projects via a transformation matrix $\mathbf{W}_p$ and adds to h_{SOTA}:

$$h_{final} = \text{Activation}(h_{SOTA} + \mathbf{W}_p h_{fused}) \tag{22}$$

This integration allows the external model to leverage knowledge-guided semantic priors captured by MVSTab, acting as a pre-trained backbone providing a warm start for complex deep learning models.

B.5 Computational Complexity Analysis

In this section, we analyze the computational efficiency of the MVSTab framework. Let B denote the batch size, M the total number of features, d the model dimension (d_model), and V the total number of views ($V = 1 + 1 + k + m$, representing numeric, categorical, statistical, and semantic views, respectively).

B.5.1 Preprocessing Complexity

The preprocessing stage involves view partitioning. We perform this only once before the training process.

- **Statistical Partitioning:** Computing the Pearson correlation matrix for M features over the entire dataset of size N takes $O(M^2N)$. The subsequent agglomerative clustering with a distance-based threshold typically requires $O(M^2 \log M)$.
- **Semantic Partitioning:** Generating feature embeddings using a pre-trained LLM requires $O(M)$ forward passes. Since the LLM is frozen and used offline, this is a constant overhead. The K-Means clustering of M embeddings in d_{llm} space takes $O(iter \cdot m \cdot M \cdot d_{llm})$. Here $iter$ is the number of iterations and m is the number of semantic clusters.

B.5.2 Training and Inference Complexity

The training and inference complexity is dominated by the shared encoding and the latent fusion mechanism.

- **Shared Multi-View Encoding:** Conventional Tabular Transformers process all M features in a single sequence. This yields $O(M^2 d)$ complexity. MVSTab partitions features into V smaller subspaces. For a subspace i with n_i features, the Transformer encoding cost is $O(B \cdot n_i^2 \cdot d)$. Since $\sum n_i \approx M$ and subspaces are typically small ($n_i \ll M$), the total encoding complexity is approximately $\sum_{i=1}^{V} O(B \cdot n_i^2 \cdot d)$. This is significantly lower than $O(B \cdot M^2 \cdot d)$. This sparse connectivity pattern effectively reduces the quadratic pressure of the self-attention mechanism.

- **Multi-Objective Optimization:** The VICReg-based WVC and CVC losses operate on the CLS tokens of each view. For V views, the complexity of computing variance, invariance, and covariance is $O(B \cdot V \cdot d + V \cdot d^2)$. The HNC Triplet Loss involves m semantic views. This incurs $O(B \cdot m \cdot d)$. These losses are linear with respect to the batch size and the number of views.
- **Latent Cross-Attention Fusion:** The fusion module employs K latent queries to attend to V view representations. The computational cost of the Multi-Head Cross-Attention (MHCA) is:

$$T_{fusion} = O(B \cdot K \cdot V \cdot d) \tag{23}$$

Since K is a small constant (e.g., $K = 4$) and is independent of M, the fusion complexity is linear with respect to the number of views V.

The total online complexity per iteration is approximately $O(B \cdot d \cdot (\sum n_i^2 + KV))$. Given that n_i is small and $K \ll M$, MVSTab demonstrates favorable scalability. The framework circumvents the $O(M^2)$ bottleneck of global attention by decomposing the feature space into semantically and statistically coherent groups. This makes it suitable for high-dimensional tabular datasets where M is large.

C Appendix C: Experimental Setup Details

D Dataset and Baselines

The following section summarizes datasets from different domains and applied to various downstream tasks. It also introduces baseline models used for comparison in different tasks.

For anomaly detection, we utilize the Gigasheet Sleep dataset assessing 374 samples and 14 features, defining anomalies as individuals experiencing extreme mental stress based on sleep and lifestyle indicators. The Heart dataset Chicco & Jurman (2020) includes 300 samples and 13 cardiovascular features, targeting heart failure deaths. The Marketing dataset Moro et al. (2014) contains 11,162 samples and 17 financial institution features, identifying failed marketing campaigns. We additionally evaluate the Bank and Donors datasets Pang et al. (2019) containing 41,189 samples with 63 features and 619,326 samples with 10 features, respectively. Baseline comparisons include KNN Ying et al. (2021), IForest Laskar et al. (2021), DeepSVDD Ruff et al. (2018), COPOD Li et al. (2020), ECOD Li et al. (2022), REPEN Pang et al. (2018), HBOS Goldstein & Dengel (2012), GOAD Bergman & Hoshen (2020), ICL Shenkar & Wolf (2022), SCARF Bahri et al. (2021), SubTab Ucar et al. (2021), VIME Yoon et al. (2020), and SemanticMask Tao et al. (2024).

Supervised classification tasks employ the Credit dataset Bischl et al. (2021) with 1,000 samples and 21 features for binary classification. The DMFT dataset Bischl et al. (2021) comprises 797 samples and 5 features for 6-class classification, while the CMC dataset Bischl et al. (2021) includes 1,473 samples and 10 features for 3-class classification. The Salary dataset Kohavi (1996) contains 48,842 samples and 14 features predicting annual incomes exceeding 50,000. Additional classification baselines include LR Cox (1958), CART Loh (2011), RF Speiser et al. (2019), XGBoost Chen & Guestrin (2016), TabNet Arik & Pfister (2021), FT-Transformer Zhou & Chen (2025), SAINT Somepalli et al. (2021), Trompt Chen et al. (2023), AMFormer Cheng et al. (2024), TabPFN Hollmann et al. (2025), and ResNet-based T-JEPA Thimonier et al. (2025).

Supervised regression evaluates sleep duration prediction using the Gigasheet Sleep dataset. The LOS dataset Barga et al. (2015) comprises 100,000 samples and 24 features forecasting hospital stay duration. The Penguin dataset Gorman et al. (2014) includes 344 samples and 11 features targeting weight information. The DMC Heart dataset Bollepalli et al. (2022) contains 14,390 samples and 50 features predicting hospital stay duration. Regression baselines add Lasso Ranstam & Cook (2018), LightGBM Shehadeh et al. (2021), NAM Agarwal et al. (2021), and TabResNet Kadra et al. (2024).

D.1 Experiment Environment

All experiments were conducted on 4090 GPUs. The framework is implemented using PyTorch. During the semantic subspace partitioning stage of different datasets, we used Mpnet-v2, BGE-M3 and gemini-embedding2. The hyperparameter configurations are detailed below. The latent

representation dimension for all feature tokens and view CLS tokens was set to 32. For the cross-attention-based fusion module, the number of latent queries was set to 4. For statistical subspace clustering, the Pearson correlation distance threshold was set to 0.7. For semantic clustering, the number of clusters was dynamically determined as $m = max(2, M/3)$, where M is the number of features. The batch size was set to 32. The initial learning rate was e^{-4}. In the HNC loss, the triplet margin was set to 1. In the WVC and CVC losses, the weights for invariance, variance, and covariance terms were set to 25, 25, and 1, respectively. α, β, γ were set to 1,1,2.

E Appendix D: Additional Experimental Results

E.1 Performance of MVSTab on classification tasks

Table 5 presents the performance of MVSTab and 11 baseline methods in supervised classification tasks. The best results for each metric are highlighted in bold. We combine MVSTab with a residual network to accomplish this task.

Table 5: The performance of MVSTab and baseline methods on supervised classification tasks, the best result for each metric is highlighted in bold.

Dataset	Credit		DMFT		CMC		Salary	
Metrics	Accuracy	AUROC	Accuracy	AUROC	Accuracy	AUROC	Accuracy	AUROC
LR	0.7004	0.7323	0.2125	0.5398	0.5254	0.7097	0.7874	0.8021
CART	0.6809	0.6429	0.1812	0.4823	0.4101	0.5488	0.7684	0.7021
RF	0.7450	0.7741	0.1625	0.4949	0.4813	0.6749	0.7916	0.8164
KNN	0.6851	0.6448	0.1812	0.5217	0.4617	0.6343	0.8190	0.8498
XGBoost	0.7650	0.7692	0.1937	0.5658	0.5220	0.7299	0.8484	0.8023
TabNet	0.7267	0.7856	0.2583	0.6021	0.5792	0.7473	0.8487	0.8917
FT-Transformer	0.7485	0.7366	0.2195	0.5907	0.5468	0.7402	0.8359	0.8931
SAINT	0.7698	0.8135	0.2247	0.5923	0.5473	0.7396	0.8790	0.8410
Trompt	0.7462	0.6150	0.2083	0.5820	0.4975	0.6475	0.8692	0.8736
AMFormer	0.7011	0.6061	0.1833	0.5691	0.4977	0.6497	0.8205	0.8924
TabPFN	0.7403	0.7894	0.2252	0.5756	0.5475	0.7276	0.8744	0.8931
T-JEPA	0.7547	0.7985	0.2326	0.5848	0.4949	0.6970	0.8642	0.8881
MVSTab	**0.7750**	**0.8219**	**0.2652**	**0.6155**	**0.5831**	**0.7621**	**0.8816**	**0.8986**

The results shown in Table 5 demonstrate the discriminative advantage of our proposed multi-view representation for supervised classification tasks. Across all datasets, our proposed ResNet-based classifier outperforms state-of-the-art tabular architectures, including Trompt, AMFormer, TabPFN, and T-JEPA. This indicates that aligning data-driven statistical views with knowledge-driven semantic views during pre-training successfully unwinds complex feature dependencies. It creates a highly separable feature space that inherently benefits downstream classification boundaries without requiring exhaustive task-specific feature engineering.

E.2 Performance of MVSTab on regression tasks

Table 6 presents the performance of MVSTab and 11 baseline methods in supervised regression tasks. The best results for each metric are highlighted in bold. We combine MVSTab with a residual network to accomplish this task.

Table 6 demonstrates the adaptability of the learned representations when paired with downstream prediction heads. The proposed framework achieves the lowest mean absolute error and root mean squared error on the Penguin, LOS and DMC Heart datasets. This surpasses both native gradient boosting trees and recent deep tabular models like T-JEPA and TabNet. On the extremely small-scale Sleep dataset, the traditional random forest model marginally outperforms the proposed approach. This outcome aligns with theoretical expectations. Ensemble decision trees possess inherent resistance to overfitting on minimal sample sizes. Deep multi-view representations require a sufficient volume of data to adequately capture and extract generalizable feature interactions.

Table 6: The performance of MVSTab and baseline methods on supervised regression tasks, the best result for each metric is highlighted in bold.

Dataset	Sleep		LOS		Penguin		DMC Heart	
Metrics	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Lasso	0.723	0.825	1.486	1.826	0.448	0.533	2.142	3.254
CART	0.056	0.091	0.768	1.256	0.339	0.438	3.008	4.928
RF	**0.055**	**0.074**	0.651	0.932	0.239	0.302	2.097	3.256
KNN	0.071	0.117	1.034	1.454	0.267	0.320	2.887	4.248
XGBoost	0.056	0.079	0.463	0.653	0.239	0.300	2.035	3.252
LightGBM	0.064	0.087	0.453	0.644	0.240	0.296	2.216	3.435
TabNet	0.196	0.396	0.849	1.147	1.721	2.013	2.219	3.296
FT-Transformer	0.090	0.134	0.547	0.753	0.260	0.312	2.219	3.296
NAM	0.178	0.253	0.554	0.763	0.304	0.383	2.129	3.298
TabResNet	0.198	0.226	0.733	1.038	0.344	0.433	2.389	3.872
TabPFN	0.172	0.204	0.875	1.127	0.274	0.343	2.623	4.054
T-JEPA	0.167	0.213	0.784	0.964	0.216	0.251	2.168	3.548
MVSTab	0.065	0.095	**0.424**	**0.636**	**0.171**	**0.205**	**1.857**	**3.143**

E.3 Parameter sensitivity analysis

We investigate the sensitivity of MVSTab to the following hyperparameters: the selection of LLM, the Quality of Metadata, the Pearson threshold of the statistical subspace, and the value of m in the semantic subspace. We used the Bank dataset from the AD task as an illustration, and the results are shown in Figure 4.

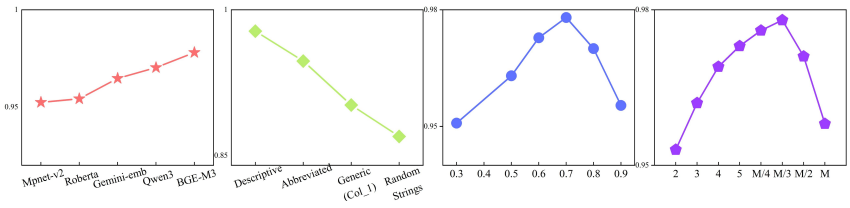

Figure 4: Sensitivity analysis of the four hyperparameters in MVSTab.

The overall effects of Gemini-embedding2, Qwen3, and BGE-M3 are similar and better than those of Mpnet-v2 and Roberta. Among them, BGE-M3 is the most accurate in constructing the semantic coordinate system. The Pearson threshold shows an inverted U-shape, with the performance peaking at 0.7 in the middle. Low thresholds will introduce incorrect aggregation, while high thresholds will regress to a weak structure state, resulting in a decrease in accuracy. At the same time, features with complete descriptions will greatly improve the semantic segmentation accuracy, thereby affecting the final performance. The more the feature names are blurred, the greater the performance will decline. The number of semantic segmentations also shows an inverted U-shape, with the peak at M/3. A too small m will destroy the semantic guidance in the coarse-grained subspace. Excessive refinement will make it difficult for LLM to distinguish.