

## Article

# GINet-DGC: Structural Inductive Biases and Dynamic Generalization Control for High-Dimensional Small-Sample Tabular Data

Xinran Zhang <sup>1,2</sup> , Yang Sheng <sup>3</sup>, Sijie Shen <sup>3</sup>, Dongjie Fan <sup>1,4</sup> and Lizhuang Liu <sup>1,2,\*</sup> 

<sup>1</sup> Department of Signal and Information Processing, Shanghai Advanced Research Institute, Chinese Academy of Sciences, Haik Road, Shanghai 201210, China; zhangxr2024@sari.ac.cn (X.Z.); fandongjie2025@sari.ac.cn (D.F.)

<sup>2</sup> University of Chinese Academy of Sciences, Yuquan Road, Beijing 100049, China

<sup>3</sup> General Mathematical-Physics Technology, Zhangheng Road, Shanghai 201210, China; shengyang@gmpt.com.cn (Y.S.); shensijie@gmpt.com.cn (S.S.)

<sup>4</sup> School of Information and Communication Engineering, North University of China, Jiancaoping District, Taiyuan 030051, China

\* Correspondence: liulz@sari.ac.cn

† <https://peopleucas.ac.cn/~0017613>.

## Abstract

Learning from high-dimensional, low-sample-size (HDLSS) data remains a persistent challenge in machine learning, as models must infer reliable patterns from limited observations while handling an excessive number of variables—a scenario particularly prevalent in biomedical applications. Such data structures render predictive modeling highly vulnerable to erratic optimization and overfitting. To address this challenge, we propose the Global Interaction Network with Dynamic Generalization Control (GINet-DGC), an artificial intelligence (AI) framework that integrates feature-wise structural priors with dynamic generalization monitoring. Rather than directly learning an unconstrained first-layer weight matrix, GINet-DGC generates task-specific weights from multi-view feature descriptors, encompassing latent semantic, global distributional, local topological, and hierarchical representations. This structure-constrained weight generation strategy effectively narrows the feature-interaction search space and acts as an inductive regularizer against noise and redundant molecular features. Furthermore, we introduce an Overfitting-aware Index (OFI) to monitor the training trajectory and effectively identify the generalization saturation point for adaptive termination. Empirical evaluations on eight public real-world biomedical HDLSS gene-expression datasets, using a repeated stratified  $5 \times 5$  cross-validation protocol, demonstrate that GINet-DGC achieves competitive and stable performance against 17 baselines. These findings support the effectiveness of the proposed framework within the evaluated public biomedical HDLSS benchmark setting.

**Keywords:** high-dimensional low-sample size; tabular representation learning; structure-constrained weight generation; dynamic generalization control



Academic Editors: Miguel Angel Cazorla and Gabriella Trucco

Received: 5 June 2026

Revised: 17 July 2026

Accepted: 24 July 2026

Published: 6 August 2026

**Copyright:** © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

## 1. Introduction

This section presents a general overview of the study by outlining the background, reviewing related works, identifying the research gaps, and summarizing the main contributions of this research.

### 1.1. Background and Motivation

Tabular data represent one of the most widely used data formats in scientific computation and industrial analysis due to their inherent capacity to represent heterogeneous observations through structured variables. Indeed, they serve as a dominant modality across numerous cutting-edge domains, playing a critical role in supporting advanced research in high-energy physics [1], computational chemistry [2], and financial forecasting [3]. Beyond these fields, structured variables are particularly crucial for predictive modeling and knowledge discovery in biomedicine and health-oriented artificial intelligence [4,5]. Despite their ubiquity and critical importance, identifying the optimal predictive model for such structured data remains an outstanding open question [6]. However, across many of these scientific disciplines, informative datasets are difficult to collect because experiments can be costly, time-consuming, or ethically constrained.

For example, in experimental particle physics, high-dimensional detector measurements are produced under prohibitively expensive acquisition conditions [7]. This data scarcity becomes even more severe in biomedical gene-expression analysis, where typical transcriptomic profiles contain tens of thousands of molecular features against merely dozens or hundreds of clinical patient samples [8]. This severe dimensional discrepancy precisely defines the high-dimensional, low-sample-size (HDLSS) nature of the data. Consequently, researchers must infer reliable patterns under an extreme regime where the feature dimension vastly exceeds the number of available observations [9,10]. In this scenario, limited samples provide insufficient supervision for the massive parameter space of deep learning models. This bottleneck inherently renders traditional networks vulnerable to stochastic noise, feature redundancy, and unstable gradients, thereby leading to severe overfitting.

Achieving effective inference from such data is thus a pressing challenge in modern machine learning and a critical milestone toward empirical scientific discovery. In this study, we focus on biomedical gene-expression classification, a prototypical benchmark that exemplifies the complexities of the HDLSS learning scenario.

### 1.2. Related Works

This section reviews the existing literature to clarify the scope of current paradigms and highlight the open challenges that motivate this study.

In the context of small-sample tabular learning, traditional tree-based ensemble methods and shallow models with regularization have served as competitive baseline approaches in many benchmark studies [10]. Representative models such as Extreme Gradient Boosting (XGBoost) [11] and Light Gradient Boosting Machine (LightGBM) [12] achieve good generalization through efficient feature splitting and regularized ensemble learning. This persistent advantage is frequently attributed to the lack of explicit inductive biases in general-purpose deep neural networks when handling heterogeneous tabular variables—a stark contrast to the robust spatial or temporal priors embedded in convolutional or recurrent architectures [6,13]. Furthermore, systematic empirical evaluations suggest that the advantages of neural networks over boosted trees are more likely to be realized in data-rich regimes with prominent and structured feature interactions [14]. However, tree-based models cannot perform end-to-end representation learning, which restricts their ability to capture complex, high-order dependencies among molecular features. To address these limitations, recent deep tabular research primarily focuses on three directions: (1) feature selection-driven representation enhancement, (2) structured parameter generation and conditional weight prediction, and (3) data distribution expansion and prior transfer.

**Feature selection-driven representation enhancement methods.** The core philosophy of this paradigm is to dynamically modulate feature contributions at the input or intermedi-

ate layers through explicit or implicit selection and weighting mechanisms, thereby refining representation quality and mitigating overfitting. For instance, AutoInt [15] captures high-order feature interactions by introducing explicit interaction layers to strengthen feature learning. TabNet [16] employs a sequential attention mechanism for sparse feature selection, achieving an elegant balance between interpretability and performance. Attention-based architectures, such as TabTransformer [17], Feature Tokenizer Transformer (FT-Transformer) [10], and Self-Attention and Intersample Attention Transformer (SAINT) [18], further expand modeling capacity by mapping heterogeneous features into a unified embedding space. To directly combat dimensionality, FsNet [19] introduces a learnable feature selection layer governed by sparsity constraints to identify salient variables and reduce parametric complexity. While these approaches exhibit robust performance competitive with tree-based models on multiple small-to-medium-scale benchmarks, their efficacy often degrades in ultra-high-dimensional biomedical contexts. Because these methods primarily operate via feature-level gating or pooling, they remain inherently vulnerable to dense feature redundancy and the curse of dimensionality, which ultimately compromises their optimization stability when deployed on data-scarce omics datasets.

#### **Structured Parameter Generation and Conditional Weight Prediction Mechanisms.**

This paradigm introduces parametric dependencies or conditional generation networks into model architectures to contract the space of free learnable parameters and explicitly inject structural priors, thereby enhancing generalization and optimization stability. This concept stems from the HyperNetworks framework [20]. In the domain of tabular learning, HyperTab [21] generates dynamic weights conditioned on feature subsets to mitigate parameter redundancy and combat overfitting. Similarly, DietNetworks [22] employ a weight prediction network to derive a compact parameterization scheme, improving robustness in small-sample regimes. The Weight Predictor Network with Feature Selection (WPFS) [23] bridges feature embeddings with a weight predictor, mapping structural dependencies among features directly into the first-layer weight matrix. More recently, GCondNet [24] uses graph-conditioned feature representations for parameter generation, demonstrating the value of feature-relation modeling in high-dimensional small-sample tasks. Despite these advancements, existing frameworks typically operate under preset training schedules, which presents opportunities to further incorporate real-time overfitting detection and adaptive model modulation. Furthermore, when scaling to ultra-high-dimensional feature spaces, managing the delicate balance between architectural expressiveness and optimization stability remains an active area of investigation, encouraging further refinement for robust HDLSS modeling.

The proposed Global Interaction Network with Dynamic Generalization Control (GINet-DGC) inherits the general paradigm of conditional weight generation from DietNetworks, WPFS, HyperTab, and GCondNet, and its use of feature-wise embeddings for first-layer weight prediction is particularly inspired by WPFS. Its novelty lies not in weight prediction itself, but in constructing a multi-view conditioning representation and combining it with Overfitting-aware Index (OFI)-based training-trajectory control. Table 1 summarizes these differences.

**Data distribution expansion and prior transfer strategies.** This class of methods aims to alleviate high variance and overfitting under small-sample conditions by expanding the training data distribution or incorporating external prior knowledge. Traditional interpolation-based approaches, such as the Synthetic Minority Over-sampling Technique (SMOTE) [25] and Adaptive Synthetic Sampling (ADASYN) [26], generate synthetic minority samples through neighborhood interpolation. In recent years, generative architectures have emerged as a prominent avenue for tabular data augmentation. For instance, the Conditional Tabular Generative Adversarial Network (CTGAN) [27] introduces conditional

generation mechanisms tailored for mixed variables, while the diffusion-based Tabular Denoising Diffusion Probabilistic Model (TabDDPM) [28] leverages score-based priors to enhance synthesis quality. However, in high-dimensional biomedical domains, synthetic augmentation can potentially amplify stochastic noise or inadvertently distort domain-specific molecular distributions. Alternatively, prior transfer strategies, such as the Tabular Prior-Data Fitted Network (TabPFN) [29] and large-scale language model transfer [30], attempt to mitigate data scarcity through cross-task knowledge migration. Recent discussions on tabular foundation models further suggest that large-scale pre-training and prior-data fitted networks can instill adaptable tabular priors across domains [31]. Nevertheless, the efficacy of these cross-domain strategies is often mediated by task consistency, as they can be susceptible to prior mismatch when deployed across heterogeneous biomedical cohorts and varying assay conditions.

In summary, while tabular learning has progressed significantly, it still faces severe challenges in high-dimensional, low-sample-size scenarios. The primary bottlenecks include high model complexity, premature overfitting, limited stability under distribution shifts, and a lack of dynamic adaptation to training states.

**Table 1.** Methodological comparison between the Global Interaction Network with Dynamic Generalization Control (GINet-DGC) and related conditional weight-generation methods.

| Method           | Conditioning Source                       | Structural Modeling                                             | Training-Trajectory Control         |
|------------------|-------------------------------------------|-----------------------------------------------------------------|-------------------------------------|
| DietNetworks     | Feature embeddings                        | Feature-specific compact parameterization                       | No method-specific trajectory index |
| WPFS             | Learned feature embeddings                | Feature-selection-aware representation                          | No method-specific trajectory index |
| HyperTab         | Feature subsets                           | Subset-level context                                            | No method-specific trajectory index |
| GCondNet         | Graph-conditioned feature representations | Feature-graph relations                                         | No method-specific trajectory index |
| GINet-DGC (Ours) | Fused multi-view descriptors              | Latent, distributional, topological, and hierarchical structure | OFI-based DGC                       |

Abbreviations: WPFS, Weight Predictor Network with Feature Selection; GINet-DGC, Global Interaction Network with Dynamic Generalization Control; OFI, Overfitting-aware Index; DGC, Dynamic Generalization Control.

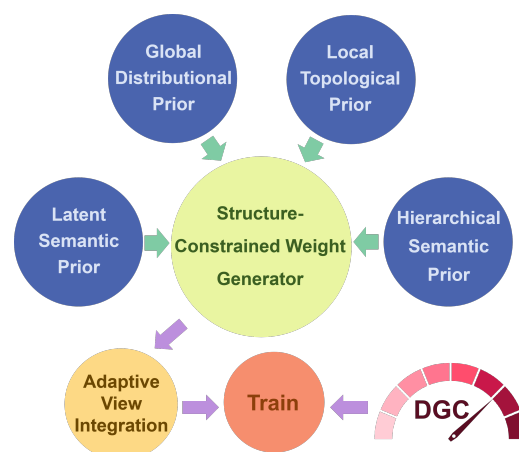
1.3. Research Gap

Although deep learning has achieved major success across various domains, whether standard deep networks constitute the optimal choice for high-dimensional, low-sample-size (HDLSS) tabular data remains a subject of ongoing debate. While deep models possess the capacity to learn flexible representations and capture intricate feature interactions, their performance frequently degrades in data-scarce or sparse regimes. For instance, Liu et al. [32] observed that standard deep neural networks (DNNs) can exhibit unstable convergence under limited samples, primarily because the feature dimension vastly exceeds the available observations. This severe dimensional imbalance often induces excessive gradient variance and impairs generalization. Similarly, Margeloiu et al. [23,24] reported that conventional neural architectures are prone to optimization difficulties on HDLSS datasets, as noisy and redundant features heavily interfere with the training trajectory. More broadly, while Shwartz-Ziv and Armon [9] argued that classical tree-based ensembles like XGBoost often outperform deep learning alternatives on tabular tasks, recent empirical evidence demonstrates that carefully constrained architectures can remain highly competitive under appropriate configurations [10]. This ongoing divergence underscores that the central challenge is not whether deep learning is applicable, but how its expressive flexibility can be effectively regularized when supervision is scarce. For omics-based biomedical tables, incorporating such constraints is particularly vital, as features represent molecular entities bound by latent biological and statistical dependencies rather than merely functioning as independent coordinates.

Despite recent exploration into parameter generation, robust neural modeling for biomedical HDLSS data still faces severe challenges. First, excessive parametric dimensions frequently induce premature overfitting. Second, pervasive molecular redundancies tend

to obscure stable, disease-related predictive patterns. Third, static training paradigms lack the capacity to monitor real-time training trajectories. These persistent limitations define the central research challenges addressed in this study. Specifically, current networks fail to explicitly constrain the feature-interaction search space while simultaneously identifying overfitting dynamically during optimization.

Therefore, we propose GINet-DGC, a neural framework engineered with structural inductive biases and adaptive trajectory-based early stopping. As illustrated in Figure 1, instead of fitting an unconstrained projection directly from scarce observations, GINet-DGC employs a Structure-Constrained Weight Generator (SCWG). This module synthesizes task-specific weights from multi-view feature embeddings to narrow the interaction search space, thereby facilitating reasoning over functional variable groups rather than treating features as isolated dimensions. Simultaneously, the Dynamic Generalization Control (DGC) strategy monitors the training trajectory via an Overfitting-aware Index (OFI) to determine adaptive termination. As shown in Figure 2, empirical comparisons across multiple benchmarks demonstrate that GINet-DGC yields competitive and robust outcomes under data-scarce regimes.



**Figure 1.** Data flow of the Global Interaction Network with Dynamic Generalization Control (GINet-DGC) framework. Raw high-dimensional features are transformed into four feature-wise structural priors: latent semantic, global distributional, local topological, and hierarchical representations. These priors are adaptively fused to condition the Structure-Constrained Weight Generator, while Dynamic Generalization Control monitors the training trajectory through the Overfitting-aware Index and terminates optimization when generalization saturates.

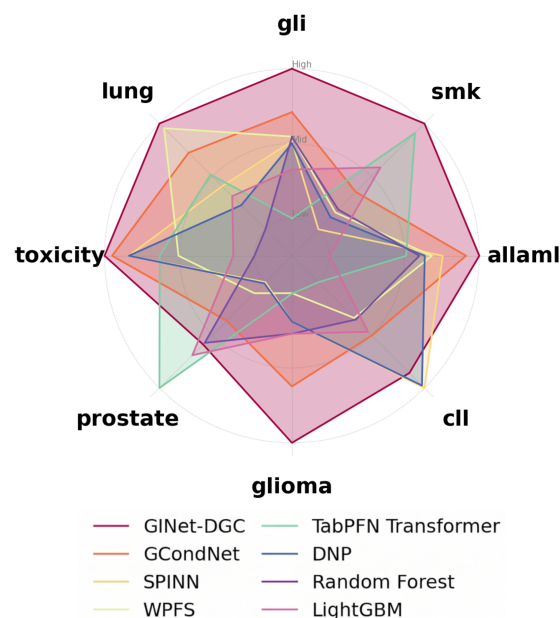
#### 1.4. Main Contributions

The main contributions of this research are summarized as follows:

- **Framework Proposal:** We introduce GINet-DGC, an end-to-end framework tailored with structural inductive biases to resolve the strict sample constraints in biomedical HDLSS tasks.
- **Weight Generation Mechanism:** We introduce a structurally constrained generator designed to generate the first-layer projection matrix by adaptively fusing latent semantic, global distributional, local topological, and hierarchical representations. This mechanism sets GINet-DGC apart from traditional conditional weight-generation approaches that rely solely on a single source.
- **Adaptive Early Stopping:** We implement an Overfitting-aware Index (OFI) alongside an adaptive training strategy to monitor generalization behavior and prevent overfitting during optimization.

- **Extensive Empirical Validation:** We comprehensively evaluate the proposed framework on eight public gene-expression datasets using repeated stratified  $5 \times 5$  cross-validation against 17 baseline models, demonstrating competitive stability.

The remainder of this paper is organized as follows: Section 2 presents the materials and methods, Section 3 reports the results, Section 4 discusses the findings, and Section 5 concludes the paper.



**Figure 2.** Performance comparison of the top eight models ranked by balanced accuracy. The radar chart illustrates the normalized accuracy of the top eight models across multiple biomedical datasets. Overall, GINet-DGC demonstrates competitive performance across the evaluated datasets, indicating stable generalization under the considered biomedical high-dimensional, low-sample-size (HDLSS) setting.

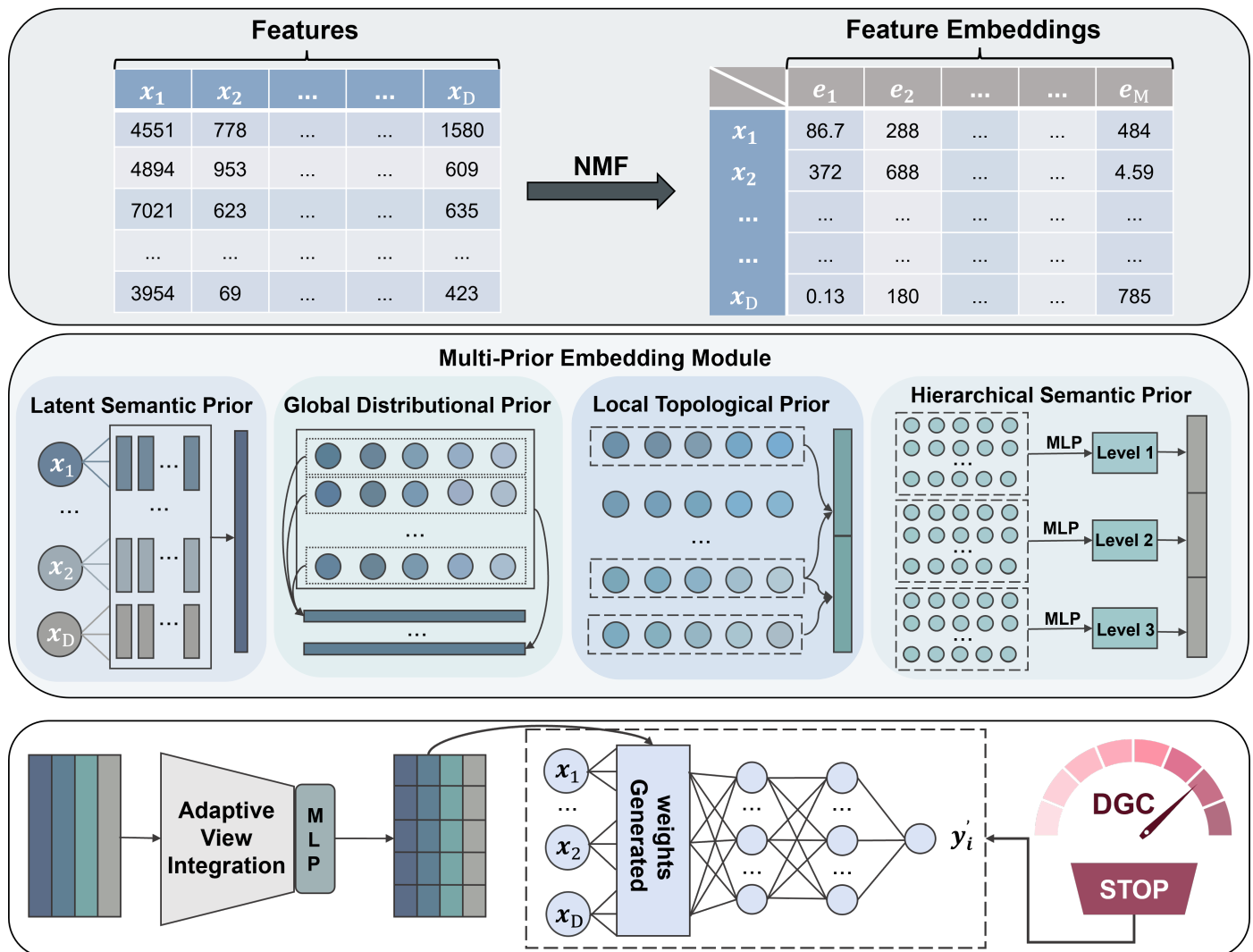
## 2. Materials and Methods

The proposed GINet-DGC framework addresses the challenge of learning from biomedical high-dimensional, low-sample-size data. In these scenarios, molecular features greatly outnumber available clinical observations, meaning that increasing model depth blindly risks severe overfitting. Therefore, our approach introduces structural inductive biases to constrain the parameter solution space. We propose a unified framework that integrates a Structure-Constrained Weight Generator and a Dynamic Generalization Control adaptive stopping strategy.

Figure 3 outlines the complete methodological workflow of this research, and Sections 1.3, 2.2–2.5 provide detailed descriptions of the proposed framework, training process, and experimental protocol.

### 2.1. Model Overview

The framework consists of two core components: the Structure-Constrained Weight Generator, which integrates multi-view structural priors from high-dimensional biomedical features and synthesizes adaptive weight representations; Dynamic Generalization Control, an adaptive early stopping strategy that monitors the training trajectory under some constraints. Figure 3 outlines the overall workflow of the proposed GINet-DGC framework.



**Figure 3.** Workflow of the GINet-DGC framework. The model first extracts Multi-View Structural Priors from high-dimensional biomedical features, including Latent Semantic, Global Distributional, Local Topological, and Hierarchical features. It then integrates these priors through Adaptive Constraint Integration to condition the Weight Generator and synthesize task-specific weights. Finally, Dynamic Generalization Control monitors the generalization trajectory and performs adaptive early stopping.

## 2.2. Structure-Constrained Weight Generator

In high-dimensional, low-sample-size (HDLSS) settings, directly optimizing the first-layer interaction matrix as unconstrained parameters often results in severe over-parameterization. Prior studies indicate that structured parameter generation can effectively improve generalization by contracting the parameter search space. Inspired by this paradigm, GINet-DGC introduces multi-view structural priors to condition the construction of the input interaction matrix. Conceptually, this mechanism functions as a structurally constrained HyperNetwork. Instead of freely optimizing weights from scarce samples, the interaction matrix is synthesized based on inherent feature descriptors. This descriptor-conditioned generation restricts the optimization search space and mitigates parameter redundancy, thereby enhancing model robustness under data-scarce regimes.

We use the following notation throughout this section. Let  $X^{\text{raw}} \in \mathbb{R}^{N_0 \times D}$  denote the original input matrix, samples with incomplete data are removed. Within each cross-

validation fold, the training split is then normalized by Min–Max Scaling so that each feature of the training matrix  $X_{\text{tr}}$  is linearly mapped into the interval  $[0, 1]$ :

$$(X_{\text{tr}})_{ij} = \frac{(X_{\text{tr}}^{\text{raw}})_{ij} - \min_i(X_{\text{tr}}^{\text{raw}})_{ij}}{\max_i(X_{\text{tr}}^{\text{raw}})_{ij} - \min_i(X_{\text{tr}}^{\text{raw}})_{ij}}, \quad (X_{\text{tr}})_{ij} \in [0, 1], \quad (1)$$

where  $X_{\text{tr}}^{\text{raw}}$  denotes the unnormalized training split, and  $i$  and  $j$  index samples and features, respectively. The minimum and maximum are computed over samples for each feature. The resulting preprocessed training matrix is denoted as  $X_{\text{tr}} \in \mathbb{R}^{N_{\text{tr}} \times D}$ , with  $N_{\text{tr}}$  training samples and  $D$  features, while  $M$  denotes the feature-embedding dimension.

**NMF-based gene embedding initialization.** Following the normalization step, Non-negative Matrix Factorization (NMF) is applied to the preprocessed training matrix  $X_{\text{tr}}$  to derive a feature-wise latent embedding matrix  $V \in \mathbb{R}_+^{D \times M}$ :

$$X_{\text{tr}} \approx WV^{\top}, \quad (2)$$

where  $M$  denotes the embedding dimension. The concept of utilizing such feature-wise low-dimensional embeddings to guide weight generation is inspired by WPFS [23]. In this formulation, each row of  $V$  functions as a compact semantic descriptor for a specific feature. Instead of treating variables as isolated coordinates, the non-negativity constraint of NMF enforces part-based, additive feature representations [33]. These localized structural priors are inherently easier to regularize, effectively stabilizing the subsequent conditional weight prediction process.

The objective of the Structure-Constrained Weight Generator is to synthesize the first-layer projection matrix  $A$  from the structural priors extracted from  $V$ . The generation process is defined as

$$A = G_{\Theta}(F(V)) \in \mathbb{R}^{D \times K}, \quad (3)$$

where  $K$  denotes the hidden width of the first classifier layer, the generator parameters are denoted by  $\Theta$ , and  $F(V)$  denotes the fused multi-view structural representation. The generated matrix  $A$  is subsequently utilized by the task classifier  $C_{\Omega}$ , whose remaining trainable parameters are denoted by  $\Omega$ .

GINet-DGC extracts structural priors from four complementary views:

1. Latent Semantic Prior: It constructs basic subspace constraints.
2. Global Distributional Prior: It captures dataset-wide statistical stability.
3. Local Topological Prior: It preserves manifold geometry among features.
4. Hierarchical Semantic Prior: It provides multi-scale nonlinear abstractions.

Then an adaptive integration mechanism fuses these multi-view priors to form a structure-constrained representation. The generation process is summarized as follows:

$$A = G_{\Theta}(\text{Fuse}(F_{\text{lat}}(V), F_{\text{dist}}(V), F_{\text{topo}}(V), F_{\text{hier}}(V))). \quad (4)$$

Here,  $F_{\text{lat}}$ ,  $F_{\text{dist}}$ ,  $F_{\text{topo}}$ , and  $F_{\text{hier}}$  denote the four prior extraction modules. The generated matrix  $A$  strictly constrains the classifier's first-layer search space. Independently, the Dynamic Generalization Control strategy monitors the training trajectory and terminates optimization when the validation behavior indicates generalization saturation.

### 2.2.1. Latent Semantic Prior

The Latent Semantic Prior module transforms the feature matrix into a compact subspace representation. Its primary role is to impose a localized structural constraint.

For each feature embedding vector  $v_i \in \mathbb{R}^{1 \times M}$  (where  $i = 1, 2, \dots, D$ ), the nonlinear mapping is defined as

$$f_{\text{lat}}(v_i) = \sigma(v_i B_{\text{lat}} + b_{\text{lat}}) \in \mathbb{R}^{1 \times M_{\text{lat}}}, \quad (5)$$

where  $\sigma(\cdot)$  is the nonlinear activation function, while  $B_{\text{lat}} \in \mathbb{R}^{M \times M_{\text{lat}}}$  and  $b_{\text{lat}} \in \mathbb{R}^{1 \times M_{\text{lat}}}$  denote the learnable weight matrix and bias vector of this module, respectively. The complete latent representation for all features is stacked as  $F_{\text{lat}}(V) = [f_{\text{lat}}(v_1), \dots, f_{\text{lat}}(v_D)]^T \in \mathbb{R}^{D \times M_{\text{lat}}}$ . This projection filters out redundant coordinate noise and maps feature-wise behaviors into a smooth, compact latent semantic space.

### 2.2.2. Global Distributional Prior

The Global Distributional Prior module captures the macroscopic statistical regularities of the feature-embedding space. Utilizing the embedding matrix  $V \in \mathbb{R}^{D \times M}$ , the module flattens all subspace variables to compute a global descriptor vector  $s \in \mathbb{R}^{1 \times 8}$ , where IQR denotes the interquartile range:

$$s = [\mu, \text{Med}, \sigma, \sigma^2, \text{Skew}, \text{Kurt}, R, \text{IQR}] \in \mathbb{R}^{1 \times 8}. \quad (6)$$

Here, the vector elements denote the global mean, median, standard deviation, variance, skewness, kurtosis, range, and interquartile range of the flattened matrix  $V$ . To calibrate the importance of different statistical properties, a learnable weight vector  $\gamma_{\text{stat}} \in \mathbb{R}^{1 \times 8}$  is introduced via element-wise multiplication.

The weighted statistical vector is expanded along the feature dimension  $D$  to construct an extended prior matrix  $S_{\text{ext}} \in \mathbb{R}^{D \times 8}$ . This matrix is mapped through two linear projection layers to extract cross-descriptor interactions:

$$F_{\text{dist}}(V) = \sigma(H_{\text{dist}} B_{\text{dist},2} + b_{\text{dist},2}) \in \mathbb{R}^{D \times M_{\text{dist}}}, \quad (7)$$

where  $H_{\text{dist}} = \sigma((S_{\text{ext}} \odot \Gamma) B_{\text{dist},1} + b_{\text{dist},1}) \in \mathbb{R}^{D \times H_{\text{dist\_mid}}}$ . The matrix  $\Gamma \in \mathbb{R}^{D \times 8}$  is broadcasted from  $\gamma_{\text{stat}}$ , and  $\odot$  denotes the Hadamard product. This layout restricts the classifier's search space based on dataset-wide regularities.

### 2.2.3. Local Topological Prior

The Local Topological Prior module models the neighborhood geometry and manifold patterns within the latent subspace [34]. It extracts structural descriptors from  $V$  via a joint soft-clustering and  $K$ -nearest neighbor (KNN) formulation.

The module maintains  $K_c$  learnable cluster centers  $C = [c_1, c_2, \dots, c_{K_c}]^T \in \mathbb{R}^{K_c \times M}$ . For each feature vector  $v_i \in \mathbb{R}^{1 \times M}$ , the soft assignment  $p_{ij}$  to center  $c_j$  is defined as

$$p_{ij} = \frac{\exp(-\|v_i - c_j\|^2 \cdot w_j / \tau)}{\sum_{j'=1}^{K_c} \exp(-\|v_i - c_{j'}\|^2 \cdot w_{j'} / \tau)}, \quad (8)$$

where  $\tau$  is a learnable temperature, and  $w_j \in \mathbb{R}$  is the  $j$ -th element of the cluster scaling vector  $w \in \mathbb{R}^{K_c}$ . Concurrently, the module extracts local density by computing Euclidean distances to the  $K_n$  nearest neighbors:

$$d_{\text{km}}(v_i) = [d_{i(1)}, d_{i(2)}, \dots, d_{i(K_n)}] \in \mathbb{R}^{1 \times K_n}. \quad (9)$$

The soft-clustering assignments and KNN distance profiles jointly form the topological descriptor  $S_{\text{topo}} \in \mathbb{R}^{D \times (K_c + K_n)}$ , capturing both coarse cluster affinity and local neighborhood geometry. The final prior is generated via a linear projection layer:

$$F_{\text{topo}}(V) = \sigma(S_{\text{topo}}B_{\text{topo}} + b_{\text{topo}}) \in \mathbb{R}^{D \times M_{\text{topo}}}, \quad (10)$$

where  $B_{\text{topo}} \in \mathbb{R}^{(K_c + K_n) \times M_{\text{topo}}}$  and  $b_{\text{topo}} \in \mathbb{R}^{1 \times M_{\text{topo}}}$  are learnable weights and biases.

#### 2.2.4. Hierarchical Semantic Prior

To capture dependencies across semantic scales, this module employs  $L$  parallel shallow subnetworks, each projecting the embedding matrix  $V$  into a scale-specific representation:

$$H_l = \sigma(\sigma(VB_{l,1} + b_{l,1})B_{l,2} + b_{l,2}) \in \mathbb{R}^{D \times M_{\text{level}}}, \quad l = 1, \dots, L, \quad (11)$$

where  $B_{l,1} \in \mathbb{R}^{M \times M_{\text{mid}}}$  and  $B_{l,2} \in \mathbb{R}^{M_{\text{mid}} \times M_{\text{level}}}$  represent level-specific parameters. All branch outputs are concatenated into a joint tensor  $H_{\text{comb}} = [H_1 \parallel H_2 \parallel \dots \parallel H_L] \in \mathbb{R}^{D \times (M_{\text{level}} \cdot L)}$ . The hierarchical prior is synthesized via a final fusion layer:

$$F_{\text{hier}}(V) = \sigma(H_{\text{comb}}B_{\text{fusion}} + b_{\text{fusion}}) \in \mathbb{R}^{D \times M_{\text{hier}}}, \quad (12)$$

where  $B_{\text{fusion}} \in \mathbb{R}^{(M_{\text{level}} \cdot L) \times M_{\text{hier}}}$  and  $b_{\text{fusion}}$  denote the learnable fusion weights and biases.

#### 2.2.5. Adaptive Prior Integration and Weight Generation

The integration layer models the non-linear interactions across the multi-view descriptors. The structural prior fields are concatenated to form a unified tensor  $F_{\text{all}}(V) = [F_{\text{lat}}(V) \parallel F_{\text{dist}}(V) \parallel F_{\text{topo}}(V) \parallel F_{\text{hier}}(V)] \in \mathbb{R}^{D \times M_{\text{all}}}$ . To adjust feature-wise contributions dynamically, a multi-head self-attention mechanism [35] maps the contextual dependencies:

$$F_{\text{fuse}} = \text{Softmax}\left(\frac{(F_{\text{all}}B_Q)(F_{\text{all}}B_K)^{\top}}{\sqrt{d_k}}\right)(F_{\text{all}}B_V) \in \mathbb{R}^{D \times M_{\text{all}}}, \quad (13)$$

where  $B_Q, B_K, B_V \in \mathbb{R}^{M_{\text{all}} \times M_{\text{all}}}$  are learnable projection weights. This Transformer-based layout enhances global dependency modeling and representational flexibility.

Finally, the weight generator maps the fused representation  $F_{\text{fuse}}$  to synthesize the target task projection matrix  $A$ :

$$A = \sigma(\sigma(F_{\text{fuse}}B_{\text{gen},1} + b_{\text{gen},1})B_{\text{gen},2} + b_{\text{gen},2})B_{\text{gen},3} + b_{\text{gen},3} \in \mathbb{R}^{D \times K}, \quad (14)$$

where  $B_{\text{gen},1}, B_{\text{gen},2}, B_{\text{gen},3}$  represent the sequential linear layers, and the final output dimension ensures that  $A$  matches the first classifier layer exactly.

### 2.3. Dynamic Generalization Control

In HDLSS tasks, overfitting is not only a final-state problem but also a trajectory problem: the model may continue reducing training loss after validation performance has saturated or begun to deteriorate. Therefore, rather than relying on standard fixed stopping rules [36,37], we introduce an adaptive early stopping strategy termed Dynamic Generalization Control. DGC closely tracks the training trajectory to determine if the optimization process has entered an overfitting regime, where subsequent parameter updates serve primarily to widen the generalization gap.

The stopping signal is the Overfitting-aware Index, which integrates three factors: accuracy discrepancy, loss discrepancy, and validation-performance deterioration. At iteration  $t$ , it is defined as

$$\text{OFI}_t = \alpha(\text{Acc}_{\text{train}}^t - \text{Acc}_{\text{val}}^t) + \beta(\text{Loss}_{\text{val}}^t - \text{Loss}_{\text{train}}^t) + \gamma r_t, \quad (15)$$

where  $r_t$  denotes the validation deterioration trend over a recent window of iterations. A low deterioration trend reduces the overfitting warning, whereas a larger deterioration trend increases the warning when it co-occurs with growing train–validation gaps.

To avoid a manually fixed stopping threshold, DGC computes an adaptive threshold from the historical OFI distribution:

$$\tau_t = \mu_{\text{OFI}}^t + \kappa \sigma_{\text{OFI}}^t, \quad (16)$$

where  $\mu_{\text{OFI}}^t$  and  $\sigma_{\text{OFI}}^t$  are the mean and standard deviation of previous OFI values, and  $\kappa$  controls threshold tolerance. Training is terminated only when  $\text{OFI}_t > \tau_t$  for  $P$  consecutive validation checks. The OFI-related hyperparameters, including  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\kappa$ , and  $P$ , are determined by automatic hyperparameter optimization rather than manual assignment. This patience-based decision separates the trajectory metric from the stopping rule and reduces sensitivity to transient validation fluctuations.

#### 2.4. Training Process

The training objective is to minimize the cross-entropy loss between predictions and true labels:

$$\mathcal{L}(\Theta, \Omega) = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_c^i \log(\hat{y}_c^i), \quad (17)$$

where  $y_c^i$  is the true label indicator and  $\hat{y}_c^i$  represents the predicted probability. The training procedure is summarized in Algorithm 1.

---

#### Algorithm 1: Training Procedure of GINet-DGC

---

**Input:** Training dataset  $\{(x_i, y_i)\}_{i=1}^N$ ; validation set; learning rate  $\lambda$ ; maximum iterations  $T_{\max}$ ; patience  $P$ .  
**Output:** Prediction probabilities  $\hat{y}$ .

- 1 **Initialization:** Initialize parameter matrices  $\Theta$  and  $\Omega$ ; compute feature embeddings  $V$  and multi-view priors  $F(V)$ ; set optimizer and counter  $q \leftarrow 0$ ;
- 2 **for**  $t = 1$  **to**  $T_{\max}$  **do**
- 3     Sample a mini-batch  $(x, y)$  from the training set;
- 4     Forward propagation: compute prediction probabilities  $\hat{y} = \text{Forward}(x)$ ;
- 5     Compute the cross-entropy loss  $\mathcal{L}(\Theta, \Omega) = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_c^i \log(\hat{y}_c^i)$ ;  
       Backward propagation: update  $\Theta$  and  $\Omega$  via gradients;
- 6     Evaluate training and validation metrics, compute  $\text{OFI}_t$ , and update adaptive threshold  $\tau_t$ ;
- 7     **if**  $\text{OFI}_t > \tau_t$  **then**
- 8          $q \leftarrow q + 1$ ;
- 9     **else**
- 10          $q \leftarrow 0$ ;
- 11     **if**  $q \geq P$  **then**
- 12         Stop training and restore the best validation checkpoint;
- 13 **return**  $\hat{y}$

---

#### 2.5. Experimental Setup

##### 2.5.1. Experimental Data

We evaluated the proposed framework on eight public high-dimensional, low-sample-size biomedical gene-expression datasets, spanning 50 to 203 samples and 3312 to 22,283 features. Detailed dataset information is provided in Appendix A.

To ensure reproducibility, we performed repeated stratified 5-fold cross-validation with five independent repeats. Feature embeddings were generated independently within

each training fold of the repeated stratified 5-fold cross-validation procedure. All structural priors were independently recalculated within each training fold. For each run, 80% of the data was allocated for development and 20% for testing. Within the development split, 10% of the samples were reserved as a validation subset for hyperparameter tuning and Dynamic Generalization Control. We report the mean balanced accuracy over all 25 test evaluations as the final performance metric.

### 2.5.2. Evaluation Metrics

We adopt balanced accuracy (BA) as the primary evaluation metric. Unlike traditional accuracy, balanced accuracy objectively reflects classification capability under class imbalance and small-sample conditions. In high-dimensional datasets, class distribution is often skewed. Standard accuracy can be misleading, as models may bias toward the majority class, obscuring their ability to recognize minority samples.

Balanced accuracy measures performance by averaging the recall of each class:

$$BA = \frac{1}{C} \sum_{c=1}^C \frac{TP_c}{TP_c + FN_c} \quad (18)$$

where  $C$  denotes the number of classes, and  $TP_c$  and  $FN_c$  represent true positives and false negatives for class  $c$ .

This metric provides a stable reflection of discriminative ability, especially in cases of severe imbalance or limited sample size.

### 2.5.3. GINet-DGC Settings

The classification backbone of GINet-DGC consists of three fully connected layers. The final layer uses Softmax for class distribution. The Structure-Constrained Weight Generator includes four prior extraction modules: a statistical feature extractor (64-dim output), a low-dimensional semantic encoder (64-dim output), a topological clustering module (32-dim output, 8 centers), and a multi-level feature module (64-dim output). The adaptive attention mechanism outputs 128-dimensional fused features. Finally, the generator produces 100-dimensional weights.

For the Dynamic Generalization Control (DGC) adaptive stopping strategy, the OFI-related hyperparameters were determined through a reproducible random-search procedure on the validation subset. Specifically, we constructed a five-dimensional search space combining continuous uniform distributions and discrete candidate sets: the weights  $\alpha$  and  $\beta$  were sampled from  $U(0.25, 2.0)$ , the trend penalty term followed  $\gamma \sim U(0.25, 1.0)$ , and the tolerance  $\kappa$  and decision patience  $P$  were randomly selected from  $\{2, 3, 4\}$  and  $\{100, 150, 200, 300, 400\}$ , respectively. Within this space, 20 distinct hyperparameter configurations were randomly sampled. On the benchmark dataset, each configuration was run for 500 training steps to rapidly evaluate its performance, resulting in an overall search cost of approximately 2.8 h on a single NVIDIA GeForce RTX 3090 graphics processing unit (GPU) across the 20 independent runs. The 20 sampled configurations were chosen as a practical trade-off between search coverage and computational cost in this relatively constrained search space. Ultimately, the random search identified the best-performing hyperparameter configuration within the predefined search space based on the average validation balanced accuracy:  $\alpha = 1.0$ ,  $\beta = 0.5$ ,  $\gamma = 0.5$ ,  $\kappa = 3$ , and  $P = 200$ . This configuration was then frozen and uniformly deployed across all scenarios to ensure a fair evaluation of model transfer performance. Because random seeds may affect the sampled candidates, this procedure aims to obtain a practical fixed configuration under a limited budget rather than a seed-invariant global optimum. Additionally, Batch Normalization

and Dropout are incorporated into the network architecture, with the Leaky Rectified Linear Unit (LeakyReLU) employed as the activation function to enhance training stability.

The model is trained using AdamW [38] with a batch size of 8. The maximum training steps are set to 5000.

### 3. Results

#### 3.1. Baselines

We compare GINet-DGC against 17 representative baselines. These cover traditional machine learning, modern neural architectures, and specialized HDLSS approaches. For fairness, all baseline methods are evaluated on the same eight public datasets and under the same repeated stratified  $5 \times 5$  cross-validation protocol, and the reported balanced accuracy is averaged over the resulting 25 evaluations. We do not directly copy performance numbers from prior publications when their datasets, generated samples, preprocessing pipelines, or split protocols are not fully aligned with ours.

For traditional methods, we selected Random Forest [39] and LightGBM [12]. These represent bagging and gradient boosting frameworks. They are known for robustness on sparse data.

For deep tabular models, we include a standard multilayer perceptron (MLP), TabNet [16], TabTransformer [17], LassoNet [40], Deep Neural Pursuit (DNP) [32], Concrete Autoencoder (CAE), and HyperTab [21]. We also include TabPFN [29], a recent foundation model for tabular learning. TabPFN leverages Bayesian priors and conditional sampling, representing a strong small-sample tabular baseline.

For feature generation and structured learning, we evaluate DietNetworks [22], FsNet [19], the Sparse-Input Neural Network (SPINN) [41], and WPFS [23]. We also evaluate GCondNet [24], the Graph Convolutional Network (GCN) [42], and the Graph Attention Network version 2 (GATv2) [43]. These methods enhance representation through mechanisms like conditional weight prediction and implicit relational modeling.

#### 3.2. Benchmark Comparison

Table 2 summarizes the mean balanced accuracy and standard deviation of all evaluated methods across the eight biomedical HDLSS datasets under repeated stratified  $5 \times 5$  cross-validation.

As shown in Table 2, GINet-DGC demonstrates competitive performance across eight high-dimensional, small-sample biomedical datasets. This supports the usefulness of its structural modeling and adaptive stopping strategy for bioinformatics prediction benchmarks. Specifically, GINet-DGC ranks among the top three on all datasets and achieves the highest balanced accuracy on gli, smk, allaml, glioma, toxicity, and lung. Its average rank is 1.50, lower than other recent competitive models such as GCondNet (3.13) and TabPFNTTransformer (5.88). Additionally, the model displays a relatively stable performance profile under repeated  $5 \times 5$  cross-validation. In particular, on the lung dataset, it records a balanced accuracy of  $98.77 \pm 0.46$ , presenting a lower variance compared with baselines such as GCondNet ( $96.64 \pm 3.10$ ) and LightGBM ( $89.79 \pm 4.40$ ), which suggests a degree of consistency across repeated experimental trials.

Regarding model categories, traditional tree-based models remain competitive in certain biomedical tasks, benefiting from their inherent robustness to lower-order statistics. However, they can underperform when compared with deep learning approaches; because tree-based methods rely on local greedy partitioning, they may struggle to fully capture complex global dependencies among high-dimensional molecular variables, which can limit their overall representation capacity.

**Table 2.** Balanced accuracy (%) across datasets under repeated stratified  $5 \times 5$  cross-validation, reported as mean  $\pm$  standard deviation.

| Method<br>N/D           | Gli<br>0.004            | Smk<br>0.009            | Allaml<br>0.01          | CLL<br>0.01             | Glioma<br>0.011          | Prostate<br>0.017       | Toxicity<br>0.03        | Lung<br>0.059           | Avg-Rank    |
|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|--------------------------|-------------------------|-------------------------|-------------------------|-------------|
| DietNetworks            | 76.42 $\pm$ 13.20       | 62.71 $\pm$ 9.40        | 92.00 $\pm$ 8.40        | 68.84 $\pm$ 9.20        | 68.00 $\pm$ 14.80        | 81.71 $\pm$ 11.00       | 82.13 $\pm$ 7.40        | 90.43 $\pm$ 6.20        | 12.38       |
| FsNet                   | 74.52 $\pm$ 11.70       | 56.27 $\pm$ 9.20        | 78.00 $\pm$ 12.90       | 66.38 $\pm$ 9.20        | 53.17 $\pm$ 12.90        | 84.74 $\pm$ 9.80        | 60.26 $\pm$ 8.10        | 91.75 $\pm$ 3.00        | 14.63       |
| DNP                     | 83.17 $\pm$ 12.10       | 66.61 $\pm$ 8.40        | 96.18 $\pm$ 5.70        | <b>85.13</b> $\pm$ 5.50 | 75.00 $\pm$ 12.80        | 88.71 $\pm$ 6.80        | 93.49 $\pm$ 6.20        | 92.81 $\pm$ 6.70        | 6.00        |
| SPINN                   | 83.39 $\pm$ 9.80        | 65.91 $\pm$ 7.60        | 96.78 $\pm$ 6.20        | <b>85.35</b> $\pm$ 5.50 | 75.00 $\pm$ 14.80        | 90.02 $\pm$ 6.80        | <b>93.50</b> $\pm$ 4.90 | 94.76 $\pm$ 4.40        | <b>4.63</b> |
| WPFS                    | 83.86 $\pm$ 9.10        | 66.89 $\pm$ 6.20        | 96.42 $\pm$ 4.20        | 79.14 $\pm$ 4.50        | 73.83 $\pm$ 16.50        | 89.15 $\pm$ 6.70        | 88.29 $\pm$ 5.30        | <b>98.43</b> $\pm$ 4.20 | 5.50        |
| TabNet                  | 64.54 $\pm$ 12.90       | 61.16 $\pm$ 9.20        | 71.64 $\pm$ 17.70       | 50.87 $\pm$ 13.80       | 50.00 $\pm$ 16.90        | 65.75 $\pm$ 17.70       | 41.38 $\pm$ 9.60        | 75.11 $\pm$ 10.20       | 16.63       |
| TabTransformer          | 78.82 $\pm$ 14.10       | 64.00 $\pm$ 9.20        | 88.38 $\pm$ 8.60        | 76.81 $\pm$ 6.80        | 63.50 $\pm$ 15.60        | 85.96 $\pm$ 11.50       | 87.67 $\pm$ 6.10        | 94.03 $\pm$ 4.70        | 10.50       |
| HyperTab                | 82.59 $\pm$ 10.35       | 65.85 $\pm$ 8.74        | 86.84 $\pm$ 9.83        | 76.70 $\pm$ 8.16        | 73.67 $\pm$ 13.32        | 87.16 $\pm$ 5.93        | 79.28 $\pm$ 6.68        | 91.36 $\pm$ 8.41        | 10.25       |
| CAE                     | 74.18 $\pm$ 11.70       | 59.96 $\pm$ 11.00       | 89.80 $\pm$ 9.20        | 71.94 $\pm$ 13.40       | 67.83 $\pm$ 17.60        | 87.60 $\pm$ 7.80        | 60.36 $\pm$ 11.30       | 85.00 $\pm$ 5.00        | 13.50       |
| LassoNet                | 53.91 $\pm$ 10.90       | 51.04 $\pm$ 8.60        | 50.80 $\pm$ 12.90       | 30.63 $\pm$ 8.70        | 29.17 $\pm$ 11.80        | 54.78 $\pm$ 10.60       | 26.67 $\pm$ 8.70        | 25.11 $\pm$ 9.80        | 18.00       |
| MLP                     | 77.72 $\pm$ 15.30       | 64.42 $\pm$ 8.40        | 91.30 $\pm$ 6.70        | 78.30 $\pm$ 9.00        | 73.00 $\pm$ 14.90        | 88.76 $\pm$ 5.50        | 93.21 $\pm$ 6.10        | 94.20 $\pm$ 4.90        | 8.88        |
| Random Forest           | 81.15 $\pm$ 8.50        | 69.84 $\pm$ 4.60        | <b>96.80</b> $\pm$ 5.60 | 76.44 $\pm$ 10.10       | 74.17 $\pm$ 10.60        | 90.35 $\pm$ 8.20        | 80.99 $\pm$ 4.50        | 88.14 $\pm$ 5.20        | 7.75        |
| LightGBM                | 80.79 $\pm$ 7.60        | <b>70.07</b> $\pm$ 5.80 | 95.36 $\pm$ 5.20        | 74.22 $\pm$ 14.60       | <b>75.50</b> $\pm$ 11.90 | <b>91.91</b> $\pm$ 4.80 | 81.26 $\pm$ 4.20        | 89.79 $\pm$ 4.40        | 7.50        |
| GCN                     | <b>84.09</b> $\pm$ 9.40 | 65.63 $\pm$ 8.00        | 80.83 $\pm$ 10.80       | 72.00 $\pm$ 8.40        | 66.23 $\pm$ 14.40        | 82.60 $\pm$ 12.50       | 76.13 $\pm$ 7.00        | 93.30 $\pm$ 4.60        | 11.38       |
| GATv2                   | 73.57 $\pm$ 12.40       | 66.06 $\pm$ 8.20        | 71.36 $\pm$ 11.60       | 57.74 $\pm$ 14.10       | 57.67 $\pm$ 15.10        | 83.23 $\pm$ 10.60       | 76.65 $\pm$ 11.20       | 93.33 $\pm$ 6.20        | 13.38       |
| GCondNet                | <b>86.36</b> $\pm$ 8.00 | 68.08 $\pm$ 7.30        | <b>97.56</b> $\pm$ 4.10 | 80.70 $\pm$ 5.50        | <b>77.67</b> $\pm$ 10.50 | 90.38 $\pm$ 5.60        | <b>95.25</b> $\pm$ 4.50 | <b>96.64</b> $\pm$ 3.10 | <b>3.13</b> |
| TabPFNTransformer       | 76.47 $\pm$ 9.95        | <b>71.28</b> $\pm$ 6.13 | 95.53 $\pm$ 4.87        | 82.69 $\pm$ 5.90        | 73.33 $\pm$ 13.29        | <b>93.36</b> $\pm$ 3.75 | 90.17 $\pm$ 5.33        | 94.89 $\pm$ 5.52        | 5.63        |
| <b>GINet-DGC (ours)</b> | <b>90.80</b> $\pm$ 5.69 | <b>72.06</b> $\pm$ 8.01 | <b>98.00</b> $\pm$ 4.08 | <b>84.02</b> $\pm$ 4.48 | <b>79.99</b> $\pm$ 6.80  | <b>91.44</b> $\pm$ 9.32 | <b>96.04</b> $\pm$ 4.27 | <b>98.77</b> $\pm$ 0.46 | <b>1.50</b> |

Note. Red, blue, and golden values indicate the best, second-best, and third-best mean performance on each dataset, respectively. Values are shown as mean  $\pm$  standard deviation. All methods are evaluated on the same public datasets under the same repeated  $5 \times 5$  cross-validation protocol. Abbreviations: N/D, ratio of the number of samples ( $N$ ) to the number of features ( $D$ ); Avg-Rank, average rank; DNP, Deep Neural Pursuit; SPINN, Sparse-Input Neural Network; WPFS, Weight Predictor Network with Feature Selection; CAE, Concrete Autoencoder; MLP, multilayer perceptron; LightGBM, Light Gradient Boosting Machine; GCN, Graph Convolutional Network; GATv2, Graph Attention Network version 2; GINet-DGC, Global Interaction Network with Dynamic Generalization Control.

Deep tabular models generally exhibit stable performance, with architectures like DNP and TabPFN demonstrating robust generalization. However, because these methods often depend on complex priors or extensive pretraining, they can display increased sensitivity to distribution shifts across different biomedical cohorts or assay conditions. Notably, GINet-DGC achieves competitive performance without relying on any external pretraining, showcasing its efficacy in a more self-contained manner.

Among feature generation methods, SPINN and GCondNet demonstrate robust efficacy, highlighting the practical value of explicit relation modeling. However, their reliance on rigid local assumptions can render them susceptible to noise. In contrast, GINet-DGC captures higher-order interactions globally while preserving local constraints, thereby enhancing robustness under the complex feature distributions characteristic of biomedical tabular data.

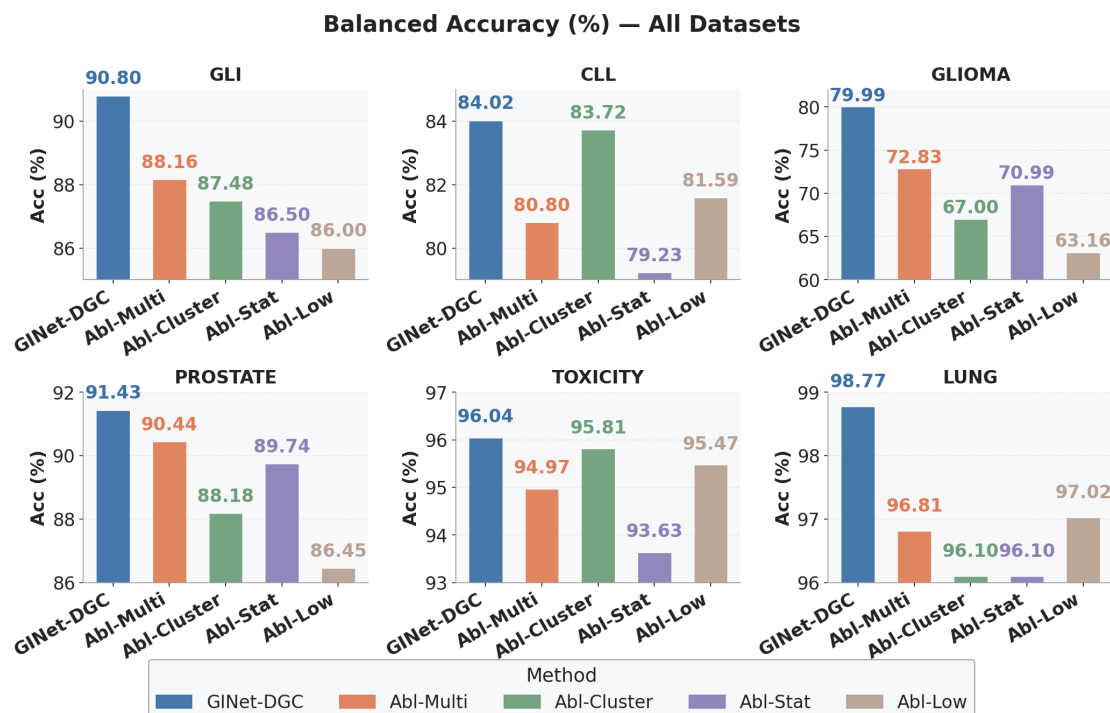
Overall, GINet-DGC provides a viable solution for data-scarce biomedical HDLSS conditions, demonstrating the practical utility of its structured global interaction design and adaptive early stopping strategy. The Friedman test confirmed significant overall differences among the 18 methods ( $\chi^2(17) = 104.46$ ,  $p = 1.32 \times 10^{-14}$ ). In the post-hoc Wilcoxon comparisons, GINet-DGC showed positive average differences against all 17 base-lines, with all unadjusted  $p$ -values below 0.05 ( $7.81 \times 10^{-3}$  to  $3.91 \times 10^{-2}$ ); it achieved 13 win/tie/loss counts of 8/0/0 and four of 7/0/1. Because no multiple-comparison correction was applied, these pairwise results should be interpreted at the nominal significance level. Detailed results are provided in Appendix D.

### 3.3. Ablation Study

#### 3.3.1. Multi-View Structural Priors

To quantify the contribution of each prior branch, we conducted an ablation study by systematically removing individual modules. Figure 4 summarizes the resulting performance variations, with the granular numerical outcomes (comprising means and standard deviations) detailed in Tables A3 and A4 to provide a more complete reference on empirical variability.

The performance degradation exhibits apparent task-dependent patterns, potentially reflecting the structural heterogeneity inherent in HDLSS data. Specifically, isolating or removing the hierarchical semantic or local topological prior tends to yield more noticeable performance decrements across the majority of tasks. This implies that different datasets may rely on distinct geometric or categorical regularities. Rather than exhibiting uniform dominance, these multi-view descriptors appear to offer complementary structural constraints. These findings support the potential utility of the adaptive fusion layer, which dynamically calibrates multi-view representations instead of assuming static, equal contributions across tasks.

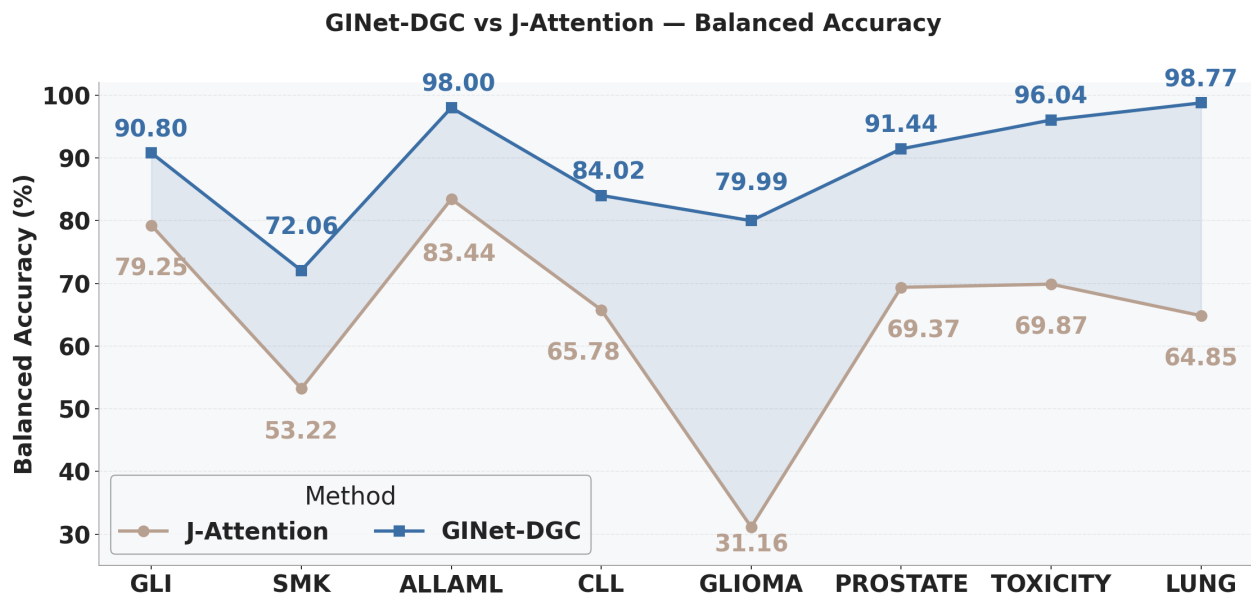


**Figure 4.** Ablation results of the Structure-Constrained Weight Generator. Each bar represents the balanced accuracy after removing a specific structural prior branch: *Abl-Low* denotes the removal of the latent semantic branch; *Abl-Stat* denotes the removal of the global distributional prior; *Abl-Cluster* denotes the removal of the local topological prior; and *Abl-Multi* denotes the removal of the hierarchical semantic prior. GINet-DGC represents the complete model. The complete architecture consistently outperforms all ablation variants, confirming the empirical utility of structured weight generation.

Quantitatively, our ablation experiments tend to reveal a general tendency toward performance decreases across the evaluated datasets when individual structural priors are omitted. For instance, the complete GINet-DGC model achieves a mean balanced accuracy of 79.99% on the GLIOMA dataset, whereas omitting the hierarchical semantic prior (*Abl-Multi*) or the local topological prior (*Abl-Cluster*) leads to lower observed means of 72.83% and 67.00%, respectively. This visible shift in empirical performance suggests that while the relative contribution of each prior varies across tasks, their collaborative integration appears to provide a supportive regularization effect, potentially helping mitigate substantial performance fluctuations in data-scarce scenarios.

### 3.3.2. Adaptive Prior Integration

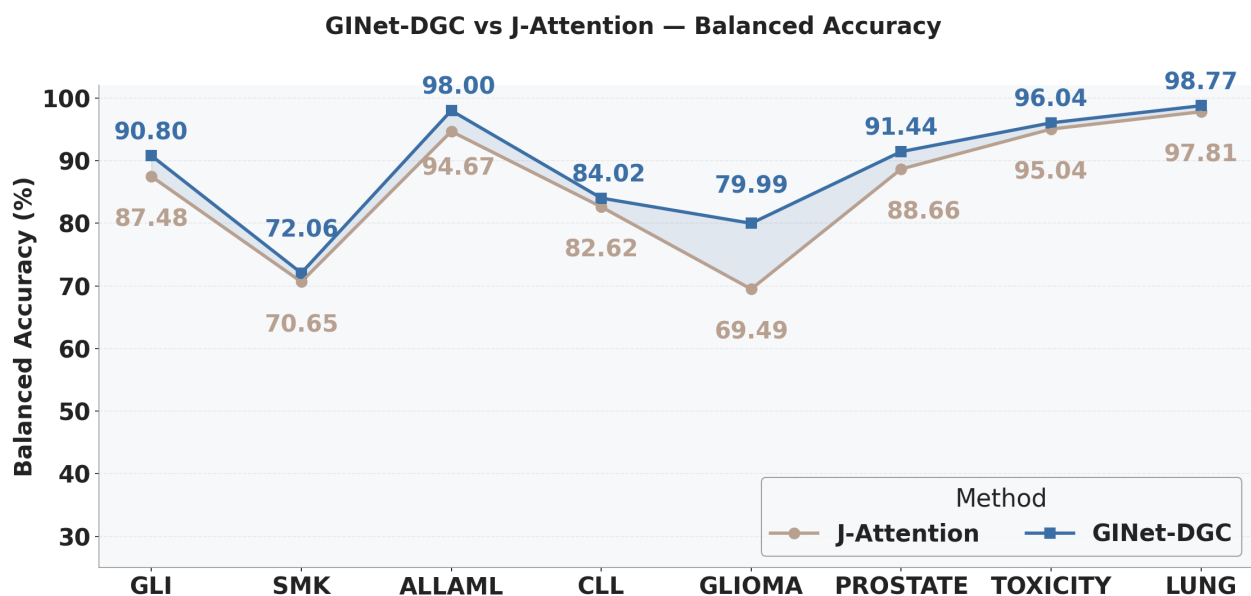
To evaluate the attention module, we conducted an ablation by removing it while keeping all feature layers. Figure 5 reports the comparison of balanced accuracy between the full model and the variant without attention, with the corresponding detailed mean and standard deviation metrics documented in Table A4.



**Figure 5.** Ablation results of the attention mechanism. Each bar represents the balanced accuracy comparison between the complete GINet-DGC model and its ablation variant without the attention module (denoted as Abl-Attention).

As shown in Figure 5, the attention mechanism contributes substantially to the overall predictive efficacy. Eliminating this component leads to a pronounced performance degradation, with the most severe decline observed on the *glioma* dataset (a decrease exceeding 40%). This underscores that the attention mechanism enables dynamic and adaptive weight allocation, thereby bolstering prominent feature channels while effectively suppressing redundant information to enhance the overall feature fusion process.

We also investigated the dependency on multi-view features via an attention-only experiment (without structural priors). Figure 6 presents the comparison between this attention-only variant (denoted as J-AttentionOnly) and the full GINet-DGC model.



**Figure 6.** Ablation results of the attention-only mechanism. The line plots illustrate the balanced accuracy comparison between the complete GINet-DGC model and the attention-only variant (denoted as J-AttentionOnly) across different datasets.

As illustrated in Figure 6, J-AttentionOnly underperforms relative to the complete architecture across the evaluated settings. This disparity indicates that self-attention alone possesses inherent limitations in sample-scarce regimes; without the guidance of structured multi-level features, it typically struggles to capture the implicit geometric or categorical structures embedded within high-dimensional tabular data.

Consequently, these findings demonstrate that the attention module operates synergistically with the multi-layer structural embedding as an integrated framework, rather than acting as an isolated performance driver.

In summary, the structured weight generation strategy bolsters model robustness within the evaluated settings. To cross-validate these findings, we extended our investigations to include comparisons with alternative initialization strategies (Appendix B), which further substantiate the empirical utility of the proposed structure-constrained generation mechanism.

To verify the efficacy of the DGC adaptive stopping strategy, we compared models utilizing this mechanism against those with conventional early stopping (CES) across eight molecular datasets.

The top panel of Figure 7 demonstrates that the integration of DGC yields observable performance gains across several benchmarks, with the most pronounced improvement observed on the *GLIOMA* dataset, where the balanced accuracy increases from 79.16% to 80.50%. Simultaneously, as illustrated in the bottom panel, DGC substantially refines computational efficiency by reducing redundant optimization steps. This efficiency gain is most prominent on the same *GLIOMA* dataset, where training iterations drop by 38.37% (from 2575 down to 1587 steps).

### 3.4. Analysis of Dynamic Generalization Control

It is worth noting that since DGC modifies the termination criterion rather than altering the underlying training trajectory, its accuracy improvements across remaining tasks are generally steady rather than dramatic. Rather than aggressively reshaping feature representations, DGC acts as a precise checkpoint selector that reliably anchors the optimal generalization state. Consequently, it achieves a well-balanced empirical outcome: steadily securing performance enhancements while simultaneously consolidating efficiency through fewer training steps.

Accordingly, the average predictive gains of DGC can be modest: when late-stage overfitting is noticeable, DGC helps stop training closer to the optimal generalization point and improves balanced accuracy, whereas when overfitting is less pronounced, its main contribution is to identify saturation and reduce redundant training steps.

To verify the universality and transferability of this mechanism, we further integrated the DGC module into another framework, GCondNet, and conducted extended experiments on the Prostate, Toxicity, and Lung datasets. As detailed in Table A5 (Appendix C), DGC consistently outperforms patience-based early stopping, standard early stopping, and the ReduceLROnPlateau scheduler under identical experimental settings, robustly demonstrating its strong architectural adaptability. In addition, we conducted a hyperparameter sensitivity analysis of the OFI-related parameters, with the results summarized in Figure A1 (Appendix C).

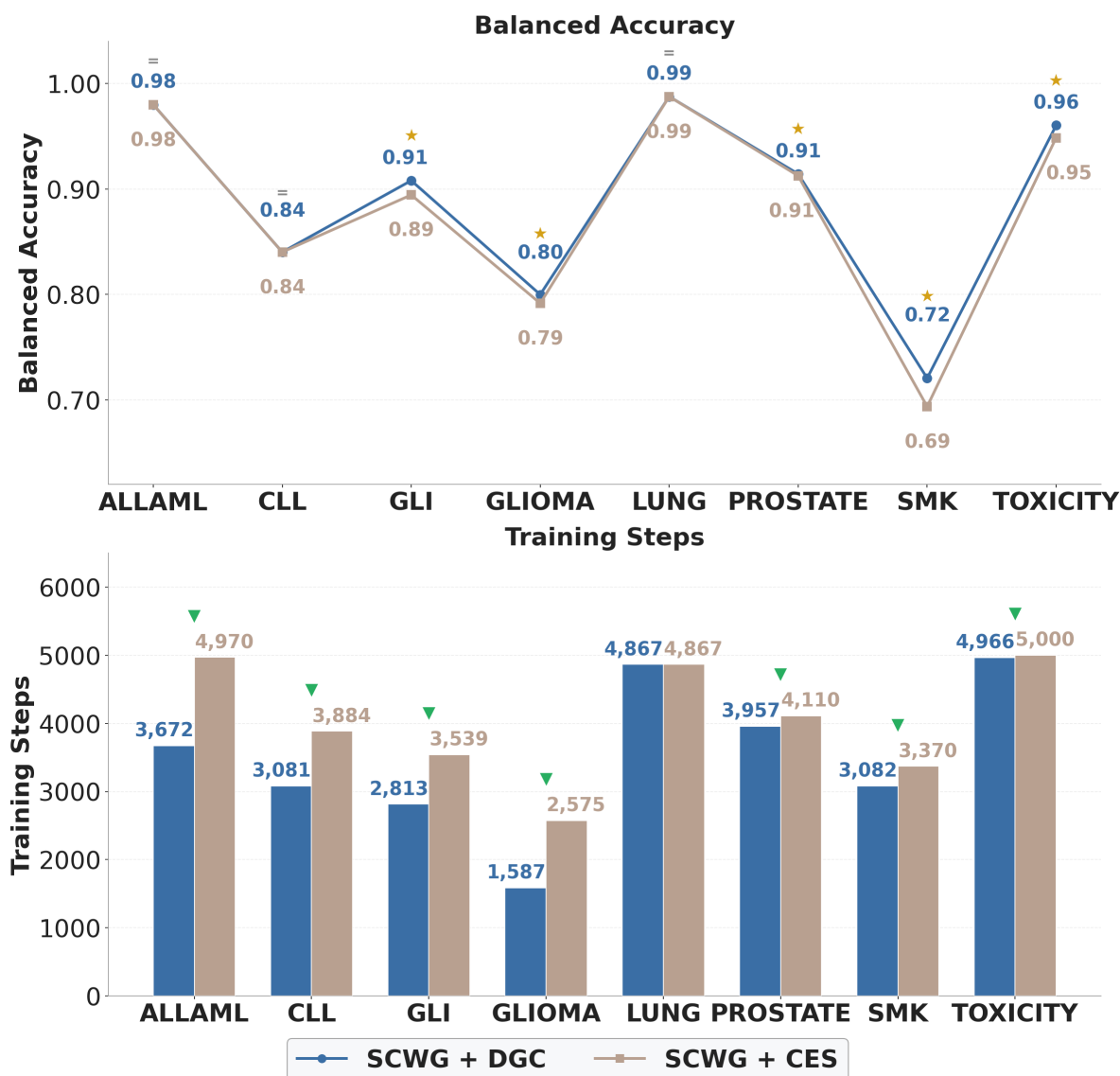
### 3.5. Training Dynamics

Figure 8 compares the training processes of GINet-DGC and MLP on the lung dataset in terms of balanced accuracy and total loss.

GINet-DGC exhibits brief early oscillation, followed by rapid stabilization. Its training and validation accuracy increase steadily after the initial fluctuation and consistently

remain higher than those of MLP. It maintains higher validation accuracy and lower loss throughout. Meanwhile, MLP converges more slowly and exhibits a larger gap between the training and validation curves, suggesting weaker generalization performance.

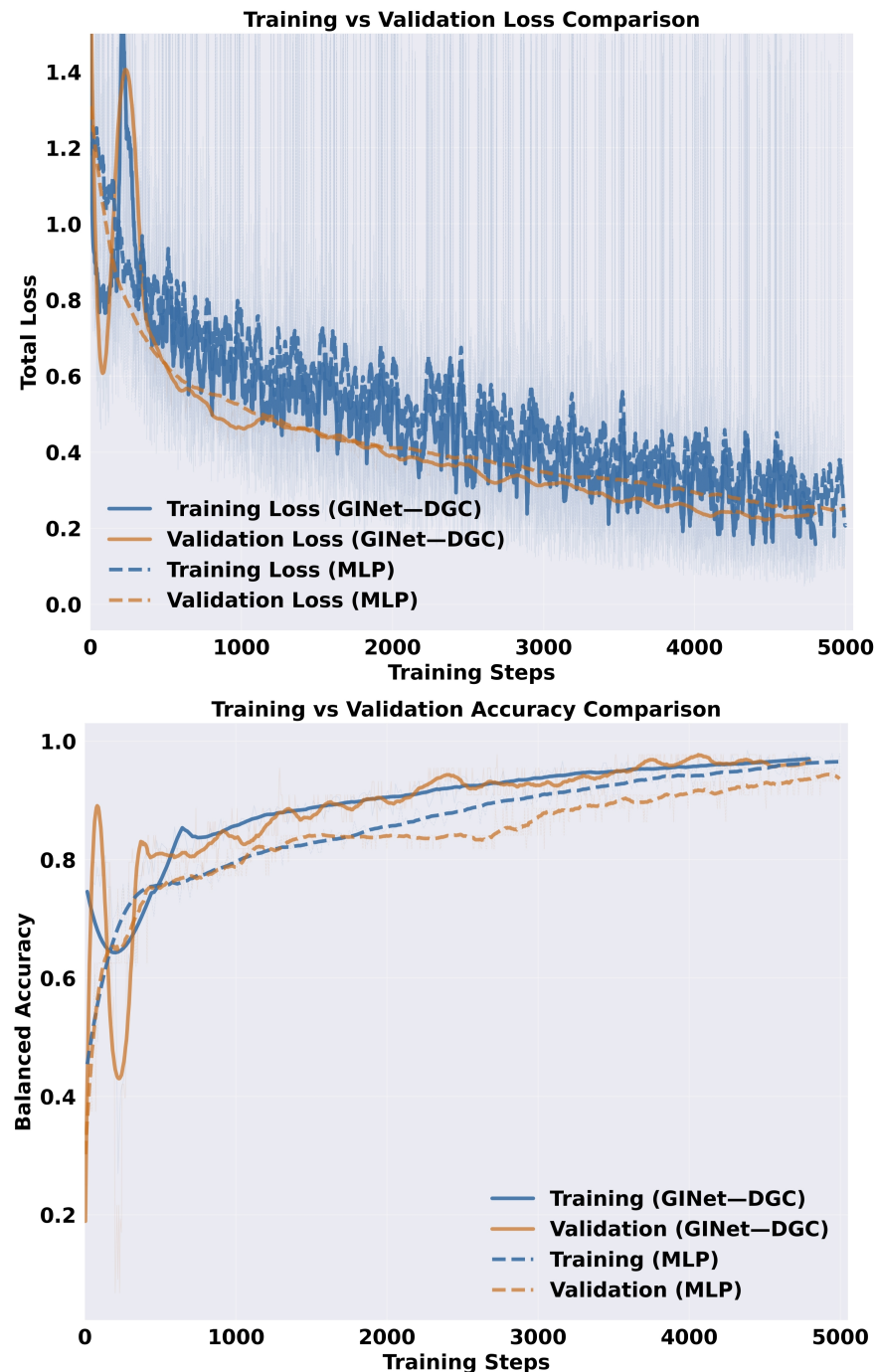
### Ablation Study: SCWG + DGC vs SCWG + Conventional Early Stopping



**Figure 7.** Ablation study on early stopping strategies. The figure compares the performance of the proposed Dynamic Generalization Control (DGC) against conventional early stopping (CES) when combined with the Structure-Constrained Weight Generator (SCWG). The top panel (line chart) illustrates the balanced accuracy, while the bottom panel (bar chart) presents the total training steps required across eight molecular datasets. Blue plots/bars denote SCWG + DGC, and taupe plots/bars denote SCWG + CES.

Prior work shows that early training dynamics can exhibit qualitatively distinct regimes depending on the learning rate, depth, and width [44]. The brief oscillation observed here is followed by rapid stabilization.

Overall, GINet-DGC demonstrates faster convergence and greater stability than MLP, with smoother optimization trajectories and a smaller training–validation gap, supporting its robustness in high-dimensional, small-sample biomedical prediction tasks.



**Figure 8.** Comparison of training dynamics between GINet-DGC and the multilayer perceptron (MLP) on the lung dataset. The upper panel presents the training and validation loss over 5000 training steps, while the lower panel illustrates the training and validation balanced accuracy. Compared with MLP, GINet-DGC converges more rapidly and exhibits more stable optimization dynamics throughout training, indicating improved convergence efficiency and generalization retention.

#### 4. Discussion

The empirical results demonstrate that GINet-DGC effectively addresses the fundamental challenges of High-Dimensional Low-Sample Size (HDLSS) learning by combining structure-constrained weight generation with adaptive generalization monitoring. Without the need for external pretraining or data augmentation, our framework attempts to leverage feature-wise structural priors to introduce beneficial mathematical constraints.

This methodology helps regularize the classification search space in scenarios where sample sizes are substantially restricted and features exhibit high redundancy.

A noteworthy structural insight from the ablation study is that individual prior branches show task-dependent importance, reflecting the distinct geometric or categorical regularities inherent in heterogeneous tabular data. This underscores the necessity of our adaptive fusion layer, which dynamically calibrates multi-view representations instead of assuming static, uniform contributions.

The underlying weight generation and trajectory monitoring mechanisms in GINet-DGC theoretically provide an open avenue for adaptation to broader tabular tasks. The framework may also be extended to other scenarios characterized by high-dimensional, small-sample constraints, such as clinical biomarker identification and related prediction tasks in alternative scientific domains.

Future research will expand in four strategic directions: (1) investigating the direct biological interpretation and visualization of the generated projection weights to map downstream molecular patterns; (2) validating the framework's scalability on larger, more diverse tabular datasets to explore its broader generalization limits; (3) developing a dynamic balancing mechanism that adaptively adjusts the relative contributions of individual OFI metric terms during training, thereby reducing the reliance on fixed hyperparameters or validation-based search procedures; and (4) systematically evaluating hyperparameter-search robustness across multiple random seeds and larger search budgets to assess the stability of the selected hyperparameter regions.

## 5. Conclusions

In this study, we proposed GINet-DGC to alleviate the optimization and over-iteration challenges commonly encountered in High-Dimensional Low-Sample Size learning. By using a Structure-Constrained Weight Generator to synthesize projection weights from multi-view structural priors, the framework helps constrain the feature-interaction search space under limited supervision. Concurrently, the Overfitting-aware Index is designed to enable adaptive training termination before performance shows clear signs of deterioration. Preliminary evaluations across eight typical biomedical datasets suggest that our method exhibits promising classification performance and relatively stable generalization behavior. These findings indicate that GINet-DGC holds promise as a viable and structurally constrained modeling strategy for data-scarce bioinformatics benchmarks. However, all experiments in the present study are limited to biomedical gene-expression datasets. Consequently, the applicability of GINet-DGC to other scientific domains has not yet been empirically established and should be regarded as an important direction for future investigation.

**Author Contributions:** Conceptualization, X.Z.; methodology, X.Z.; investigation, X.Z.; writing—original draft preparation, X.Z.; methodological guidance, comparative experiment solving, and experimental recording, Y.S.; comparative experiments and experimental-result recording, S.S.; data collection, manuscript revision, and content review, D.F.; supervision, experimental guidance, and manuscript-writing guidance, L.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by Shanghai Municipal Commission of Economy and Informatization, Grant No. E1421H1E01.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The datasets used in this study are publicly available benchmarks described in Section 2.5.1 and Appendix A. No new datasets were generated. The source code for replication is available at [https://github.com/user123h/Global\\_Interaction](https://github.com/user123h/Global_Interaction) (accessed on 1 June 2026).

**Acknowledgments:** During the preparation of this manuscript, the authors used ChatGPT, version GPT-5.5 Instant (OpenAI) solely to improve the language quality and clarity of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

**Conflicts of Interest:** Authors Yang Sheng and Sijie Shen were employed by the company General Mathematical-Physics Technology. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Abbreviations

|       |                                        |
|-------|----------------------------------------|
| AI    | Artificial Intelligence                |
| BA    | Balanced Accuracy                      |
| DGC   | Dynamic Generalization Control         |
| DNN   | Deep Neural Network                    |
| HDLSS | High-Dimensional Low-Sample Size       |
| NMF   | Non-negative Matrix Factorization      |
| OFI   | Overfitting-aware Index                |
| SCWG  | Structure-Constrained Weight Generator |

## Appendix A. Detailed Dataset Description

To support reproducibility, this appendix reports the sample size and feature dimensionality of the eight public biomedical gene-expression datasets used in this study. All datasets originate from publicly available bioinformatics benchmark studies [24], and their standardized versions can be obtained from online repositories.

These datasets are widely used in biomedical benchmarking for gene-expression pattern recognition, cancer type diagnosis support, and drug-toxicity prediction. They are characterized by limited sample sizes and extremely high feature dimensionality, making them representative HDLSS benchmarks. The detailed dataset statistics are summarized in Table A1.

**Table A1.** Summary of datasets used in this study.

| Name     | Size | Features |
|----------|------|----------|
| Allaml   | 72   | 7129     |
| CLL      | 111  | 11,340   |
| Gli      | 85   | 22,283   |
| Glioma   | 50   | 4434     |
| Lung     | 203  | 3312     |
| Prostate | 102  | 5966     |
| Smk      | 187  | 19,993   |
| Toxicity | 171  | 5748     |

The dataset selection follows a reproducibility-oriented criterion. We only include real-world datasets whose feature matrices and labels are publicly accessible and can be independently reconstructed without relying on author-generated samples, private preprocessing artifacts, or non-public data augmentation.

## Appendix B. Explanation of Ablation Effectiveness

To evaluate the robustness of the proposed model against parameter initialization sensitivity, we compared four commonly used initialization methods (Kaiming, NMF, principal component analysis (PCA), and Xavier) with the proposed Structure-Constrained Weight Generator (SCWG) of GINet-DGC. For fair comparison, all baselines were trained under identical configurations, including consistent optimization hyperparameters and

the same DGC adaptive stopping strategy. The balanced accuracy results on eight HDLSS datasets are presented in Table A2.

Experimental results demonstrate that GINet-DGC achieves competitive performance compared with conventional initialization strategies. This indicates that the model's structural priors effectively constrain the solution space, mitigating sensitivity to random initialization and ensuring stable convergence. In contrast, general-purpose methods like Kaiming and Xavier show instability in specific tasks, while simple matrix decomposition lacks sufficient expressiveness. Overall, the SCWG mechanism demonstrates robustness akin to a regularization prior, supporting its reliability in data-scarce regimes.

**Table A2.** Performance comparison against initialization strategies on eight datasets.

| Initialization   | Gli                 | Smk                 | Allaml              | CLL                 | Glioma              | Prostate            | Toxicity            | Lung                |
|------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| Kaiming          | 83.45 ± 5.18        | 70.78 ± 7.46        | 88.06 ± 4.63        | 73.32 ± 6.70        | 65.50 ± 7.41        | 90.47 ± 7.97        | 83.51 ± 7.74        | 95.20 ± 2.21        |
| NMF              | 82.43 ± 8.85        | 66.14 ± 2.56        | 90.15 ± 3.08        | 70.50 ± 9.15        | 64.67 ± 8.34        | 88.67 ± 8.14        | 84.71 ± 6.28        | 95.16 ± 2.33        |
| PCA              | 79.53 ± 10.34       | 60.66 ± 7.89        | 93.78 ± 5.99        | 77.91 ± 6.82        | 63.83 ± 18.21       | 89.27 ± 5.81        | 93.97 ± 3.54        | 95.95 ± 1.84        |
| Xavier           | 87.33 ± 8.10        | 70.67 ± 7.75        | 93.78 ± 5.10        | 81.47 ± 5.73        | 63.83 ± 8.16        | <b>92.27 ± 5.94</b> | 83.75 ± 7.75        | 96.88 ± 1.80        |
| <b>GINet-DGC</b> | <b>90.80 ± 5.69</b> | <b>72.06 ± 8.01</b> | <b>98.00 ± 4.08</b> | <b>84.02 ± 4.48</b> | <b>79.99 ± 6.80</b> | 91.44 ± 9.32        | <b>96.04 ± 4.27</b> | <b>98.77 ± 0.46</b> |

*Note.* Values denote balanced accuracy (%) reported as mean ± standard deviation. Bold values indicate the best-performing result in each column, while the bold model name identifies GINet-DGC as the model proposed in this study. GINet-DGC achieves favorable overall performance in this comparison, supporting the usefulness of its structure-constrained weight generation mechanism.

Specifically, Table A3 reports the granular ablation outcomes of individual structural prior branches across representative datasets, incorporating both means and standard deviations to comprehensively evaluate empirical variability. The results indicate that omitting any prior branch tends to not only induce a noticeable degradation in mean performance but also widen the standard deviation bounds on specific tasks, particularly for the Glioma and Prostate datasets. This behavior implies that the hierarchical semantic, local topological, global distributional, and latent semantic priors likely provide complementary regularizing constraints that contribute to stabilizing predictions.

Concurrently, Table A4 documents the comparative metrics for attention-related variants across eight datasets, integrating mean values and standard deviations to reflect experimental stability. The empirical data reveal that the substantial decline in mean accuracy under the Abl-Attention configuration is coupled with a marked inflation of standard deviations, which frequently exceeds 10% or even 15%. Such pronounced volatility demonstrates that the attention module plays a critical role in stabilizing the integration of these priors, whereas J-AttentionOnly alone appears insufficient to fully substitute for the complete architecture.

**Table A3.** Detailed ablation results of structural prior branches.

| Dataset  | GINet-DGC           | Abl-Multi    | Abl-Cluster   | Abl-Stat      | Abl-Low       |
|----------|---------------------|--------------|---------------|---------------|---------------|
| Gli      | <b>90.80 ± 5.69</b> | 88.16 ± 8.14 | 87.48 ± 7.91  | 86.50 ± 8.10  | 86.00 ± 7.04  |
| CLL      | <b>84.02 ± 4.48</b> | 80.80 ± 3.78 | 83.72 ± 4.34  | 79.23 ± 4.99  | 81.59 ± 5.04  |
| Glioma   | <b>79.99 ± 6.80</b> | 72.83 ± 9.66 | 67.00 ± 12.17 | 70.99 ± 14.31 | 63.16 ± 15.35 |
| Prostate | <b>91.44 ± 9.32</b> | 90.44 ± 8.39 | 88.18 ± 8.69  | 89.74 ± 7.99  | 86.45 ± 8.20  |
| Toxicity | <b>96.04 ± 4.27</b> | 94.97 ± 6.31 | 95.81 ± 3.95  | 93.63 ± 4.11  | 95.47 ± 3.79  |
| Lung     | <b>98.77 ± 0.46</b> | 96.81 ± 2.94 | 96.10 ± 2.78  | 96.10 ± 2.64  | 97.02 ± 2.99  |

*Note.* Values denote balanced accuracy (%) reported as mean ± standard deviation. Abl-Multi, Abl-Cluster, Abl-Stat, and Abl-Low denote variants removing the hierarchical semantic, local topological, global distributional, and latent semantic prior branches, respectively.

**Table A4.** Detailed ablation results of attention-related variants on eight datasets.

| Dataset  | GINet-DGC    | Abl-Attention | J-AttentionOnly |
|----------|--------------|---------------|-----------------|
| Gli      | 90.80 ± 5.69 | 79.25 ± 11.73 | 87.48 ± 7.53    |
| CLL      | 84.02 ± 4.48 | 68.78 ± 14.10 | 82.62 ± 8.85    |
| Glioma   | 79.99 ± 6.80 | 31.17 ± 15.77 | 69.49 ± 8.33    |
| Prostate | 91.44 ± 9.32 | 69.36 ± 15.97 | 88.66 ± 8.14    |
| Toxicity | 96.04 ± 4.27 | 69.87 ± 10.08 | 95.04 ± 3.79    |
| Lung     | 98.77 ± 0.46 | 65.36 ± 19.72 | 97.81 ± 3.00    |
| Allaml   | 98.00 ± 4.08 | 83.44 ± 12.84 | 94.67 ± 5.10    |
| Smk      | 72.06 ± 8.01 | 53.22 ± 5.48  | 70.65 ± 7.75    |

Note. Values denote balanced accuracy (%) reported as mean ± standard deviation. Abl-Attention denotes the variant without the attention module, and J-AttentionOnly denotes the attention-only variant.

## Appendix C. DGC Strategy Transferability

### Appendix C.1. DGC Analysis

To verify the universality and transferability of the proposed Dynamic Generalization Control (DGC) strategy, we integrated DGC into the GCondNet framework and conducted comparative experiments on the Prostate, Toxicity, and Lung datasets. Table A5 compares DGC with patience-based early stopping, standard early stopping, and ReduceLROnPlateau under the same experimental settings. The reported balanced accuracies are averaged over five-fold cross-validation.

**Table A5.** Performance comparison of DGC and alternative stopping strategies on three datasets.

| Dataset  | DGC          | Patience-Based Early Stopping | Standard Early Stopping | ReduceLROnPlateau |
|----------|--------------|-------------------------------|-------------------------|-------------------|
| Prostate | 90.78 ± 5.65 | 90.38 ± 4.92                  | 88.27 ± 4.04            | 89.18 ± 3.56      |
| Toxicity | 94.86 ± 2.71 | 94.33 ± 4.38                  | 94.37 ± 1.76            | 93.12 ± 4.73      |
| Lung     | 95.77 ± 4.59 | 95.20 ± 3.10                  | 94.46 ± 4.85            | 92.85 ± 3.49      |

Note. Values denote balanced accuracy (%) reported as mean ± standard deviation. The best performance on each dataset is highlighted in bold.

As shown in Table A5, DGC consistently achieves the best balanced accuracy across the three datasets. This result indicates that the proposed adaptive stopping criterion transfers effectively to another backbone and can improve generalization without changing the underlying network architecture.

### Appendix C.2. Hyperparameter Sensitivity Analysis

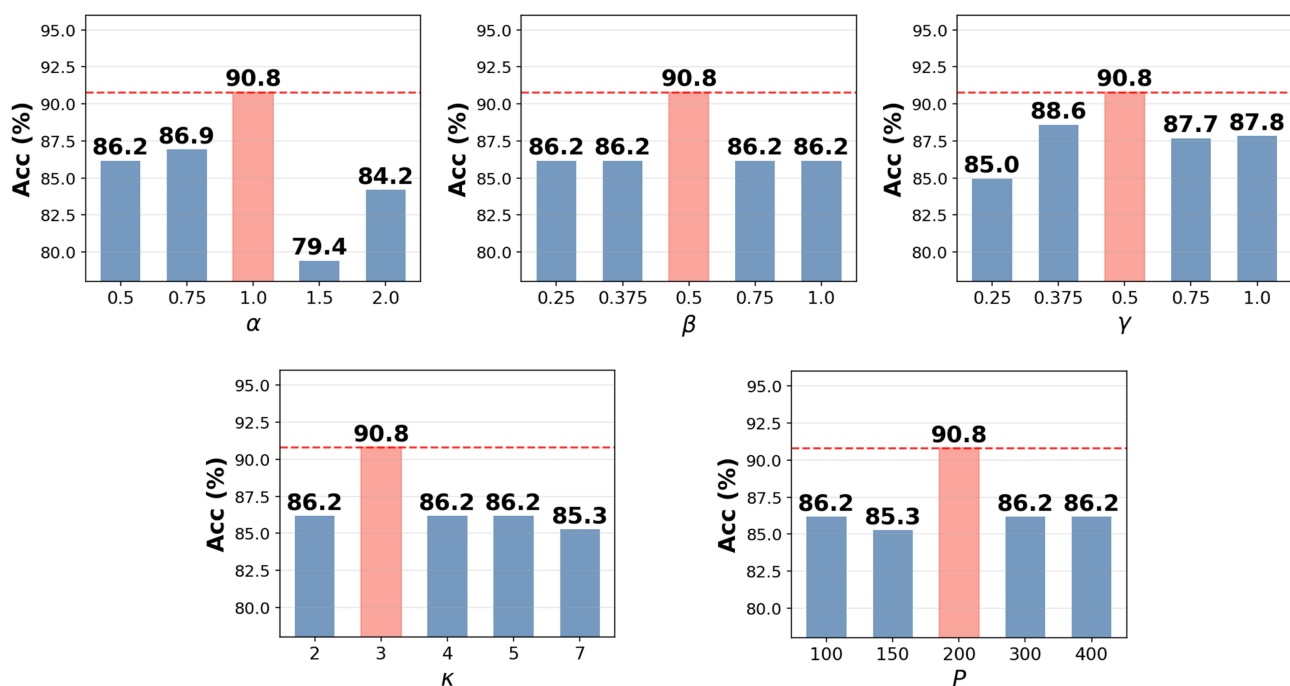
To examine how OFI hyperparameters affect the behavior of DGC, we conducted a one-factor-at-a-time sensitivity study on the GLI dataset. For each parameter, the target value was varied while the remaining hyperparameters were fixed at the optimal configuration obtained via random search.

Figure A1 illustrates the single-parameter sensitivity profiles of the five primary DGC hyperparameters evaluated. The empirical curves reveal clear operational trade-offs and stability characteristics across different dimensions:

- **Accuracy Discrepancy Weight ( $\alpha$ ):** This parameter follows an inverted-U pattern, with performance peaking at the baseline  $\alpha = 1.0$  (90.80%). Decreasing  $\alpha$  to 0.5 or 0.75 reduces accuracy to 86.17% and 86.92%, respectively, while increasing it to 1.5 drops performance to 79.42%. This suggests that while accuracy divergence should be sufficiently monitored to track potential overfitting, an excessive weight can over-respond to transient fluctuations and trigger premature stopping.
- **Loss Discrepancy Weight ( $\beta$ ):**  $\beta$  exhibits a stable plateau across several tested values. While the baseline setting  $\beta = 0.5$  yields 90.80%, the alternative configurations (0.25, 0.375, 0.75, and 1.0) consistently maintain a steady performance baseline of 86.17%. This indicates that once the loss-gap constraint is incorporated, DGC maintains robust performance against localized scaling variations within the explored range.

- **Validation Deterioration Trend Weight ( $\gamma$ ):**  $\gamma$  shows moderate sensitivity, with the setting  $\gamma = 0.5$  achieving the best results (90.80%). Lowering  $\gamma$  to 0.25 limits the model's awareness of recent trajectory trends, reducing accuracy to 84.95%. Conversely, increasing the trend penalty to 0.75 or 1.0 leads to slight degradation (87.68% and 87.83%), as a higher penalty tends to interpret temporary validation plateaus as persistent overfitting.
- **Threshold Tolerance ( $\kappa$ ):**  $\kappa$  presents a steady trend, sustaining a value of 86.17% across most evaluated points ( $\kappa = 2, 4$ , and 5) and slightly retreating to 85.26% at  $\kappa = 7$ . Peak performance remains at the baseline  $\kappa = 3$  (90.80%), confirming that this parameter provides a reasonable and viable boundary for trajectory partitioning.
- **Stopping Patience ( $P$ ):** The patience parameter  $P$  acts as a temporal filter. Restricting patience ( $P = 150$ ) or extending it ( $P = 300$  or 400) limits accuracy to a suboptimal zone (85.26% to 86.17%) compared to the optimum at  $P = 200$  (90.80%). This suggests that an appropriate temporal window is essential: lower patience may cause premature termination, whereas larger patience allows late-stage updates to degrade generalizable features.

### Hyperparameter Sensitivity on GLI Dataset



**Figure A1.** Sensitivity analysis of OFI hyperparameters ( $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\kappa$ , and  $P$ ) on the GLI dataset. The plots show balanced accuracy across the tested values, with the best configuration identified by random search marked as the reference baseline.

In summary, while certain parameters exhibit noticeable sensitivity due to the high-variance nature of small-sample datasets, the DGC framework demonstrates consistent and competitive overall performance once properly configured.

### Appendix D. Statistical Analysis

To provide a clearer assessment of the benchmark results, we conducted a two-step non-parametric statistical analysis. First, a Friedman test was performed to evaluate whether overall performance differences existed among the  $k = 18$  evaluated algorithms

across the  $N = 8$  datasets. The test rejected the null hypothesis with  $\chi^2(17) = 104.46$  and  $p = 1.32 \times 10^{-14}$ , suggesting statistically significant differences within the overall pool of methods. Second, post-hoc pairwise comparisons between GINet-DGC and each baseline were performed using Wilcoxon signed-rank tests. Table A6 summarizes the raw  $p$ -values, average performance differences, and win/tie/loss counts. GINet-DGC recorded 13 comparisons with win/tie/loss counts of 8/0/0 and four comparisons with counts of 7/0/1. These results indicate a consistent trend of improvement over the baseline methods across the evaluated datasets. No correction for multiple pairwise comparisons was applied, and the reported  $p$ -values are unadjusted.

**Table A6.** Wilcoxon signed-rank test for pairwise comparisons between GINet-DGC and each baseline across the eight datasets.

| Baseline           | $p$ -Value            | Avg Diff (%) | W/T/L |
|--------------------|-----------------------|--------------|-------|
| GCondNet           | $7.81 \times 10^{-3}$ | +2.31        | 8/0/0 |
| WPFS               | $7.81 \times 10^{-3}$ | +4.39        | 8/0/0 |
| MLP                | $7.81 \times 10^{-3}$ | +6.28        | 8/0/0 |
| Random Forest      | $7.81 \times 10^{-3}$ | +6.65        | 8/0/0 |
| HyperTab           | $7.81 \times 10^{-3}$ | +8.46        | 8/0/0 |
| DietNetworks       | $7.81 \times 10^{-3}$ | +11.11       | 8/0/0 |
| TabTransformer     | $7.81 \times 10^{-3}$ | +9.00        | 8/0/0 |
| GCN                | $7.81 \times 10^{-3}$ | +11.29       | 8/0/0 |
| CAE                | $7.81 \times 10^{-3}$ | +14.31       | 8/0/0 |
| FsNet              | $7.81 \times 10^{-3}$ | +18.25       | 8/0/0 |
| GATv2              | $7.81 \times 10^{-3}$ | +16.44       | 8/0/0 |
| TabNet             | $7.81 \times 10^{-3}$ | +28.83       | 8/0/0 |
| LassoNet           | $7.81 \times 10^{-3}$ | +48.63       | 8/0/0 |
| DNP                | $1.56 \times 10^{-2}$ | +3.75        | 7/0/1 |
| LightGBM           | $1.56 \times 10^{-2}$ | +6.53        | 7/0/1 |
| SPINN              | $2.34 \times 10^{-2}$ | +3.30        | 7/0/1 |
| TabPFNTTransformer | $3.91 \times 10^{-2}$ | +4.17        | 7/0/1 |

Note. Avg Diff (%) denotes the average balanced-accuracy difference between GINet-DGC and the corresponding baseline across the eight datasets. W/T/L denotes the win/tie/loss counts of GINet-DGC against each baseline. The reported  $p$ -values are unadjusted.

References

1. Baldi, P.; Sadowski, P.; Whiteson, D. Searching for Exotic Particles in High-Energy Physics with Deep Learning. *Nat. Commun.* **2014**, *5*, 4308. [CrossRef] [PubMed]

2. Keith, J.A.; Vassilev-Galindo, V.; Cheng, B.; Chmiela, S.; Gastegger, M.; Müller, K.R.; Tkatchenko, A. Combining Machine Learning and Computational Chemistry for Predictive Insights into Chemical Systems. *Chem. Rev.* **2021**, *121*, 9816–9872. [CrossRef] [PubMed]

3. Feng, F.; He, X.; Wang, X.; Luo, C.; Liu, Y.; Chua, T.S. Temporal Relational Ranking for Stock Prediction. *ACM Trans. Inf. Syst.* **2019**, *37*, 1–30. [CrossRef] [CrossRef]

4. Wang, Z.; Gao, C.; Xiao, C.; Sun, J. MediTab: Scaling Medical Tabular Data Predictors via Data Consolidation, Enrichment, and Refinement. In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, Jeju, Republic of Korea, 3–9 August 2024.

5. Ruan, Y.; Lan, X.; Tan, D.J.; Abdullah, H.R.; Feng, M. P-Transformer: A Prompt-Based Multimodal Transformer Architecture for Medical Tabular Data. In *International Conference on Artificial Intelligence in Medicine*; Springer Nature: Cham, Switzerland, 2025; pp. 375–385.

6. Borisov, V.; Haussermann, J.; Szwarc, T.; Bühler, R.; Müller, S.; Kasneci, E. Deep Neural Networks and Tabular Data: A Survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 7499–7519. [CrossRef] [PubMed]

7. Golling, T.; Heinrich, L.; Kagan, M.; Klein, S.; Leigh, M.; Osadchy, M.; Raine, J.A. Masked Particle Modeling on Sets: Towards Self-Supervised High Energy Physics Foundation Models. *arXiv* **2024**, arXiv:2401.13537.

8. Zou, J.; Huss, M.; Abid, A.; Mohammadi, P.; Torkamani, A.; Telenti, A. A Primer on Deep Learning in Genomics. *Nat. Genet.* **2019**, *51*, 12–18. [CrossRef] [PubMed]

9. Shwartz-Ziv, R.; Armon, A. Tabular Data: Deep Learning Is Not All You Need. *arXiv* **2021**, arXiv:2106.03253. [CrossRef]

10. Gorishniy, Y.; Rubachev, I.; Khrulkov, V.; Babenko, A. Revisiting Deep Learning Models for Tabular Data. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 18932–18943.

11. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: New York, NY, USA, 2016; pp. 785–794. [\[CrossRef\]](#)
12. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3146–3154.
13. Grinsztajn, L.; Oyallon, E.; Varoquaux, G. Why Do Tree-Based Models Still Outperform Deep Learning on Typical Tabular Data? *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 507–520. [\[CrossRef\]](#)
14. McElfresh, D.; Khandagale, S.; Valverde, J.; Prasad, C., V.; Feuer, B.; Hegde, C.; Ramakrishnan, G.; Goldblum, M.; White, C. When Do Neural Nets Outperform Boosted Trees on Tabular Data? *arXiv* **2024**, arXiv:2305.02997. [\[CrossRef\]](#)
15. Song, W.; Shi, C.; Xiao, Z.; Duan, Z.; Xu, Y.; Zhang, M.; Tang, J. AutoInt: Automatic Feature Interaction Learning via Self-Attentive Neural Networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*; ACM: New York, NY, USA, 2019; pp. 1161–1170. [\[CrossRef\]](#)
16. Arik, S.O.; Pfister, T. TabNet: Attentive Interpretable Tabular Learning. *arXiv* **2020**, arXiv:1908.07442. [\[CrossRef\]](#)
17. Huang, X.; Khetan, A.; Cvitkovic, M.; Karnin, Z. TabTransformer: Tabular Data Modeling Using Contextual Embeddings. *arXiv* **2020**, arXiv:2012.06678. [\[CrossRef\]](#)
18. Somepalli, G.; Goldblum, M.; Schwarzschild, A.; Bruss, C.B.; Goldstein, T. SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-Training. *arXiv* **2021**, arXiv:2106.01342. [\[CrossRef\]](#)
19. Singh, D.; Clemente-González, H.; Petrovich, M.; Kawakami, E.; Yamada, M. FsNet: Feature Selection Network on High-dimensional Biological Data. *arXiv* **2020**, arXiv:2001.08322. [\[CrossRef\]](#)
20. Ha, D.; Dai, A.; Le, Q.V. HyperNetworks. In *Proceedings of the International Conference on Learning Representations*, Toulon, France, 24–26 April 2017.
21. Wydmański, W.; Bulenok, O.; Śmieja, M. HyperTab: Hypernetwork Approach for Deep Learning on Small Tabular Datasets. *arXiv* **2023**, arXiv:2304.03543. [\[CrossRef\]](#)
22. Romero, A.; Carrier, P.L.; Erraqabi, A.; Sylvain, T.; Auvolet, A.; Dejoie, E.; Legault, M.A.; Dubé, M.P.; Hussin, J.G.; Bengio, Y. Diet Networks: Thin Parameters for Fat Genomics. *arXiv* **2017**, arXiv:1611.09340. [\[CrossRef\]](#)
23. Margeloiu, A.; Simidjievski, N.; Liò, P.; Jamnik, M. Weight Predictor Network with Feature Selection for Small Sample Tabular Biomedical Data. *arXiv* **2022**, arXiv:2211.15616. [\[CrossRef\]](#)
24. Margeloiu, A.; Simidjievski, N.; Liò, P.; Jamnik, M. GCondNet: A Novel Method for Improving Neural Networks on Small High-Dimensional Tabular Data. *arXiv* **2024**, arXiv:2211.06302. Available online: <https://arxiv.org/abs/2211.06302> (accessed on 1 June 2026). [\[CrossRef\]](#)
25. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [\[CrossRef\]](#)
26. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. In *Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*; IEEE: Piscataway, NJ, USA, 2008; pp. 1322–1328. [\[CrossRef\]](#)
27. Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; Veeramachaneni, K. Modeling Tabular Data Using Conditional GAN. *arXiv* **2019**, arXiv:1907.00503. [\[CrossRef\]](#)
28. Kotelnikov, A.; Baranchuk, D.; Rubachev, I.; Babenko, A. TabDDPM: Modelling Tabular Data with Diffusion Models. In *Proceedings of the International Conference on Machine Learning*, Baltimore, MD, USA, 17–23 July 2022; pp. 17564–17579.
29. Hollmann, N.; Müller, S.; Purucker, L.; Krishnakumar, A.; Körfer, M.; Hoo, S.B.; Schirrmeister, R.T.; Hutter, F. Accurate Predictions on Small Data with a Tabular Foundation Model. *Nature* **2025**, *637*, 319–326. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Gardner, J.; Perdomo, J.C.; Schmidt, L. Large Scale Transfer Learning for Tabular Data via Language Modeling. *arXiv* **2024**, arXiv:2406.12031. [\[CrossRef\]](#)
31. van Breugel, B.; van der Schaar, M. Position: Why Tabular Foundation Models Should Be a Research Priority. In *Proceedings of the 41st International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2024; Volume 235, pp. 48976–48993.
32. Liu, B.; Wei, Y.; Zhang, Y.; Yang, Q. Deep Neural Networks for High Dimension, Low Sample Size Data. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017)*, Melbourne, Australia, 19–25 August 2017; pp. 2287–2293. Available online: <https://www.ijcai.org/> (accessed on 1 June 2026). [\[CrossRef\]](#)
33. Lee, D.D.; Seung, H.S. Learning the Parts of Objects by Non-Negative Matrix Factorization. *Nature* **1999**, *401*, 788–791. [\[CrossRef\]](#) [\[PubMed\]](#)
34. Xie, J.; Girshick, R.; Farhadi, A. Unsupervised Deep Embedding for Clustering Analysis. *arXiv* **2016**, arXiv:1511.06335. [\[CrossRef\]](#)
35. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2023**, arXiv:1706.03762. [\[CrossRef\]](#)
36. Prechelt, L. Early Stopping—But When? In *Neural Networks: Tricks of the Trade*; Orr, G.B., Müller, K.-R., Eds.; Lecture Notes in Computer Science, Volume 1524; Springer: Berlin/Heidelberg, Germany, 1998; pp. 55–69. [\[CrossRef\]](#)
37. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [\[CrossRef\]](#) [\[PubMed\]](#)

38. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
39. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
40. Lemhadri, I.; Ruan, F.; Abraham, L.; Tibshirani, R. LassoNet: A Neural Network with Feature Sparsity. *J. Mach. Learn. Res.* **2021**, *22*, 1–29.
41. Feng, J.; Simon, N. Sparse-Input Neural Networks for High-Dimensional Nonparametric Regression and Classification. *arXiv* **2019**, arXiv:1711.07592. [[CrossRef](#)]
42. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv* **2017**, arXiv:1609.02907. [[CrossRef](#)]
43. Brody, S.; Alon, U.; Yahav, E. How Attentive Are Graph Attention Networks? *arXiv* **2022**, arXiv:2105.14491. [[CrossRef](#)]
44. Kalra, D.S.; Barkeshli, M. Phase Diagram of Early Training Dynamics in Deep Neural Networks: Effect of the Learning Rate, Depth, and Width. *arXiv* **2023**, arXiv:2302.12250. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.