Rethinking Multi-Instance Learning through Graph-Driven Fusion: A Dual-Path Approach to Adaptive Representation

Yu-Xuan Zhang[1], Zhengchun Zhou[1*], Weisha Liu[2], Mingxing Zhang[3]

[1]School of Information Science and Technology, Southwest Jiaotong University
[2]SWJTU-Leeds Joint School, Southwest Jiaotong University
[3]School of Mathematics, Southwest Jiaotong University
inki.yinji@gmail.com, zzc@swjtu.edu.cn, lws_weiss@my.swjtu.edu.cn, mxz@swjtu.edu.cn

Theoretical Analysis

We establish theoretical guarantees for GDF-MIL through four fundamental theorems that form the mathematical backbone of our framework.

Theorem 1 (Stability and Manifold Preservation of ABMM.)**.** *Let $B_i \in \mathbb{R}^{n_i \times d}$ be a bag with $\|B_i\|_F \leq R$, where R is the upper bound of the Frobenius norm of B_i. Consider ABMM that projects B_i into a compact representation $\mathcal{P}(B_i) \in \mathbb{R}^{K_C \times d_K}$ via an encoder and soft clustering. Then ABMM holds the Lipschitz stability, that is there exists a constant $L > 0$ such that for any two bags B_i, B_j, we have*

$$\|\mathcal{P}(B_i) - \mathcal{P}(B_j)\|_F \leq L\|B_i - B_j\|_F, \tag{1}$$

where $L = \|W_{E_2}\|_2 \left(1 + \frac{2}{\tau}\|W_{E_1}\|_2\right)$.

Proof. Let $B_i^E = \mathcal{A}_L(B_i W_{E_1}) W_{E_2}$ is the encoder output for B_i and P_i be the soft assignment matrix obtained via Gumbel-Softmax, where $W_{E_1} \in \mathbb{R}^{d \times d_K}$ and $W_{E_2} \in \mathbb{R}^{d_K \times K_C}$ are the weight matrices of the encoder, and $\mathcal{A}_L$ is the LeakyReLU activation function.

As Gumbel-Softmax is $2/\tau$-Lipschitz continuous (Jang, Gu, and Poole 2017), we have

$$\begin{aligned} \|P_i - P_j\|_F &\leq \frac{2}{\tau}\|B_i^E - B_j^E\|_F \\ &\leq \frac{2}{\tau}\|W_{E_2}\|_2\|W_{E_1}\|_2\|B_i - B_j\|_F. \end{aligned} \tag{2}$$

The ABMM projection is given by $\mathcal{P}(B_i) = P_i^\top B_i^A$. Applying triangle inequality, we have:

$$\begin{aligned} \|\mathcal{P}(B_i) - \mathcal{P}(B_j)\|_F &\leq \|W_{E_2}\|_2\|B_i - B_j\|_F \\ &+ \|P_i^\top - P_j^\top\|_F \cdot \|B_j^A\|_F, \end{aligned} \tag{3}$$

Under bounded norms $\|P_i^\top\|_2 \leq 1$, $\|B_j^A\|_F \leq \|W_{E_1}\|_2 R$, and combining with encoder Lipschitz, we obtain:

$$\begin{aligned} \|\mathcal{P}(B_i) - \mathcal{P}(B_j)\|_F &\leq L\|B_i - B_j\|_F \\ L &= \|W_{E_2}\|_2 \left(1 + \frac{2}{\tau}\|W_{E_1}\|_2\right). \end{aligned} \tag{4}$$

Thus, ABMM is Lipschitz continuous with the claimed constant. □

Theorem 2 (Per-Bag Time Complexity of GDF-MIL)**.** *Let n denote the cardinality of the bag B_i, d the instance feature dimension, d_K the projected feature dimension, K_C the number of clusters in ABMM, and K_N the number of neighbors in DGSL. Then, the overall time complexity of GDF-MIL for a single bag is:*

$$\mathcal{O}(nd_K + K_C^2 d_K + K_C K_N d_K + d_K^2).$$

When $K_C, K_N, d_K \ll n$, this simplifies to $\mathcal{O}(nd_K)$.

Proof. The total complexity is composed of the following:

1) ABMM: Encoder projection takes $\mathcal{O}(ndd_K)$, and soft clustering costs $\mathcal{O}(nK_C d_K)$. As d is usually comparable to d_K, the total cost is $\mathcal{O}(nd_K + nK_C d_K)$.

2) DGSL: Similarity computation among K_C cluster centroids costs $\mathcal{O}(K_C^2 d_K)$. Top-K_N neighbor selection and sparse aggregation via SAGEConv cost $\mathcal{O}(K_C K_N d_K)$.

3) DPFF: Attention over n instances takes $\mathcal{O}(nd_K)$; attention over K_C graph nodes takes $\mathcal{O}(K_C d_K)$ and the fusion process costs $\mathcal{O}(d_K^2)$.

Summing all components yields:

$$\mathcal{O}(nd_K + nK_C d_K + K_C^2 d_K + K_C K_N d_K + d_K^2). \tag{5}$$

As K_C, K_N, d_K are typically much smaller than n, we can omit lower-order terms and obtain $\mathcal{O}(nd_K)$. □

Experiments

This section further comprehensively evaluates GDF-MIL through performance comparison, time cost comparison, convergence analysis, and statistical significance tests.

Performance Comparison

The performance of GDF-MIL and the comparison algorithms on the 4 datasets is shown in Tables 2-13. To provide an intuitive understanding of the overall classification performance, we plotted the violin plot of all algorithms using the ACC metric, as show in Figure 1. In the figure, the symmetrical colored area represents the kernel density estimate, where wider regions indicate higher sample frequency. The embedded black boxplot provides summary statistics:

[*] Corresponding author

Symbol	Meaning
$\mathcal{D} = \{B_i\}_{i=1}^{N}$	The dataset with N bags
$B_i = \{\boldsymbol{x}_{ij}\}_{j=1}^{n_i}$	The i-th bag with n_i instances
$Y_i \in \mathcal{Y} = \{1, \ldots, C\}$	The label of B_i
$\hat{Y}_i$	The predicted label of B_i
$\boldsymbol{x}_{ij} \in \mathbb{R}^d$	The j-th instance in B_i
$\boldsymbol{r}_i$	The bag-level feature of B_i
$\boldsymbol{g}_i$	The graph-level feature of B_i
$\boldsymbol{b}_i$	The fused feature of B_i
N	The number of bags in the dataset
n_i	The number of instances in bag B_i
d	The dimension of $\boldsymbol{x}_{ij}$
C	The number of classes
$\mathcal{A}_L$	The LeakyReLU activation function
$\mathcal{A}_S$	The Sigmoid activation function
$\mathcal{A}_N$	The layer norm
$P_i \in \mathbb{R}^{n_i \times K_C}$	The assignment matrix of soft clustering
$S_i \in \mathbb{R}^{K_C \times K_C}$	The similarity matrix of instances in bag B_i
K_C	The number of soft clusters
K_N	The number of neighbors in graph learning
g_k	The Gumbel noise for differentiable clustering
$\mathcal{N}_i(k)$	The k-th nearest neighbors of each instance
$\mathcal{L}$	Cross-entropy loss function

Table 1: The meaning of some key symbols.

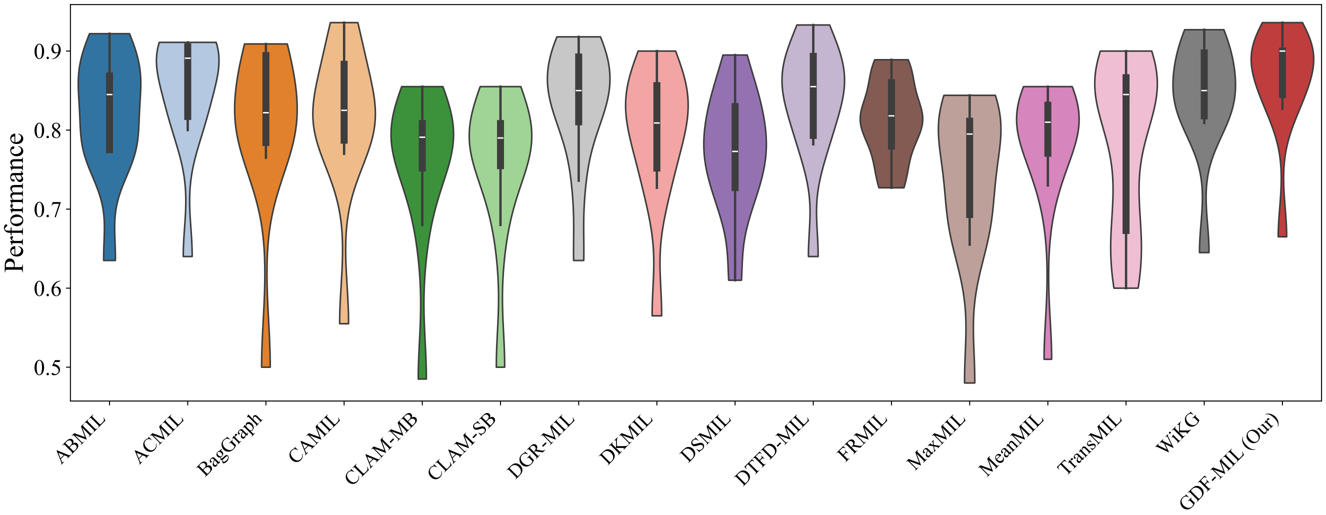


Figure 1: Violin plot of all compared algorithms on 24 datasets in four fields. Note that the MILGNN, RGMIL, and TADGraph algorithms are excluded from this figure, as they failed to achieve effective classification results on multiple datasets (e.g., Web).

the box range corresponds to the first and third quartiles, respectively, and the white horizontal line denotes the median. The results show that our algorithm has the highest median and better classification performance. Next, we will focus on analyzing the experiment results in the table.

For the text datasets, GDF-MIL achieves the best performance on all datasets, demonstrating its strong generalization capability. For example, on the news.aa dataset, GDF-MIL outperformed all baselines by approximately 4% in terms of ACC, F1-score, and AUC. On news.rsh, it achieved perfect classification performance (100% across all metrics). The reason for such achievements is that GDF-MIL fully integrates the strengths of existing intra-bag context feature extraction and bag topology structure mining methods. By achieving an adaptive balance between these two paths, it enhances classification performance and enables efficient

Algorithm	News.aa			News.cwx		
	ACC	F1-Score	AUC	ACC	F1-Score	AUC
ABMIL	.870 ± .076	.862 ± .083	.861 ± .087	.870 ± .084	.866 ± .088	.873 ± .086
ACMIL	.570 ± .196	.483 ± .255	.607 ± .153	.570 ± .164	.518 ± .195	.624 ± .110
BagGraph†	.870 ± .057	.867 ± .059	.871 ± .058	.820 ± .067	.796 ± .102	.799 ± .105
CAMIL	.850 ± .100	.843 ± .110	.850 ± .111	.820 ± .076	.813 ± .088	.825 ± .092
CLAM-MB	.850 ± .087	.846 ± .088	.853 ± .085	.770 ± .091	.760 ± .097	.775 ± .096
CLAM-SB	.840 ± .065	.834 ± .068	.841 ± .065	.790 ± .074	.782 ± .078	.803 ± .066
DGR-MIL	.890 ± .074	.885 ± .081	.890 ± .082	.870 ± .057	.855 ± .079	.849 ± .084
DKMIL†	.670 ± .168	.630 ± .206	.678 ± .119	.810 ± .042	.802 ± .049	.811 ± .057
DSMIL	.510 ± .167	.403 ± .222	.560 ± .133	.570 ± .179	.449 ± .227	.577 ± .145
DTFD-MIL	.800 ± .127	.782 ± .142	.800 ± .121	.800 ± .122	.771 ± .146	.778 ± .135
FRMIL	.720 ± .045	.698 ± .047	.715 ± .035	.710 ± .089	.680 ± .117	.717 ± .099
MaxMIL	.890 ± .082	.885 ± .086	.885 ± .090	.830 ± .027	.822 ± .039	.830 ± .047
MeanMIL	.660 ± .185	.558 ± .261	.643 ± .194	.750 ± .272	.728 ± .308	.796 ± .209
RGMIL†	.890 ± .055	.887 ± .058	.892 ± .063	.790 ± .074	.778 ± .085	.783 ± .087
TAD-Graph†	.510 ± .089	.354 ± .071	.511 ± .025	.520 ± .115	.363 ± .097	.514 ± .032
TransMIL	.610 ± .108	.577 ± .094	.607 ± .078	.650 ± .061	.558 ± .124	.601 ± .077
WiKG†	.890 ± .055	.887 ± .057	.888 ± .061	.880 ± .027	.871 ± .039	.871 ± .051
GDF-MIL (Ours)	**.940 ± .042**	**.938 ± .043**	**.937 ± .045**	**.900 ± .035**	**.895 ± .043**	**.897 ± .048**

Table 2: Performance comparison of GDF-MIL with 18 rivals on text datasets. The symbols † indicate the graph-based methods.

Algorithm	News.mf			News.rsb		
	ACC	F1-Score	AUC	ACC	F1-Score	AUC
ABMIL	.700 ± .146	.685 ± .163	.711 ± .121	.860 ± .114	.857 ± .116	.867 ± .099
ACMIL	.490 ± .074	.342 ± .048	.509 ± .020	.680 ± .189	.651 ± .234	.722 ± .149
BagGraph†	.720 ± .125	.693 ± .129	.721 ± .106	.890 ± .082	.889 ± .082	.893 ± .073
CAMIL	.710 ± .114	.701 ± .113	.710 ± .115	.830 ± .115	.826 ± .118	.845 ± .092
CLAM-MB	.660 ± .129	.598 ± .187	.654 ± .116	.850 ± .106	.844 ± .109	.852 ± .089
CLAM-SB	.650 ± .100	.537 ± .179	.605 ± .110	.850 ± .106	.847 ± .106	.860 ± .082
DGR-MIL	.740 ± .114	.685 ± .190	.709 ± .137	.910 ± .082	.909 ± .082	.913 ± .072
DKMIL†	.750 ± .087	.737 ± .089	.751 ± .072	.860 ± .102	.858 ± .103	.869 ± .082
DSMIL	.560 ± .129	.413 ± .172	.544 ± .099	.520 ± .115	.427 ± .172	.570 ± .103
DTFD-MIL	.590 ± .178	.537 ± .238	.633 ± .149	.770 ± .076	.760 ± .077	.782 ± .057
FRMIL	.620 ± .076	.571 ± .076	.618 ± .048	.750 ± .079	.733 ± .075	.754 ± .048
MaxMIL	.700 ± .079	.686 ± .090	.702 ± .082	.860 ± .119	.859 ± .119	.867 ± .103
MeanMIL	.580 ± .179	.505 ± .227	.617 ± .143	.650 ± .166	.617 ± .206	.685 ± .121
RGMIL†	.790 ± .074	.779 ± .079	.791 ± .074	.850 ± .061	.844 ± .063	.846 ± .064
TAD-Graph†	.480 ± .076	.323 ± .035	.500 ± .000	.600 ± .050	.437 ± .110	.542 ± .066
TransMIL	.650 ± .079	.597 ± .103	.614 ± .094	.550 ± .061	.496 ± .094	.549 ± .049
WiKG†	.740 ± .055	.732 ± .053	.737 ± .058	.870 ± .091	.869 ± .090	.874 ± .081
GDF-MIL (Ours)	**.840 ± .074**	**.833 ± .075**	**.833 ± .076**	**.930 ± .057**	**.929 ± .057**	**.933 ± .049**

Table 3: Performance comparison of GDF-MIL with 18 rivals on text datasets. The symbols † indicate the graph-based methods.

model construction. For image and web page datasets, GDF-MIL only fails to achieve the best result on the Elephant dataset, but its ACC and other metrics are only 1% lower than the first place. In addition, MIL-GNN, RGMIL, and TAD-Graph cannot perform effective classification on the Web dataset. This may be due to the high dimensionality and sparsity of this dataset. For the musk dataset, the best algorithms are ABMIL and DTFD-MIL, which are relatively early MIL studies and can effectively capture the key knowledge in the bag. Nevertheless, GDF-MIL remains one of the top-performing methods on this dataset. In summary, our GDF-MIL demonstrates its extraordinary performance and scalability on all the above datasets.

Time Cost Comparison

We conducted time cost comparison experiments across 24 datasets in four domains: text, web, image, and medicine. Due to the varying characteristics of these datasets, such as high feature sparsity and dimensionality, we evaluated each domain separately. The experiment uses 5-fold cross vali-

Algorithm		News.rsh			News.sc	
	ACC	F1-Score	AUC	ACC	F1-Score	AUC
ABMIL	.920 ± .057	.920 ± .057	.928 ± .050	.860 ± .089	.855 ± .093	.860 ± .089
ACMIL	.770 ± .214	.745 ± .263	.795 ± .173	.630 ± .182	.473 ± .266	.592 ± .205
BagGraph†	.900 ± .035	.898 ± .036	.901 ± .032	.830 ± .076	.825 ± .076	.823 ± .073
CAMIL	.890 ± .082	.889 ± .082	.900 ± .064	.880 ± .057	.879 ± .056	.885 ± .041
CLAM-MB	.900 ± .071	.900 ± .071	.908 ± .060	.810 ± .129	.809 ± .129	.820 ± .122
CLAM-SB	.910 ± .074	.910 ± .074	.917 ± .061	.840 ± .096	.839 ± .096	.850 ± .093
DGR-MIL	.950 ± .035	.949 ± .036	.949 ± .036	.850 ± .079	.848 ± .079	.851 ± .076
DKMIL†	.920 ± .057	.918 ± .058	.921 ± .055	.860 ± .096	.858 ± .096	.858 ± .089
DSMIL	.740 ± .198	.699 ± .253	.752 ± .185	.660 ± .129	.568 ± .210	.637 ± .131
DTFD-MIL	.870 ± .076	.856 ± .090	.851 ± .086	.780 ± .091	.763 ± .093	.777 ± .073
FRMIL	.850 ± .100	.847 ± .100	.852 ± .094	.770 ± .115	.765 ± .117	.775 ± .106
MaxMIL	.950 ± .035	.949 ± .035	.950 ± .033	.870 ± .076	.869 ± .075	.874 ± .069
MeanMIL	.840 ± .114	.837 ± .115	.856 ± .093	.690 ± .119	.645 ± .162	.689 ± .129
RGMIL†	.850 ± .071	.846 ± .074	.852 ± .072	.870 ± .110	.868 ± .111	.873 ± .107
TAD-Graph†	.800 ± .170	.790 ± .179	.819 ± .151	.510 ± .108	.386 ± .151	.530 ± .067
TransMIL	.620 ± .076	.578 ± .094	.610 ± .069	.520 ± .084	.472 ± .053	.503 ± .060
WiKG†	.930 ± .057	.929 ± .057	.938 ± .051	.860 ± .082	.855 ± .086	.859 ± .080
GDF-MIL (Ours)	**1.00 ± .000**	**1.00 ± .000**	**1.00 ± .000**	**.920 ± .067**	**.916 ± .072**	**.914 ± .076**

Table 4: Performance comparison of GDF-MIL with 18 rivals on text datasets. The symbols † indicate the graph-based methods.

Algorithm		News.sm			News.src	
	ACC	F1-Score	AUC	ACC	F1-Score	AUC
ABMIL	.890 ± .042	.887 ± .041	.888 ± .041	.840 ± .065	.837 ± .066	.840 ± .065
ACMIL	.640 ± .156	.491 ± .235	.590 ± .169	.690 ± .167	.661 ± .213	.719 ± .126
BagGraph†	.850 ± .094	.838 ± .104	.840 ± .091	.840 ± .074	.836 ± .078	.843 ± .074
CAMIL	.850 ± .079	.847 ± .078	.851 ± .076	.860 ± .042	.855 ± .041	.854 ± .044
CLAM-MB	.850 ± .071	.848 ± .073	.860 ± .074	.810 ± .065	.808 ± .064	.822 ± .056
CLAM-SB	.860 ± .082	.856 ± .083	.861 ± .078	.830 ± .045	.829 ± .043	.840 ± .037
DGR-MIL	.860 ± .124	.854 ± .134	.862 ± .123	.860 ± .055	.857 ± .056	.857 ± .053
DKMIL†	.900 ± .094	.897 ± .096	.899 ± .095	.820 ± .097	.809 ± .118	.821 ± .115
DSMIL	.630 ± .152	.551 ± .219	.641 ± .142	.670 ± .192	.600 ± .267	.673 ± .176
DTFD-MIL	.750 ± .146	.729 ± .170	.758 ± .142	.800 ± .061	.784 ± .084	.788 ± .088
FRMIL	.760 ± .108	.747 ± .123	.759 ± .111	.800 ± .087	.784 ± .104	.794 ± .083
MaxMIL	.880 ± .057	.877 ± .057	.878 ± .060	.810 ± .108	.808 ± .109	.812 ± .110
MeanMIL	.720 ± .120	.700 ± .133	.741 ± .105	.640 ± .213	.584 ± .276	.674 ± .180
RGMIL†	.850 ± .061	.844 ± .063	.851 ± .058	.880 ± .076	.878 ± .076	.881 ± .075
TAD-Graph†	.520 ± .076	.341 ± .034	.500 ± .000	.490 ± .065	.328 ± .030	.492 ± .019
TransMIL	.610 ± .055	.555 ± .117	.606 ± .068	.590 ± .042	.540 ± .101	.579 ± .064
WiKG†	.880 ± .084	.877 ± .083	.876 ± .083	.860 ± .082	.856 ± .084	.857 ± .085
GDF-MIL (Ours)	**.920 ± .067**	**.918 ± .067**	**.919 ± .067**	**.910 ± .055**	**.908 ± .055**	**.907 ± .056**

Table 5: Performance comparison of GDF-MIL with 18 rivals on text datasets. The symbols † indicate the graph-based methods.

dation, and each fold validation is run 100 times. The final running time is the average running time (ms) of each epoch, as shown in Tables 14–17. Experiment results show that GDF-MIL achieves comparable runtime efficiency to BagGraph, the fastest existing graph-based MIL method, while significantly outperforms high-performance baselines such as WiKG and RGMIL. Moreover, GDF-MIL exhibits favorable time efficiency across all compared MIL methods. In addition, ABMIL shows an excellent time advantage over the other compared algorithms, which is unmatched by any other algorithm. However, GDF-MIL maintains a reasonable runtime compared to other strong baselines, and the additional cost is well compensated by its significant performance gains.

Convergence Analysis

Figure 2-3 shows the convergence performance differences between GDF-MIL and its rivals. For the text dataset, TransMIL is the most convergent algorithm. Despite a high initial loss, it converges within approximately 5 epochs. How-

Algorithm	News.ss			News.tpg		
	ACC	F1-Score	AUC	ACC	F1-Score	AUC
ABMIL	.860 ± .074	.858 ± .074	.864 ± .075	.780 ± .097	.773 ± .105	.782 ± .096
ACMIL	.690 ± .185	.621 ± .268	.693 ± .179	.580 ± .168	.438 ± .227	.567 ± .149
BagGraph†	.870 ± .067	.866 ± .070	.870 ± .071	.790 ± .055	.778 ± .064	.783 ± .059
CAMIL	.840 ± .074	.834 ± .078	.837 ± .068	.780 ± .084	.774 ± .088	.779 ± .091
CLAM-MB	.850 ± .079	.848 ± .079	.856 ± .065	.810 ± .065	.806 ± .066	.814 ± .065
CLAM-SB	.860 ± .082	.859 ± .082	.867 ± .067	.790 ± .096	.786 ± .098	.800 ± .094
DGR-MIL	.850 ± .061	.843 ± .062	.840 ± .060	.790 ± .082	.780 ± .092	.786 ± .095
DKMIL†	.730 ± .157	.703 ± .180	.732 ± .152	.640 ± .156	.604 ± .203	.662 ± .124
DSMIL	.590 ± .156	.440 ± .219	.562 ± .140	.580 ± .115	.471 ± .190	.571 ± .097
DTFD-MIL	.790 ± .119	.786 ± .120	.800 ± .102	.690 ± .108	.648 ± .148	.684 ± .110
FRMIL	.780 ± .057	.772 ± .056	.776 ± .059	.680 ± .076	.661 ± .080	.690 ± .039
MaxMIL	.890 ± .096	.888 ± .096	.885 ± .095	.790 ± .082	.785 ± .084	.798 ± .075
MeanMIL	.610 ± .171	.557 ± .214	.639 ± .140	.610 ± .164	.574 ± .177	.638 ± .141
RGMIL†	.880 ± .104	.879 ± .103	.879 ± .100	.780 ± .091	.778 ± .091	.785 ± .092
TAD-Graph†	.480 ± .076	.323 ± .035	.500 ± .000	.490 ± .082	.343 ± .061	.510 ± .022
TransMIL	.610 ± .042	.593 ± .066	.603 ± .064	.630 ± .157	.626 ± .157	.646 ± .161
WiKG†	.900 ± .061	.899 ± .061	.904 ± .061	.800 ± .079	.795 ± .083	.796 ± .085
GDF-MIL (Ours)	**.930 ± .057**	**.929 ± .057**	**.928 ± .057**	**.830 ± .057**	**.823 ± .061**	**.823 ± .062**

Table 6: Performance comparison of GDF-MIL with 18 rivals on text datasets. The symbols † indicate the graph-based methods.

Algorithm	News.tpmid			News.tpmis		
	ACC	F1-Score	AUC	ACC	F1-Score	AUC
ABMIL	.850 ± .100	.844 ± .105	.846 ± .104	.790 ± .096	.788 ± .097	.799 ± .098
ACMIL	.550 ± .158	.420 ± .214	.564 ± .142	.580 ± .135	.431 ± .197	.554 ± .121
BagGraph†	.840 ± .082	.834 ± .086	.838 ± .083	.750 ± .132	.733 ± .145	.744 ± .149
CAMIL	.880 ± .076	.877 ± .078	.878 ± .078	.800 ± .100	.797 ± .101	.809 ± .099
CLAM-MB	.860 ± .042	.854 ± .048	.858 ± .049	.710 ± .096	.702 ± .108	.734 ± .090
CLAM-SB	.860 ± .042	.858 ± .042	.862 ± .041	.720 ± .076	.714 ± .075	.733 ± .073
DGR-MIL	.870 ± .084	.860 ± .091	.857 ± .093	.810 ± .042	.808 ± .041	.816 ± .035
DKMIL†	.840 ± .042	.838 ± .041	.843 ± .039	.760 ± .124	.732 ± .170	.748 ± .138
DSMIL	.680 ± .168	.650 ± .213	.713 ± .126	.530 ± .135	.419 ± .187	.559 ± .110
DTFD-MIL	.810 ± .108	.803 ± .113	.812 ± .111	.720 ± .067	.689 ± .087	.702 ± .064
FRMIL	.770 ± .125	.765 ± .126	.780 ± .109	.650 ± .094	.628 ± .106	.663 ± .060
MaxMIL	.860 ± .089	.858 ± .090	.861 ± .088	.800 ± .094	.797 ± .096	.814 ± .089
MeanMIL	.800 ± .141	.794 ± .147	.813 ± .120	.620 ± .168	.571 ± .211	.641 ± .157
RGMIL†	.810 ± .065	.808 ± .066	.813 ± .065	.830 ± .084	.830 ± .083	.843 ± .084
TAD-Graph†	.530 ± .104	.385 ± .126	.530 ± .067	.480 ± .076	.323 ± .035	.500 ± .000
TransMIL	.630 ± .076	.558 ± .138	.612 ± .086	.620 ± .027	.564 ± .049	.598 ± .036
WiKG†	.850 ± .071	.848 ± .072	.857 ± .071	.780 ± .115	.770 ± .131	.782 ± .129
GDF-MIL (Ours)	**.890 ± .065**	**.885 ± .068**	**.881 ± .068**	**.860 ± .082**	**.856 ± .085**	**.860 ± .085**

Table 7: Performance comparison of GDF-MIL with 18 rivals on text datasets. The symbols † indicate the graph-based methods.

ever, its poor performance on the text dataset suggests severe overfitting. Our GDF-MIL's convergence performance fluctuates slightly, but this does not affect its high performance. On the Web dataset, GDF-MIL and TransMIL significantly outperformed other algorithms in convergence performance. On other datasets, particularly the Musk dataset, GDF-MIL also significantly outperformed rivals. However, its performance on the current dataset was not leading. This suggests that a key direction for future research on GDF-MIL is to explore strategies for mitigating overfitting.

Statistical Significance Tests

To further analyze the algorithm performance, we used the Friedman test (Demšar 2006; Qian et al. 2023) with a significance level of 5% to assess whether the mean measurement order of each item is significantly different under the null hypothesis for all algorithms. The Friedman test yields test statistics of 163.0562 with a p-value of 0.0000 for the text datasets, indicating statistically significant differences among the compared algorithms and requiring a post-hoc analysis. Note that we did not perform this test on MILGNN

Algorithm	News.trm ACC	News.trm F1-Score	News.trm AUC	Elephant ACC	Elephant F1-Score	Elephant AUC
ABMIL	.790 ± .096	.788 ± .097	.799 ± .098	.870 ± .074	.867 ± .075	.864 ± .076
ACMIL	.550 ± .127	.470 ± .143	.582 ± .064	.895 ± .054	.892 ± .056	.893 ± .054
BagGraph†	.860 ± .055	.853 ± .058	.851 ± .053	.765 ± .070	.763 ± .069	.764 ± .069
CAMIL	.800 ± .079	.795 ± .082	.803 ± .078	.825 ± .061	.823 ± .061	.823 ± .062
CLAM-MB	.780 ± .135	.757 ± .159	.777 ± .132	.775 ± .145	.765 ± .161	.797 ± .113
CLAM-SB	.780 ± .135	.757 ± .159	.777 ± .132	.780 ± .146	.771 ± .163	.801 ± .114
DGR-MIL	.770 ± .135	.717 ± .215	.745 ± .168	.850 ± .079	.847 ± .082	.851 ± .081
DKMIL†	.750 ± .146	.728 ± .168	.758 ± .132	.875 ± .031	.873 ± .031	.872 ± .033
DSMIL	.720 ± .125	.642 ± .197	.680 ± .145	.895 ± .062	.892 ± .065	.891 ± .068
DTFD-MIL	.760 ± .207	.739 ± .234	.770 ± .166	.900 ± .053	.898 ± .055	.897 ± .056
FRMIL	.710 ± .096	.691 ± .098	.713 ± .080	.830 ± .082	.829 ± .083	.832 ± .084
MaxMIL	.800 ± .146	.794 ± .150	.809 ± .120	.815 ± .142	.812 ± .148	.829 ± .117
MeanMIL	.700 ± .162	.684 ± .179	.713 ± .140	.810 ± .110	.808 ± .112	.825 ± .090
MIL-GNN†	N/A	N/A	N/A	.870 ± .048	.868 ± .048	.868 ± .049
RGMIL†	.840 ± .096	.837 ± .098	.843 ± .101	**.910 ± .052**	**.908 ± .052**	**.907 ± .052**
TAD-Graph†	.520 ± .091	.358 ± .070	.511 ± .025	.755 ± .211	.729 ± .059	.772 ± .171
TransMIL	.640 ± .096	.602 ± .108	.636 ± .090	.890 ± .038	.886 ± .042	.887 ± .045
WiKG†	.780 ± .115	.770 ± .131	.782 ± .129	.860 ± .052	.857 ± .053	.857 ± .053
GDF-MIL (Ours)	**.890 ± .065**	**.884 ± .069**	**.877 ± .069**	.900 ± .047	.899 ± .047	.902 ± .047

Table 8: Performance comparison of GDF-MIL with 18 rivals on text and image datasets. The symbols † indicate the graph-based methods.

Algorithm	Tiger ACC	Tiger F1-Score	Tiger AUC	Fox ACC	Fox F1-Score	Fox AUC
ABMIL	.845 ± .067	.844 ± .065	.844 ± .061	.635 ± .052	.630 ± .053	.645 ± .051
ACMIL	.850 ± .040	.846 ± .038	.846 ± .039	.640 ± .045	.635 ± .048	.648 ± .045
BagGraph†	.850 ± .047	.848 ± .046	.848 ± .046	.500 ± .092	.384 ± .133	.538 ± .073
CAMIL	.780 ± .062	.776 ± .061	.779 ± .059	.555 ± .048	.527 ± .037	.572 ± .044
CLAM-MB	.795 ± .102	.794 ± .102	.807 ± .091	.485 ± .055	.364 ± .067	.524 ± .029
CLAM-SB	.790 ± .111	.787 ± .114	.805 ± .094	.500 ± .064	.391 ± .083	.538 ± .036
DGR-MIL	.815 ± .084	.814 ± .084	.814 ± .084	.635 ± .022	.607 ± .056	.627 ± .047
DKMIL†	.835 ± .052	.829 ± .053	.830 ± .052	.565 ± .096	.518 ± .136	.596 ± .060
DSMIL	.820 ± .069	.815 ± .068	.817 ± .068	.610 ± .068	.600 ± .068	.621 ± .070
DTFD-MIL	.860 ± .058	.856 ± .057	.857 ± .057	.640 ± .076	.636 ± .081	.648 ± .076
FRMIL	.820 ± .069	.815 ± .068	.817 ± .068	.805 ± .082	.804 ± .082	.812 ± .078
MaxMIL	.795 ± .057	.794 ± .057	.804 ± .051	.480 ± .048	.355 ± .052	.519 ± .020
MeanMIL	.805 ± .060	.804 ± .060	.812 ± .058	.510 ± .068	.421 ± .087	.547 ± .038
MIL-GNN†	.825 ± .047	.823 ± .044	.823 ± .041	.615 ± .082	.561 ± .121	.616 ± .088
RGMIL†	.870 ± .060	.867 ± .060	.866 ± .062	.660 ± .042	.648 ± .054	.667 ± .041
TAD-Graph†	.850 ± .040	.846 ± .041	.845 ± .044	.520 ± .057	.389 ± .081	.528 ± .038
TransMIL	.870 ± .054	.868 ± .054	.866 ± .052	.600 ± .075	.579 ± .084	.607 ± .068
WiKG†	.850 ± .079	.848 ± .081	.849 ± .082	.645 ± .067	.626 ± .071	.643 ± .054
GDF-MIL (Ours)	**.870 ± .048**	**.868 ± .047**	**.869 ± .047**	**.665 ± .034**	**.659 ± .033**	**.668 ± .029**

Table 9: Performance comparison of GDF-MIL with 18 rivals on image datasets. The symbols † indicate the graph-based methods.

since it was not able to classify this dataset effectively.

Specifically, we adopt the Nemenyi test (Demšar 2006), the most widely used post-hoc method following the Friedman test. Specifically, to find the crucial value at a significance level of 0.05, we first calculate the average ranking based on the performance results in the text datasets. Figure 4 displays the experiment results, where the two subfigures represent statistical analysis results based on various performance comparison experiments. Each critical difference (CD) plot contains a CD value, which indicates that there is no significant difference between algorithms with average rankings lower than this value. Therefore, it verifies the optimal performance of our algorithm under the statistics. For different methods, GDF-MIL emerges as the optimal solution to the MIL problem.

Related Works

Multi-Instance Learning (MIL)

MIL was first introduced by Dietterich *et al.* (Dietterich, Lathrop, and Lozano-Pérez 1997) for their drug activity prediction task and was later extended to deep learning by Ramon *et al.* (Ramon and De Raedt 2000) and Zhou *et al.* (Zhou and Zhang 2002). In the following decade, traditional MIL approaches dominated this field. For example, Zhang *et al.* (Zhang and Zhou 2009) employed a clustering algorithm where key nodes within the bag space are identified, effectively transforming the MIL problem into a single instance classification problem based on the distance between the bag and these representative nodes. Wei *et al.* (Wei, Wu, and Zhou 2016). proposed a large-scale MIL algorithm that leveraged the vector of locally aggregated descriptors and Fisher vectors to extract statistical information from the instance space and map bags into vector representations. Wu *et al.* (Wu et al. 2018) designed a discriminative optimization preparation to further improve the performance and interpretability of the MIL method. In 2018, Wang *et al.* (Wang et al. 2018) revisited deep MIL, and Ilse *et al.* (Ilse, Tomczak, and Welling 2018) proposed an attention-based method, which spurred rapid advancements in deep MIL algorithms. Cui *et al.* (Cui, Chen, and Su 2025) employed an adaptive memory module that estimates overall data distribution by analyzing multi-scale frequency-domain information. Li *et al.* (Li, Li, and Eliceiri 2021) modeled instance relationships through a trainable distance metric and employed self-supervised contrastive learning to extract high-quality representations of bags. Zhang *et al.* (Zhang et al. 2022) developed a double-tier framework that leverages the inherent capabilities of the model, incorporating the concept of pseudo-bags to essentially augment the training bag information. Lin *et al.* (Ling et al. 2024) introduced a mask denoising mechanism to improve attention allocation, thereby improving the detection ability of bag. Fourkioti *et al.* (Fourkioti, De Vries, and Bakal 2024) developed a neighbor-constrained attention mechanism that combines contextual information to enhance the detection and classification of local tumors. Tang *et al.* (Tang et al. 2024b) proposed a re-embedded regional Transformer for online re-embedding of instance features, which effectively captures fine-grained local features and establishes connections across different regions. Chen *et al.* (Chen et al. 2024b) proposed a Bayesian non-parametric model based on the Dirichlet process to estimate the uncertainty of predictions. He *et al.* (He et al. 2025) introduced a pseudo-label attention mechanism to enhance the performance of MIL models. Xie *et al.* (Xie et al. 2025) proposed a prototype similarity-guided feature fusion and hard instance mining approach. Based on this, many excellent deep MIL algorithms have been proposed and used to solve practical applications such as medical image classification (Chen et al. 2025; Li et al. 2024; Xiong et al. 2023; Tang et al. 2024a; Zhu et al. 2024), text classification (Xiao, Liu, and Hao 2024; Yang et al. 2021), and video anomaly detection (Chen et al. 2024a; Liu et al. 2024). These advancements have significantly promoted the rapid development of the MIL field.

Graph-based MIL

Graph-based MIL has become a highly active research area, primarily due to its powerful ability to model intra-bag topology. The core idea of graph-based MIL is to construct a graph structure for each bag, where instances are represented as nodes and their relationships as edges. One prominent approach is the fully connected graph method (Pal et al. 2022; Wang et al. 2025). While designed to comprehensively capture topological structures, this method inherently incurs high computational costs. The second category focuses on key or top-K instances (Zhang et al. 2024; Li et al. 2024), which build graphs faster by using the most informative instances but inevitably ignore the potential topological structure in the bag, thus harming the classification performance. Therefore, this paper aims to share a novel graph-based MIL architecture that preserves the core ideas of the above two methods and enhances scalability by adaptively fusing bag-level and graph-level features.

References

Chen, J.; Li, L.; Su, L.; Zha, Z.-j.; and Huang, Q. 2024a. Prompt-enhanced multiple instance learning for weakly supervised video anomaly detection. In *CVPR*, 18319–18329.

Chen, T.; Wen, J.; Shen, X.; Shen, J.; Deng, J.; Zhao, M.; Xu, L.; Wu, C.; Yu, B.; Yang, M.; et al. 2025. Whole slide image based deep learning refines prognosis and therapeutic response evaluation in lung adenocarcinoma. *NPJ Digital Medicine*, 8(1): 69.

Chen, Y.; Chan, T. H.; Yin, G.; Jiang, Y.; and Yu, L. 2024b. cDP-MIL: Robust multiple instance learning via cascaded Dirichlet process. In *ECCV*, 232–250.

Cui, X.; Chen, W.; and Su, J. 2025. A multiscale frequency domain causal framework for enhanced pathological analysis. In *ICLR*, 1–16.

Demšar, J. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7: 1–30.

Dietterich, T. G.; Lathrop, R. H.; and Lozano-Pérez, T. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2): 31–71.

Fourkioti, O.; De Vries, M.; and Bakal, C. 2024. CAMIL: Context-Aware Multiple Instance Learning for Cancer Detection and Subtyping in Whole Slide Images. In *ICLR*, 1–16.

He, J.; Wang, P.; Cai, J.; Tang, D.; Yao, S.; and Liu, R. 2025. Pseudo-label attention-based multiple instance learning for whole slide image classification. *Engineering Applications of Artificial Intelligence*, 142: 109908.

Ilse, M.; Tomczak, J.; and Welling, M. 2018. Attention-based deep multiple instance learning. In *ICML*, 2127–2136.

Jang, E.; Gu, S.; and Poole, B. 2017. Categorical reparametrization with Gumble-Softmax. In *ICLR*, 1–12.

Algorithm		**Web4**			**Web5**	
	ACC	**F1-Score**	**AUC**	**ACC**	**F1-Score**	**AUC**
ABMIL	.864 ± .072	.759 ± .090	.750 ± .134	.864 ± .072	.759 ± .090	.750 ± .134
ACMIL	.891 ± .052	.818 ± .043	.802 ± .098	.909 ± .032	.845 ± .069	.803 ± .077
BagGraph†	.909 ± .032	.843 ± .044	.827 ± .097	.909 ± .032	.851 ± .068	.817 ± .080
CAMIL	**.900 ± .020**	**.762 ± .163**	**.734 ± .138**	.900 ± .038	.837 ± .072	.801 ± .074
CLAM-MB	.855 ± .087	.742 ± .110	.738 ± .144	.845 ± .052	.687 ± .164	.670 ± .120
CLAM-SB	.855 ± .087	.742 ± .110	.738 ± .144	.855 ± .038	.722 ± .105	.690 ± .089
DGR-MIL	**.900 ± .020**	**.762 ± .163**	**.734 ± .138**	.909 ± .045	.851 ± .085	.817 ± .091
DKMIL†	.809 ± .138	.577 ± .213	.600 ± .137	.800 ± .089	.532 ± .223	.580 ± .179
DSMIL	.773 ± .111	.434 ± .036	.500 ± .000	.764 ± .020	.433 ± .007	.500 ± .000
DTFD-MIL	.855 ± .059	.677 ± .124	.655 ± .098	.873 ± .050	.759 ± .132	.727 ± .116
FRMIL	.873 ± .075	.806 ± .067	.811 ± .098	.818 ± .056	.722 ± .093	.719 ± .091
MaxMIL	.827 ± .093	.620 ± .126	.614 ± .092	.800 ± .025	.571 ± .077	.577 ± .043
MeanMIL	.855 ± .075	.688 ± .137	.666 ± .103	.836 ± .052	.667 ± .155	.653 ± .112
TransMIL	.855 ± .038	.735 ± .072	.719 ± .066	.845 ± .041	.695 ± .121	.677 ± .124
WiKG†	.891 ± .025	.749 ± .153	.720 ± .126	**.918 ± .050**	**.871 ± .078**	**.841 ± .092**
GDF-MIL (Ours)	**.877 ± .069**	**.900 ± .020**	**.762 ± .163**	**.734 ± .138**	.909 ± .032	.851 ± .068

Table 10: Performance comparison of GDF-MIL with 18 rivals on web datasets. The symbols † indicate the graph-based methods.

Algorithm		**Web6**			**Web7**	
	ACC	**F1-Score**	**AUC**	**ACC**	**F1-Score**	**AUC**
ABMIL	.782 ± .075	.777 ± .073	.786 ± .054	.773 ± .056	.760 ± .060	.778 ± .072
ACMIL	.909 ± .056	.778 ± .203	.765 ± .190	.800 ± .076	.791 ± .081	.804 ± .075
BagGraph†	.909 ± .032	.816 ± .093	.783 ± .116	.809 ± .059	.801 ± .056	.810 ± .066
CAMIL	**.936 ± .041**	.873 ± .109	**.858 ± .128**	.809 ± .059	.800 ± .052	.804 ± .049
CLAM-MB	.800 ± .061	.496 ± .059	.531 ± .043	.745 ± .069	.738 ± .066	.762 ± .052
CLAM-SB	.800 ± .061	.496 ± .059	.531 ± .043	.755 ± .076	.745 ± .072	.765 ± .051
DGR-MIL	.918 ± .038	.792 ± .204	.790 ± .200	.809 ± .038	.796 ± .042	.803 ± .052
DKMIL†	.800 ± .109	.523 ± .205	.567 ± .149	.727 ± .107	.623 ± .222	.668 ± .155
DSMIL	.791 ± .076	.469 ± .064	.517 ± .037	.755 ± .052	.705 ± .105	.709 ± .093
DTFD-MIL	.845 ± .052	.624 ± .164	.622 ± .126	.782 ± .075	.750 ± .089	.750 ± .098
FRMIL	.800 ± .109	.677 ± .152	.696 ± .141	.727 ± .056	.700 ± .083	.729 ± .079
MaxMIL	.800 ± .076	.504 ± .087	.537 ± .051	.655 ± .105	.604 ± .082	.629 ± .048
MeanMIL	.827 ± .038	.571 ± .116	.580 ± .088	.764 ± .059	.736 ± .098	.756 ± .083
TransMIL	.809 ± .081	.683 ± .098	.661 ± .074	.655 ± .069	.630 ± .070	.641 ± .059
WiKG†	.927 ± .061	.826 ± .212	.823 ± .192	.818 ± .032	.808 ± .023	.809 ± .013
GDF-MIL (Ours)	**.936 ± .025**	**.876 ± .089**	.857 ± .118	**.827 ± .050**	**.822 ± .045**	**.829 ± .040**

Table 11: Performance comparison of GDF-MIL with 18 rivals on web datasets. The symbols † indicate the graph-based methods.

Li, B.; Li, Y.; and Eliceiri, K. W. 2021. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *CVPR*, 14318–14328.

Li, J.; Chen, Y.; Chu, H.; Sun, Q.; Guan, T.; Han, A.; and He, Y. 2024. Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In *CVPR*, 11323–11332.

Ling, X.; Ouyang, M.; Wang, Y.; Chen, X.; Yan, R.; Chu, H.; Cheng, J.; Guan, T.; Tian, S.; Liu, X.; et al. 2024. Agent aggregator with mask denoise mechanism for histopathology whole slide image analysis. In *ACM MM*, 2795–2803.

Liu, Y.; Ren, J.; Xu, J.; Bai, X.; Kaur, R.; and Xia, F. 2024. Multiple instance learning for cheating detection and localization in online examinations. *IEEE Transactions on Cognitive and Developmental Systems*, 1–12.

Pal, S.; Valkanas, A.; Regol, F.; and Coates, M. 2022. Bag graph: Multiple instance learning using bayesian graph neural networks. In *AAAI*, 7922–7930.

Qian, K.; Min, X.-Y.; Cheng, Y.; and Min, F. 2023. Weight matrix sharing for multi-label learning. *Pattern Recognition*, 136: 109156.

Algorithm	**Web8**			**Web9**		
	ACC	**F1-Score**	**AUC**	**ACC**	**F1-Score**	**AUC**
ABMIL	.773 ± .056	.760 ± .060	.778 ± .072	.782 ± .075	.777 ± .073	.786 ± .054
ACMIL	.836 ± .025	.829 ± .032	.830 ± .035	.811 ± .078	.794 ± .099	.797 ± .100
BagGraph†	.782 ± .093	.779 ± .097	.790 ± .089	.791 ± .076	.785 ± .078	.797 ± .071
CAMIL	.827 ± .059	.823 ± .063	.832 ± .061	.809 ± .075	.798 ± .089	.800 ± .087
CLAM-MB	.773 ± .072	.768 ± .074	.788 ± .058	.791 ± .076	.752 ± .134	.758 ± .123
CLAM-SB	.755 ± .069	.694 ± .173	.724 ± .142	.791 ± .089	.788 ± .088	.799 ± .073
DGR-MIL	**.845 ± .076**	**.840 ± .081**	**.846 ± .072**	.736 ± .126	.721 ± .135	.728 ± .119
DKMIL†	**.845 ± .076**	**.840 ± .081**	**.846 ± .072**	.736 ± .243	.718 ± .276	.773 ± .167
DSMIL	.700 ± .119	.622 ± .182	.665 ± .139	.718 ± .188	.688 ± .214	.733 ± .138
DTFD-MIL	.827 ± .038	.817 ± .044	.817 ± .050	.782 ± .050	.778 ± .052	.793 ± .041
FRMIL	.773 ± .107	.755 ± .120	.768 ± .118	.745 ± .061	.729 ± .078	.762 ± .074
MaxMIL	.700 ± .025	.690 ± .030	.719 ± .031	.691 ± .109	.678 ± .120	.714 ± .111
MeanMIL	.818 ± .064	.815 ± .067	.827 ± .064	.791 ± .089	.789 ± .089	.808 ± .082
TransMIL	.727 ± .140	.695 ± .168	.703 ± .160	.645 ± .059	.626 ± .067	.660 ± .064
WiKG†	.809 ± .050	.799 ± .053	.801 ± .051	.818 ± .056	.801 ± .083	.797 ± .086
GDF-MIL (Ours)	**.845 ± .052**	.835 ± .069	.839 ± .077	**.845 ± .052**	**.835 ± .069**	**.839 ± .077**

Table 12: Performance comparison of GDF-MIL with 18 rivals on web datasets. The symbols † indicate the graph-based methods.

Algorithm	**Musk1**			**Musk2**		
	ACC	**F1-Score**	**AUC**	**ACC**	**F1-Score**	**AUC**
ABMIL	.922 ± .084	.921 ± .085	.921 ± .085	**.910 ± .042**	**.905 ± .038**	**.917 ± .048**
ACMIL	.911 ± .063	.906 ± .066	.899 ± .070	.895 ± .054	.892 ± .056	.893 ± .054
BagGraph†	.822 ± .072	.816 ± .072	.813 ± .067	.830 ± .045	.801 ± .054	.794 ± .057
CAMIL	.833 ± .111	.825 ± .117	.826 ± .115	.770 ± .091	.695 ± .176	.702 ± .158
CLAM-MB	.811 ± .030	.803 ± .037	.801 ± .037	.680 ± .135	.539 ± .163	.596 ± .097
CLAM-SB	.811 ± .063	.799 ± .081	.799 ± .083	.680 ± .164	.561 ± .239	.649 ± .191
DGR-MIL	.867 ± .063	.860 ± .069	.861 ± .072	.870 ± .084	.858 ± .088	.874 ± .090
DKMIL†	.900 ± .046	.892 ± .058	.887 ± .068	.860 ± .042	.821 ± .100	.823 ± .115
DSMIL	.833 ± .088	.826 ± .097	.831 ± .099	.850 ± .079	.825 ± .117	.825 ± .109
DTFD-MIL	**.933 ± .025**	**.931 ± .025**	**.928 ± .024**	.900 ± .087	.890 ± .091	.897 ± .090
FRMIL	.889 ± .039	.885 ± .041	.884 ± .047	.870 ± .027	.858 ± .027	.873 ± .050
MaxMIL	.844 ± .061	.836 ± .067	.831 ± .068	.700 ± .141	.602 ± .197	.650 ± .143
MeanMIL	.833 ± .056	.825 ± .062	.821 ± .063	.730 ± .120	.671 ± .149	.705 ± .148
MIL-GNN†	.900 ± .025	.897 ± .024	.893 ± .021	.800 ± .071	.795 ± .064	.792 ± .059
RGMIL†	.867 ± .030	.863 ± .028	.864 ± .029	.860 ± .042	.859 ± .042	.860 ± .042
TAD-Graph†	.811 ± .030	.796 ± .047	.800 ± .052	.910 ± .022	.886 ± .044	.883 ± .046
GDF-MIL (Ours)	.900 ± .025	.894 ± .028	.886 ± .031	.900 ± .079	.892 ± .078	.895 ± .086

Table 13: Performance comparison of GDF-MIL with 18 rivals on medicine datasets. The symbols † indicate the graph-based methods.

Ramon, J.; and De Raedt, L. 2000. Multi instance neural networks. In *ICML Workshop*, 53–60.

Tang, W.; Yang, Y.-F.; Wang, Z.; Zhang, W.; and Zhang, M.-L. 2024a. Multi-instance partial-label learning with margin adjustment. In *NeurIPS*, 1–24.

Tang, W.; Zhou, F.; Huang, S.; Zhu, X.; Zhang, Y.; and Liu, B. 2024b. Feature re-embedding: Towards foundation model-level performance in computational pathology. In *CVPR*, 11343–11352.

Wang, F.; Xin, J.; Zhao, W.; Jiang, Y.; Yeung, M.; Wang, L.; and Yu, L. 2025. TAD-Graph: Enhancing Whole Slide Image Analysis via Task-Aware Subgraph Disentanglement. *IEEE Transactions on Medical Imaging*, 1–13.

Wang, X.; Yan, Y.; Tang, P.; Bai, X.; and Liu, W. 2018. Revisiting multiple instance neural networks. *Pattern Recognition*, 74: 15–24.

Wei, X.-S.; Wu, J.; and Zhou, Z.-H. 2016. Scalable algorithms for multi-instance learning. *IEEE Transactions on Neural Networks and Learning Systems*, 28(4): 975–987.

Wu, J.; Pan, S.; Zhu, X.; Zhang, C.; and Wu, X. 2018. Multi-instance learning with discriminative bag mapping. *IEEE*

Algorithm	News.aa	News.cwx	News.mf	News.rsb	News.rsh	News.sc	News.sm
ABMIL	193.66	178.14	195.80	185.08	171.30	186.72	179.48
ACMIL	535.15	320.22	334.15	309.53	295.27	319.94	328.22
CAMIL	1148.24	842.11	854.52	1125.30	1119.80	1139.57	1116.86
CLAMMB	480.08	478.70	495.30	451.70	433.59	471.92	447.00
CLAMSB	218.24	226.38	222.29	218.25	187.77	210.72	198.05
DGRMIL	2032.36	1284.18	1327.61	1259.72	1241.01	1272.95	1241.06
DSMIL	300.22	262.68	273.23	255.10	241.43	271.76	274.85
DTFDMIL	909.03	934.07	952.37	952.54	916.04	969.63	942.44
FRMIL	378.53	407.54	428.41	410.38	397.35	418.60	406.16
MaxMIL	162.50	150.35	168.76	143.03	133.31	146.73	155.28
MeanMIL	157.70	153.83	172.22	140.58	129.06	143.11	149.92
TransMIL	1604.29	1236.43	1271.74	1221.36	1242.12	1244.01	1485.52
BagGraph†	732.42	709.79	683.39	686.18	668.19	689.87	688.88
DKMIL†	1155.68	1137.49	1157.61	1141.60	1130.93	1127.09	1129.27
RGMIL†	10404.10	9615.16	10029.53	10342.26	9887.44	9248.66	9158.67
TADGraph†	1737.73	1761.76	1829.61	1770.48	1675.23	1832.29	1764.74
WiKG†	3056.30	3262.06	3323.66	3273.81	3231.04	3313.29	3032.26
GDAMIL	773.78	708.70	725.17	704.92	646.83	686.44	679.00

Table 14: Time cost comparison on the text datasets. Note that the algorithms marked with † are graph-based algorithms. In addition, MILGNN cannot complete the classification task on the text datasets.

Algorithm	News.src	News.ss	News.tpg	News.tpmid	News.tpmis	News.trm
ABMIL	191.37	183.25	184.15	181.58	189.99	192.10
ACMIL	335.19	315.67	332.90	309.38	324.70	341.87
CAMIL	1138.61	1134.42	1134.49	1132.51	1146.28	1145.77
CLAMMB	480.91	481.57	452.73	463.67	486.44	474.26
CLAMSB	226.58	207.40	201.57	208.60	211.99	215.66
DGRMIL	1260.44	1252.16	1253.41	1250.48	1257.04	1262.21
DSMIL	259.24	268.33	266.17	269.94	278.93	276.42
DTFDMIL	999.85	1024.46	1023.14	1013.55	1020.85	957.74
FRMIL	422.09	412.12	412.09	410.29	423.13	421.02
MaxMIL	161.14	159.53	159.47	159.51	162.19	175.33
MeanMIL	149.90	154.88	152.73	149.44	169.36	152.78
TransMIL	1514.29	1551.92	1484.55	1259.99	1271.02	1266.04
BagGraph†	738.51	1262.66	1909.24	1910.43	1936.71	1933.86
DKMIL†	1123.87	1102.58	1103.56	1101.05	1126.08	1115.24
RGMIL†	9267.99	9207.88	9221.77	9126.23	9253.83	9208.95
TADGraph†	1894.04	1845.73	1807.47	1794.49	1884.64	1835.39
WiKG†	3052.43	3070.94	3053.07	3102.35	3111.84	3096.62
GDAMIL	722.06	709.75	722.88	688.41	768.97	736.03

Table 15: Time cost comparison on the text datasets. Note that the algorithms marked with † are graph-based algorithms. In addition, MILGNN cannot complete the classification task on the text datasets.

Transactions on Knowledge and Data Engineering, 30(6): 1065–1080.

Xiao, Y.; Liu, B.; and Hao, Z. 2024. Multi-instance nonparallel tube learning. *IEEE Transactions on Neural Networks and Learning Systems*, 1–12.

Xie, Y.; Liu, Z.; Zhao, J.; and Ma, J. 2025. PHIM-MIL: Multiple instance learning with prototype similarity-guided feature fusion and hard instance mining for whole slide image classification. *Information Fusion*, 117: 102847.

Xiong, C.; Chen, H.; Sung, J. J.; and King, I. 2023. Diagnose like a pathologist: Transformer-enabled hierarchical attention-guided multiple instance learning for whole slide image classification. In *IJCAI*, 1587–1595.

Yang, M.; Zhang, Y.-X.; Wang, X.; and Min, F. 2021. Multi-instance ensemble learning with discriminative bags. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(9): 5456–5467.

Zhang, H.; Meng, Y.; Zhao, Y.; Qiao, Y.; Yang, X.; Coupland, S. E.; and Zheng, Y. 2022. DTFD-MIL: Double-tier feature distillation multiple instance learning for

Algorithm	**Web4**	**Web5**	**Web6**	**Web7**	**Web8**	**Web9**
ABMIL	983.34	1051.01	1037.15	1047.30	974.44	1034.34
ACMIL	1900.85	2011.66	2014.73	2085.95	2373.20	2689.56
CAMIL	3287.01	3478.43	3531.50	4041.55	3750.10	4121.15
CLAMMB	1172.68	1306.96	1316.84	1321.08	1242.16	1301.23
CLAMSB	1494.25	1566.94	1596.47	1588.34	1481.56	1546.61
DGRMIL	2565.32	2729.01	2804.16	2737.25	2530.69	2715.06
DSMIL	1498.81	1568.17	1556.64	1587.56	1490.20	1214.39
DTFDMIL	2034.90	2114.31	2126.84	2114.85	2022.38	2079.82
FRMIL	1635.59	1717.78	1675.13	1729.35	1667.79	1721.56
MaxMIL	1404.08	1474.80	1483.44	1572.09	1621.15	1504.05
MeanMIL	1428.33	1492.32	1500.72	1581.04	1621.13	1514.97
TransMIL	2097.16	2166.68	2175.36	2169.21	2069.37	2144.85
BagGraph†	1528.10	1654.35	1663.45	1683.78	1575.52	1630.58
DKMIL†	3816.68	3978.97	4128.77	4034.43	3453.54	3928.52
WiKG†	4623.71	4789.57	4855.18	4836.04	4767.72	4863.63
GDAMIL	1856.04	2023.88	1982.70	2033.11	1919.79	1945.91

Table 16: Time cost comparison on the web datasets. Note that MILGNN, RGMIL, TADGraph cannot complete the classification task on the web datasets.

Algorithm	**Elephant**	**Tiger**	**Fox**	**Musk1**	**Musk2**
ABMIL	319.42	354.64	319.82	159.25	240.44
ACMIL	437.93	431.89	435.98	198.79	270.60
CAMIL	1951.00	1949.15	1989.71	903.25	1070.68
CLAMMB	478.47	531.98	475.84	237.51	329.60
CLAMSB	422.10	471.37	380.41	199.54	332.33
DGRMIL	2887.51	2893.81	2891.51	1324.29	1520.53
DSMIL	455.26	449.95	444.13	207.25	269.12
DTFDMIL	1619.15	1452.03	1467.22	0.00	838.91
FRMIL	1251.94	1281.21	1247.81	507.17	565.74
MaxMIL	281.29	270.87	287.26	119.68	188.79
MeanMIL	588.50	300.24	283.81	125.18	198.50
MILGNN†	588.50	582.12	579.73	259.29	354.58
TransMIL	2464.40	2462.47	2466.39	1118.07	1302.08
BagGraph†	2566.91	2220.36	2531.13	1130.56	1296.41
DKMIL†	2553.51	2598.31	2511.43	1171.86	1425.78
RGMIL†	19340.55	582.12	17893.84	8314.80	2238.56
TADGraph†	2995.04	19738.07	2964.13	1403.98	1477.79
WiKG*	6132.34	6159.73	6103.14	2814.96	3187.87
GDAMIL	1401.00	1438.26	1444.33	615.42	794.24

Table 17: Time cost comparison on the image and medicine datasets.

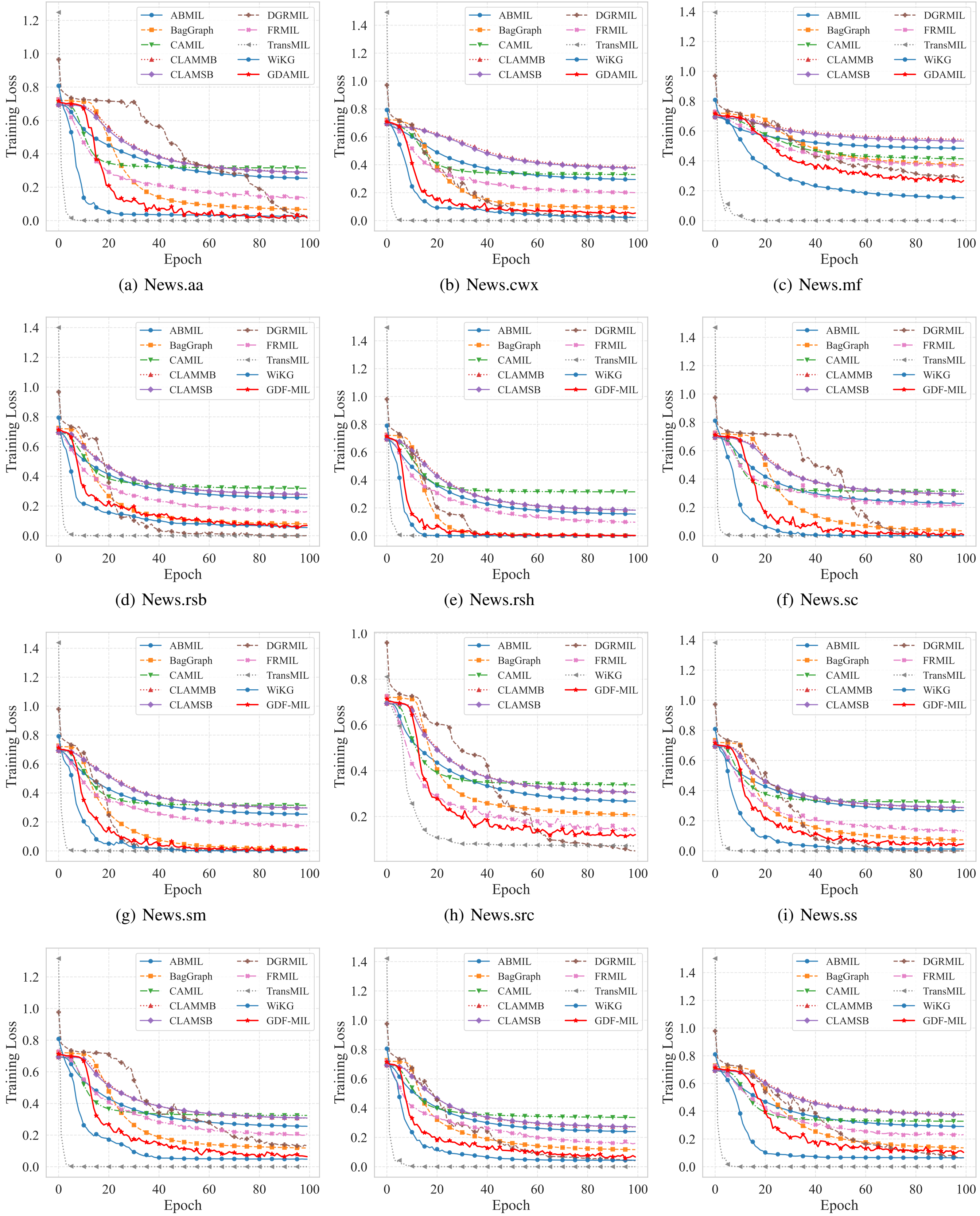
histopathology whole slide image classification. In *CVPR*, 18802–18812.

Zhang, M.-L.; and Zhou, Z.-H. 2009. Multi-instance clustering with applications to multi-instance prediction. *Applied Intelligence*, 31: 47–68.

Zhang, Y.-X.; Zhou, Z.; He, X.; Adhikary, A. R.; and Dutta, B. 2024. Data-driven knowledge fusion for deep multi-instance learning. *IEEE Transactions on Neural Networks and Learning Systems*, 8292–8306.

Zhou, Z.-H.; and Zhang, M.-L. 2002. Neural networks for multi-instance learning. In *ICIIT*, 455–459.

Zhu, W.; Chen, X.; Qiu, P.; Sotiras, A.; Razi, A.; and Wang, Y. 2024. DGR-MIL: Exploring diverse global representation in multiple instance learning for whole slide image classification. In *ECCV*, 333–351.



(a) News.aa

(b) News.cwx

(c) News.mf

(d) News.rsb

(e) News.rsh

(f) News.sc

(g) News.sm

(h) News.src

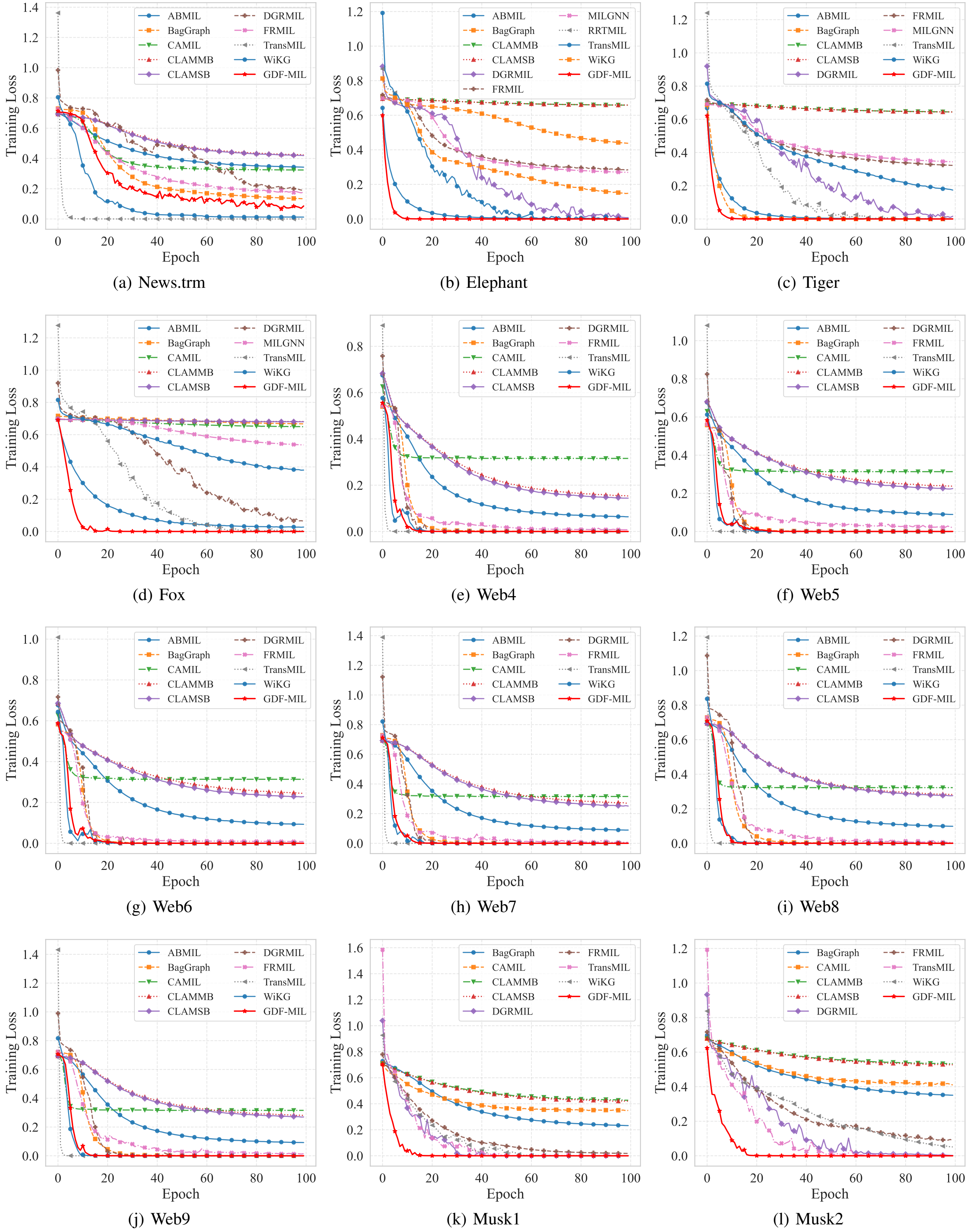
(i) News.ss

(j) News.tpg

(k) News.tpmid

(l) News.tpmis

Figure 2: Convergence analysis of GDF-MIL and rivals. Note that we have removed some algorithms with poor convergence.



(a) News.trm

(b) Elephant

(c) Tiger

(d) Fox

(e) Web4

(f) Web5

(g) Web6

(h) Web7

(i) Web8

(j) Web9

(k) Musk1

(l) Musk2

Figure 3: Convergence analysis of GDF-MIL and rivals. Note that we have removed some algorithms with poor convergence.

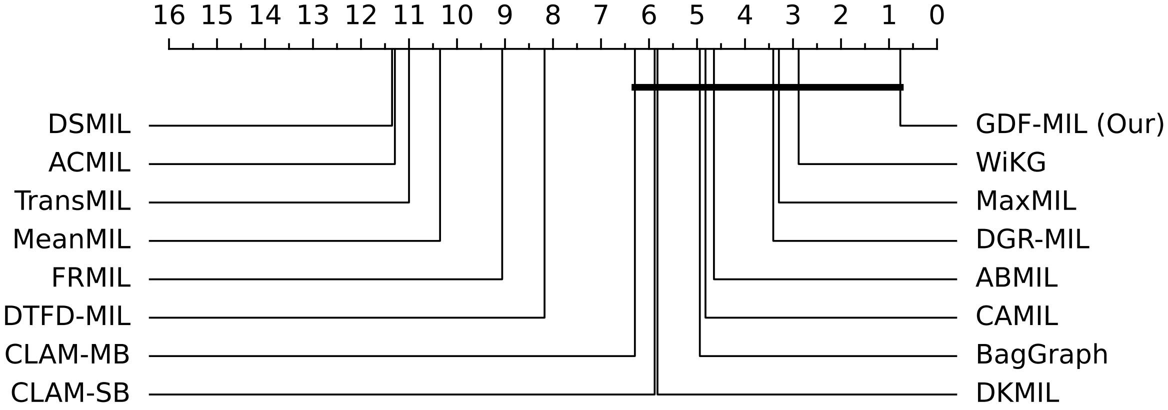

16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
DSMIL
ACMIL
TransMIL
MeanMIL
FRMIL
DTFD-MIL
CLAM-MB
CLAM-SB
GDF-MIL (Our)
WiKG
MaxMIL
DGR-MIL
ABMIL
CAMIL
BagGraph
DKMIL

Figure 4