Supplementary:
Discriminative Deep Generalized Dependency Analysis for Multi-View Data

Debamita Kumar and Pradipta Maji *

SI Derivation of Variational Lower Bound

In this section, the detailed derivation of variational lower bound, reported in (15) of the main manuscript, is presented.

$$
\begin{aligned}
\mathcal{L}_v &= \sum_{\mathbf{h}} Q(\mathbf{h}|\mathbf{v},\mathbf{y})\{\ln P(\mathbf{v},\mathbf{h},\mathbf{y}) - \ln Q(\mathbf{h}|\mathbf{v},\mathbf{y})\} = \sum_{\mathbf{h}} Q(\mathbf{h}|\mathbf{v},\mathbf{y})\{-E(\mathbf{v},\mathbf{h},\mathbf{y}) - \ln Z\} - \sum_{\mathbf{h}} Q(\mathbf{h}|\mathbf{v},\mathbf{y}) \ln Q(\mathbf{h}|\mathbf{v},\mathbf{y}) \\
&= \sum_{\mathbf{h}} Q(\mathbf{h}|\mathbf{v},\mathbf{y})\{-E_a(\mathbf{v},\mathbf{h},\mathbf{y}) - E_b(\mathbf{h}^{L_1 1},\cdots,\mathbf{h}^{L_1 m},\cdots,\mathbf{h}^{L_1 M}) - \ln Z\} - \sum_{\mathbf{h}} Q(\mathbf{h}|\mathbf{v},\mathbf{y}) \ln Q(\mathbf{h}|\mathbf{v},\mathbf{y}) \\
&= \sum_{\mathbf{h}} Q(\mathbf{h}|\mathbf{v},\mathbf{y}) \Bigg\{ \sum_{m=1}^{M}\sum_{i=1}^{V^m}\sum_{j=1}^{H^{1m}} v_i^m w_{ij}^{1m} h_j^{1m} + \sum_{l=1}^{L_1}\sum_{m=1}^{M}\sum_{c=1}^{Y}\sum_{j=1}^{H^{lm}} y_c u_{cj}^{lm} h_j^{lm} + \sum_{l=1}^{L_1-1}\sum_{m=1}^{M}\sum_{j=1}^{H^{lm}}\sum_{k=1}^{H^{(l+1)m}} h_j^{lm} w_{jk}^{(l+1)m} h_k^{(l+1)m} \\
&+ \sum_{m=1}^{M}\sum_{j=1}^{H^{L_1 m}}\sum_{k=1}^{H^1} h_j^{L_1 m} w_{jk}^{(L_1+1)m} h_k^1 + \sum_{l=1}^{L_2}\sum_{c=1}^{Y}\sum_{j=1}^{H^l} y_c u_{cj}^l h_j^l + \sum_{l=1}^{(L_2-1)}\sum_{j=1}^{H^l}\sum_{k=1}^{H^{(l+1)}} h_j^l w_{jk}^l h_k^{(l+1)} + \sum_{l=1}^{L_1}\sum_{m=1}^{M}\sum_{j=1}^{H^{lm}} b_j^{lm} h_j^{lm} \\
&+ \sum_{l=1}^{L_2}\sum_{j=1}^{H^l} b_j^l h_j^l - \sum_{m=1}^{M}\sum_{j=1}^{H^{L_1 m}} \lambda_m {h_j^{L_1 m}}^2 + \sum_{m=1}^{M-1}\sum_{r=(m+1)}^{M} \gamma_{mr} \Bigg\{ \sum_{j=1}^{H^{L_1 m}} (h_j^{L_1 m} - \overline{h}_j^{L_1 m})(h_j^{L_1 r} - \overline{h}_j^{L_1 r}) \Bigg\}^2 \Bigg\} + \sum_{m=1}^{M}\sum_{i=1}^{V^m} a_i^m v_i^m \\
&+ \sum_{c=1}^{Y} d_c y_c + \sum_{m=1}^{M} \lambda_m - \sum_{\mathbf{h}} Q(\mathbf{h}|\mathbf{v},\mathbf{y}) \ln Q(\mathbf{h}|\mathbf{v},\mathbf{y}) - \ln Z \\
&= \sum_{m=1}^{M}\sum_{i=1}^{V^m}\sum_{j=1}^{H^{1m}} \sum_{h_j^{1m}\in\{0,1\}} q(h_j^{1m}|\mathbf{v},\mathbf{y}) v_i^m w_{ij}^{1m} h_j^{1m} + \sum_{l=1}^{L_1}\sum_{m=1}^{M}\sum_{c=1}^{Y}\sum_{j=1}^{H^{lm}}\sum_{h_j^{lm}} q(h_j^{lm}|\mathbf{v},\mathbf{y}) y_c u_{cj}^{lm} h_j^{lm} \\
&+ \sum_{l=1}^{L_1-1}\sum_{m=1}^{M}\sum_{j=1}^{H^{lm}}\sum_{k=1}^{H^{(l+1)m}}\sum_{h_j^{lm}}\sum_{h_k^{(l+1)m}} q(h_j^{lm}|\mathbf{v},\mathbf{y}) q(h_k^{(l+1)m}|\mathbf{v},\mathbf{y}) h_j^{lm} w_{jk}^{(l+1)m} h_k^{(l+1)m} \\
&+ \sum_{m=1}^{M}\sum_{j=1}^{H^{L_1 m}}\sum_{k=1}^{H^1}\sum_{h_j^{L_1 m}}\sum_{h_k^1} q(h_j^{L_1 m}|\mathbf{v},\mathbf{y}) q(h_k^1|\mathbf{v},\mathbf{y}) h_j^{L_1 m} w_{jk}^{(L_1+1)m} h_k^1 + \sum_{l=1}^{L_2}\sum_{c=1}^{Y}\sum_{j=1}^{H^l}\sum_{h_j^l} q(h_j^l|\mathbf{v},\mathbf{y}) y_c u_{cj}^l h_j^l \\
&+ \sum_{l=1}^{(L_2-1)}\sum_{j=1}^{H^l}\sum_{k=1}^{H^{(l+1)}}\sum_{h_j^l}\sum_{h_k^{(l+1)}} q(h_j^l|\mathbf{v},\mathbf{y}) q(h_k^{(l+1)}|\mathbf{v},\mathbf{y}) h_j^l w_{jk}^{(l+1)} h_k^{(l+1)} + \sum_{l=1}^{L_1}\sum_{m=1}^{M}\sum_{j=1}^{H^{lm}}\sum_{h_j^{lm}} q(h_j^{lm}|\mathbf{v},\mathbf{y}) b_j^{lm} h_j^{lm} \\
&+ \sum_{l=1}^{L_2}\sum_{j=1}^{H^l}\sum_{h_j^l} q(h_j^l|\mathbf{v},\mathbf{y}) b_j^l h_j^l - \sum_{m=1}^{M}\sum_{j=1}^{H^{L_1 m}}\sum_{h_j^{L_1 m}} q(h_j^{L_1 m}|\mathbf{v},\mathbf{y}) \lambda_m {h_j^{L_1 m}}^2 + \sum_{m=1}^{M}\sum_{i=1}^{V^m} a_i^m v_i^m + \sum_{c=1}^{Y} d_c y_c + \sum_{m=1}^{M} \lambda_m \\
&+ \sum_{m=1}^{(M-1)}\sum_{r=(m+1)}^{M} \gamma_{mr} \Bigg\{ \sum_{j=1}^{H^{L_1 m}}\sum_{h_j^{L_1 m}}\sum_{h_j^{L_1 r}} q(h_j^{L_1 m}|\mathbf{v},\mathbf{y}) q(h_j^{L_1 r}|\mathbf{v},\mathbf{y}) (h_j^{L_1 m} - \overline{h}_j^{L_1 m})(h_j^{L_1 r} - \overline{h}_j^{L_1 r}) \Bigg\}^2
\end{aligned}
$$

*The authors are with the Biomedical Imaging and Bioinformatics Lab, Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. E-mail: {debamita_r,pmaji}@isical.ac.in.

$$
\begin{aligned}
&- \sum_{\mathbf{h}} \left\{ \prod_{l=1}^{L_1} \prod_{m=1}^{M} \prod_{j=1}^{H^{lm}} q(h_j^{lm}|\mathbf{v},\mathbf{y}) \prod_{l=1}^{L_2} \prod_{j=1}^{H^l} q(h_j^l|\mathbf{v},\mathbf{y}) \left(\sum_{l=1}^{L_1} \sum_{m=1}^{M} \sum_{j=1}^{H^{lm}} \ln q(h_j^{lm}|\mathbf{v},\mathbf{y}) + \sum_{l=1}^{L_2} \sum_{j=1}^{H^l} \ln q(h_j^l|\mathbf{v},\mathbf{y}) \right) \right\} - \ln Z \\
&= \sum_{m=1}^{M} \sum_{i=1}^{V^m} \sum_{j=1}^{H^{1m}} v_i^m w_{ij}^{1m} \mu_j^{1m} + \sum_{l=1}^{L_1} \sum_{m=1}^{M} \sum_{c=1}^{Y} \sum_{j=1}^{H^{lm}} y_c u_{cj}^{lm} \mu_j^{lm} + \sum_{l=1}^{(L_1-1)} \sum_{m=1}^{M} \sum_{j=1}^{H^{lm}} \sum_{k=1}^{H^{(l+1)m}} \mu_j^{lm} w_{jk}^{(l+1)m} \mu_k^{(l+1)m} \\
&+ \sum_{m=1}^{M} \sum_{j=1}^{H^{L_1 m}} \sum_{k=1}^{H^1} \mu_j^{L_1 m} w_{jk}^{(L_1+1)m} \mu_k^1 + \sum_{l=1}^{L_2} \sum_{c=1}^{Y} \sum_{j=1}^{H^l} y_c u_{cj}^l \mu_j^l + \sum_{l=1}^{(L_2-1)} \sum_{j=1}^{H^l} \sum_{k=1}^{H^{(l+1)}} \mu_j^l w_{jk}^{(l+1)} \mu_k^{(l+1)} + \sum_{m=1}^{M} \sum_{i=1}^{V^m} a_i^m v_i^m \\
&+ \sum_{l=1}^{L_1} \sum_{m=1}^{M} \sum_{j=1}^{H^{lm}} b_j^{lm} \mu_j^{lm} + \sum_{l=1}^{L_2} \sum_{j=1}^{H^l} b_j^l \mu_j^l + \sum_{c=1}^{Y} d_c y_c + \sum_{m=1}^{M} \lambda_m \left(1 - \sum_{j=1}^{H^{L_1 m}} \mu_j^{L_1 m} \right) - \ln Z \\
&+ \sum_{m=1}^{M-1} \sum_{r=(m+1)}^{M} \sum_{j=1}^{H^{L_1 m}} \gamma_{mr} (\mu_j^{L_1 m} - 2\mu_j^{L_1 m} \overline{h}_j^{L_1 m} + \overline{h}_j^{L_1 m})(\mu_j^{L_1 r} - 2\mu_j^{L_1 r} \overline{h}_j^{L_1 r} + \overline{h}_j^{L_1 r}) \\
&+ 2 \sum_{m=1}^{M-1} \sum_{r=(m+1)}^{M} \sum_{j=1}^{H^{L_1 m}-1} \sum_{k=(j+1)}^{H^{L_1 m}} \gamma_{mr} (\mu_j^{L_1 m} - \overline{h}_j^{L_1 m})(\mu_j^{L_1 r} - \overline{h}_j^{L_1 r})(\mu_k^{L_1 m} - \overline{h}_k^{L_1 m})(\mu_k^{L_1 r} - \overline{h}_k^{L_1 r}) \\
&- \sum_{l=1}^{L_1} \sum_{m=1}^{M} \sum_{j=1}^{H^{lm}} \left\{ \mu_j^{lm} \ln \mu_j^{lm} + (1 - \mu_j^{lm}) \ln(1 - \mu_j^{lm}) \right\} - \sum_{l=1}^{L_2} \sum_{j=1}^{H^l} \left\{ \mu_j^l \ln \mu_j^l + (1 - \mu_j^l) \ln(1 - \mu_j^l) \right\}. \qquad (1)
\end{aligned}
$$

SII Derivation of Update Rule for Mean Field Parameters

The mean field parameters ($\boldsymbol{\mu}$) of the proposed D2GDA model can be updated by maximizing the variational lower bound $\mathcal{L}_v$ with respect to $\boldsymbol{\mu}$ for a fixed parameter set $\boldsymbol{\theta}$. So, differentiating $\mathcal{L}_v$ with respect to the mean field parameters for $1 \leq l < L_1$ and $\forall m$ leads to

$$
\begin{aligned}
&\frac{\partial \mathcal{L}_v}{\partial \mu_j^{lm}} = 0 \\
&\Rightarrow \sum_{k=1}^{H^{(l-1)m}} \mu_k^{(l-1)m} w_{kj}^{lm} + \sum_{k=1}^{H^{(l+1)m}} w_{jk}^{(l+1)m} \mu_k^{(l+1)m} + \sum_{p=1}^{Y} y_p u_{pj}^{lm} + b_j^{lm} - \ln \mu_j^{lm} - 1 + \ln(1 - \mu_j^{lm}) + 1 = 0 \\
&\Rightarrow \ln \left(\frac{1 - \mu_j^{lm}}{\mu_j^{lm}} \right) = - \sum_{k=1}^{H^{(l-1)m}} \mu_k^{(l-1)m} w_{kj}^{lm} - \sum_{k=1}^{H^{(l+1)m}} w_{jk}^{(l+1)m} \mu_k^{(l+1)m} - \sum_{p=1}^{Y} y_p u_{pj}^{lm} - b_j^{lm} \\
&\Rightarrow \frac{1}{\mu_j^{lm}} - 1 = exp \left\{ - \sum_{k=1}^{H^{(l-1)m}} \mu_k^{(l-1)m} w_{kj}^{lm} - \sum_{k=1}^{H^{(l+1)m}} w_{jk}^{(l+1)m} \mu_k^{(l+1)m} - \sum_{p=1}^{Y} y_p u_{pj}^{lm} - b_j^{lm} \right\} \\
&\Rightarrow \mu_j^{lm} = \sigma \left(\sum_{k=1}^{H^{(l-1)m}} \mu_k^{(l-1)m} w_{kj}^{lm} + \sum_{k=1}^{H^{(l+1)m}} w_{jk}^{(l+1)m} \mu_k^{(l+1)m} + \sum_{p=1}^{Y} y_p u_{pj}^{lm} + b_j^{lm} \right), \qquad (2)
\end{aligned}
$$

where $\sigma(x) = \dfrac{1}{1 + e^{-x}}$ is the sigmoid function.

Similarly, the update rule for the mean field parameters for layer L_1 and $1 \leq l \leq L_2$ can be obtained as

$$
\begin{aligned}
\mu_j^{L_1 m} &= \sigma \Bigg(\sum_{k=1}^{H^{(L_1-1)m}} \mu_k^{(L_1-1)m} w_{kj}^{L_1 m} + \sum_{k=1}^{H^1} w_{jk}^{(L_1+1)m} \mu_k^1 - \lambda_m + \sum_{r \neq m=1}^{M} \gamma_{mr} (1 - 2\overline{h}_j^{L_1 m})(\mu_j^{L_1 r} - 2\mu_j^{L_1 r} \overline{h}_j^{L_1 r} + \overline{h}_j^{L_1 r}) \\
&+ 2 \sum_{r \neq m=1}^{M} \gamma_{mr} \sum_{k \neq j=1}^{H^{L_1 m}} (\mu_j^{L_1 r} - \overline{h}_j^{L_1 r})(\mu_k^{L_1 m} - \overline{h}_k^{L_1 m})(\mu_k^{L_1 r} - \overline{h}_k^{L_1 r}) + \sum_{c=1}^{Y} y_c u_{cj}^{L_1 m} + b_j^{L_1 m} \Bigg), \ \forall m; \qquad (3)
\end{aligned}
$$

$$
\mu_j^l = \sigma \left(\sum_{k=1}^{H^{(l-1)}} \mu_k^{(l-1)} w_{kj}^{(l-1)} + \sum_{k=1}^{H^{(l+1)}} w_{jk}^l \mu_k^{(l+1)} + \sum_{c=1}^{Y} y_c u_{cj}^l + b_j^l \right), \quad \text{for } 1 \leq l \leq L_2. \qquad (4)
$$

SIII Derivation of Data-Dependent Expectations

The data-dependent expectations can be obtained by differentiating the variational lower bound L_v with respect to the parameter set $\boldsymbol{\theta}$ at equilibrium state of the proposed D2GDA model, which are given by

$$\frac{\partial \mathcal{L}_v}{\partial w_{ij}^{1m}} = v_i^m \mu_j^{1m}; \quad \frac{\partial \mathcal{L}_v}{\partial w_{jk}^{lm}} = \mu_j^{(l-1)m} \mu_k^{lm}, \text{ for } 1 < l \leq L_1; \quad \frac{\partial \mathcal{L}_v}{\partial w_{jk}^{(L_1+1)m}} = \mu_j^{L_1 m} \mu_k^1; \quad \frac{\partial \mathcal{L}_v}{\partial w_{jk}^{l}} = \mu_j^{l} \mu_k^{(l+1)}, \text{ for } 1 \leq l < L_2;$$

$$\frac{\partial \mathcal{L}_v}{\partial u_{cj}^{lm}} = y_c \mu_j^{lm}, \text{ for } 1 \leq l \leq L_1; \quad \frac{\partial \mathcal{L}_v}{\partial u_{cj}^{l}} = y_c \mu_j^{l}, \text{ for } 1 \leq l \leq L_2; \quad \frac{\partial \mathcal{L}_v}{\partial a_i^m} = v_i^m; \quad \frac{\partial \mathcal{L}_v}{\partial b_j^{lm}} = \mu_j^{lm}, \text{ for } 1 \leq l \leq L_1;$$

$$\frac{\partial \mathcal{L}_v}{\partial d_c} = y_c; \quad \frac{\partial \mathcal{L}_v}{\partial b_j^{l}} = \mu_j^{l}, \text{ for } 1 \leq l \leq L_2; \quad \frac{\partial \mathcal{L}_v}{\partial \lambda_m} = \left(1 - \sum_{j=1}^{H^{L_1 m}} \mu_j^{L_1 m}\right); \text{ and}$$

$$\frac{\partial \mathcal{L}_v}{\partial \gamma_{mr}} = \sum_{j=1}^{H^{L_1 m}} (\mu_j^{L_1 m} - 2\mu_j^{L_1 m} \overline{h}_j^{L_1 m} + \overline{h}_j^{L_1 m})(\mu_j^{L_1 r} - 2\mu_j^{L_1 r} \overline{h}_j^{L_1 r} + \overline{h}_j^{L_1 r})$$
$$+ 2 \sum_{j=1}^{H^{L_1 m}-1} \sum_{k=(j+1)}^{H^{L_1 m}} (\mu_j^{L_1 m} - \overline{h}_j^{L_1 m})(\mu_j^{L_1 r} - \overline{h}_j^{L_1 r})(\mu_k^{L_1 m} - \overline{h}_k^{L_1 m})(\mu_k^{L_1 r} - \overline{h}_k^{L_1 r}) \quad \forall m, r. \tag{5}$$

SIV Derivation of Conditional Distributions

The conditional distribution, corresponding to the proposed D2GDA model, over hidden nodes of layer $1 \leq l < L_1$ and $\forall m$ for a fixed parameter set $\boldsymbol{\theta}$ is given by

$$\begin{aligned}
&P(h_j^{lm} | \mathbf{h}^{(l-1)m}, \mathbf{h}^{(l+1)m}, \mathbf{y}) \\
&= \frac{e^{h_j^{lm}\left(\sum_{k=1}^{H^{(l-1)m}} h_k^{(l-1)m} w_{kj}^{lm} + \sum_{k=1}^{H^{(l+1)m}} w_{jk}^{(l+1)m} h_k^{(l+1)m} + \sum_{p=1}^{Y} y_p u_{pj}^{lm} + b_j^{lm}\right)}}{\sum_{\mathbf{h}^{lm}} e^{h_j^{lm}\left(\sum_{k=1}^{H^{(l-1)m}} h_k^{(l-1)m} w_{kj}^{lm} + \sum_{k=1}^{H^{(l+1)m}} w_{jk}^{(l+1)m} h_k^{(l+1)m} + \sum_{p=1}^{Y} y_p u_{pj}^{lm} + b_j^{lm}\right)}} \\
&= \frac{e^{\left\{\sum_{k=1}^{H^{(l-1)m}} h_k^{(l-1)m} w_{kj}^{lm} + \sum_{k=1}^{H^{(l+1)m}} w_{jk}^{(l+1)m} h_k^{(l+1)m} + \sum_{p=1}^{Y} y_p u_{pj}^{lm} + b_j^{lm}\right\}}}{1 + e^{\left\{\sum_{k=1}^{H^{(l-1)m}} h_k^{(l-1)m} w_{kj}^{lm} + \sum_{k=1}^{H^{(l+1)m}} w_{jk}^{(l+1)m} h_k^{(l+1)m} + \sum_{p=1}^{Y} y_p u_{pj}^{lm} + b_j^{lm}\right\}}} \\
&= \sigma\left(\sum_{k=1}^{H^{(l-1)m}} h_k^{(l-1)m} w_{kj}^{lm} + \sum_{k=1}^{H^{(l+1)m}} w_{jk}^{(l+1)m} h_k^{(l+1)m} + \sum_{p=1}^{Y} y_p u_{pj}^{lm} + b_j^{lm}\right)
\end{aligned} \tag{6}$$

Similarly, the conditional distributions over visible and hidden nodes for the rest of the layers of the proposed model can be obtained as

$$P(h_j^{L_1 m} | \mathbf{h}^{(L_1-1)m}, \mathbf{h}^1, \mathbf{h}_{-j}^{L_1 m}, \mathbf{h}^{L_1 r}, \mathbf{y}) = \sigma\Bigg(\sum_{k=1}^{H^{(L_1-1)m}} h_k^{(L_1-1)m} w_{kj}^{L_1 m} + \sum_{k=1}^{H^1} w_{jk}^{(L_1+1)m} h_k^1 + \sum_{c=1}^{Y} y_c u_{cj}^{L_1 m} + b_j^{L_1 m}$$
$$- \lambda_m + \sum_{r \neq m=1}^{M} \gamma_{mr}(1 - 2\overline{h}_j^{L_1 m})\left(h_j^{L_1 r} - \overline{h}_j^{L_1 r}\right)^2 + 2 \sum_{r \neq m=1}^{M} \sum_{k \neq j=1}^{H^{L_1 m}} \gamma_{mr}(h_j^{L_1 r} - \overline{h}_j^{L_1 r})(h_k^{L_1 m} - \overline{h}_k^{L_1 m})(h_k^{L_1 r} - \overline{h}_k^{L_1 r})\Bigg), \ \forall m; \tag{7}$$

where $\mathbf{h}_{-j}^{L_1 m}$ represents modality-specific hidden representation, corresponding to L_1-th layer of m-th modality, consisting values of all the nodes except $h_j^{L_1 m}$.

$$P(h_j^l|\mathbf{h}^{(l-1)},\mathbf{h}^{(l+1)},\mathbf{y}) = \sigma\Bigg(\sum_{k=1}^{H^{(l-1)}} h_k^{(l-1)} w_{kj}^{(l-1)} + \sum_{c=1}^{Y} y_c u_{cj}^l + b_j^l + \sum_{k=1}^{H^{(l+1)}} w_{jk}^l h_k^{(l+1)}\Bigg), \text{ for } 1 \le l \le L_2; \tag{8}$$

$$P(v_i^m|\mathbf{h}^{1m}) = \sigma\left(\sum_{j=1}^{H^{1m}} w_{ij}^{1m} h_j^{1m} + a_i^m\right), \ \forall m; \tag{9}$$

$$P(y_c|\mathbf{h}^{11},\cdots,\mathbf{h}^{L_1 M},\mathbf{h}^1,\cdots,\mathbf{h}^{L_2}) = \frac{e^{X_c}}{\sum_{\bar{c}=1}^{Y} e^{X_{\bar{c}}}}; \text{ where, } X_c = \sum_{l=1}^{L_c}\sum_{m=1}^{M}\sum_{j=1}^{H^{lm}} u_{cj}^{lm} h_j^{lm} + \sum_{l=1}^{L_2}\sum_{j=1}^{H^l} u_{cj}^l h_j^l + d_c. \tag{10}$$

SV Derivation of Data-Independent Expectations

The data-independent expectations can be obtained by differentiation energy function $E(\mathbf{v},\mathbf{h},\mathbf{y})$, reported in (13) of the main paper, with respect to the model parameter set $\boldsymbol{\theta}$, where the state variables, sampled from the model distribution, are denoted with superscript tilde (e.g., $\tilde{v}$):

$$\frac{\partial E}{\partial w_{ij}^{1m}} = \tilde{v}_i^m \tilde{h}_j^{1m}; \quad \frac{\partial E}{\partial w_{jk}^{lm}} = \tilde{h}_j^{(l-1)m}\tilde{h}_k^{lm}, \text{ for } 1 < l \le L_1; \quad \frac{\partial E}{\partial w_{jk}^{(L_1+1)m}} = \tilde{h}_j^{L_1 m}\tilde{h}_k^1; \quad \frac{\partial E}{\partial w_{jk}^l} = \tilde{h}_j^l \tilde{h}_k^{(l+1)}, \text{ for } 1 \le l < L_2;$$

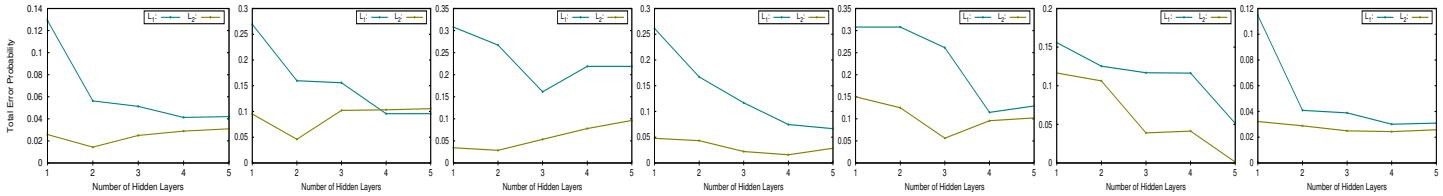
$$\frac{\partial E}{\partial u_{cj}^{lm}} = \tilde{y}_c \tilde{h}_j^{lm}, \text{for } 1 \le l \le L_1; \quad \frac{\partial E}{\partial u_{cj}^l} = \tilde{y}_c \tilde{h}_j^l, \text{for } 1 \le l \le L_2; \quad \frac{\partial E}{\partial a_i^m} = \tilde{v}_i^m; \quad \frac{\partial E}{\partial b_j^{lm}} = \tilde{h}_j^{lm}, \text{for } 1 \le l \le L_1;$$

$$\frac{\partial E}{\partial d_c} = \tilde{y}_c; \quad \frac{\partial E}{\partial b_j^l} = \tilde{h}_j^l, \text{for } 1 \le l \le L_2; \quad \frac{\partial E}{\partial \lambda_m} = \left(1 - \sum_{j=1}^{H^{L_1 m}} (\tilde{h}_j^{L_1 m})^2\right); \text{ and}$$

$$\frac{\partial E}{\partial \gamma_{mr}} = \left\{\sum_{j=1}^{H^{L_1 m}} (\tilde{h}_j^{L_1 m} - \bar{\tilde{h}}_j^{L_1 m})(\tilde{h}_j^{L_1 r} - \bar{\tilde{h}}_j^{L_1 r})\right\}^2 \ \forall m, r. \tag{11}$$

SVI Error Plots

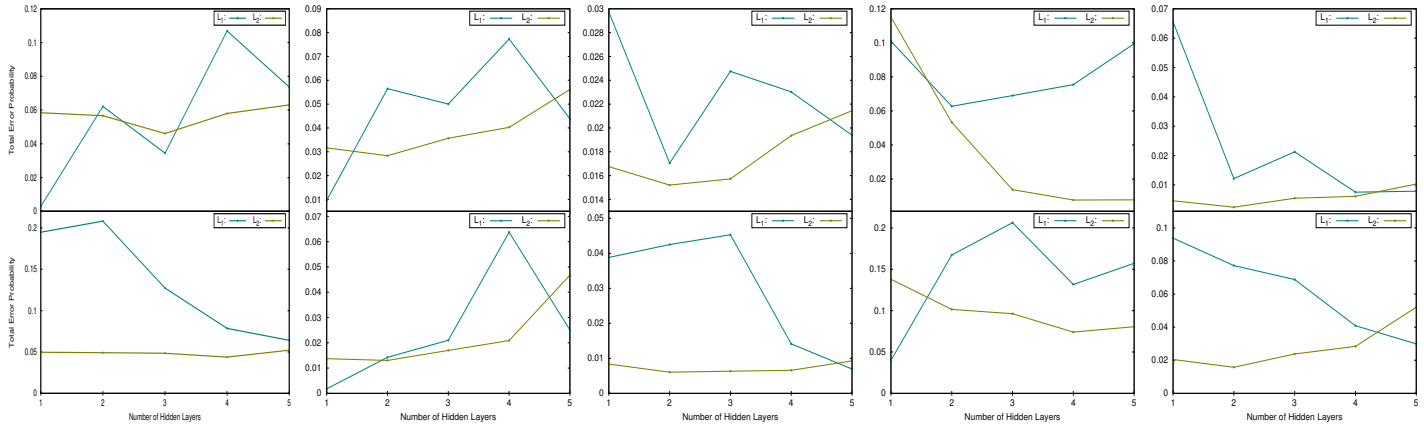
In existing literature, the architecture of a deep framework is heuristically determined for all the databases under consideration. Hence, it does not take into account the diversities present in the nature of the problem as well as the complexities associated with the given data sets. However, in the proposed method, an upper bound on the error probability is estimated in terms of the architecture of the model, which enables the framework to select an optimal architecture for the analysis of the given multi-view data. In the current study, greedy layer-wise pretraining [18] is performed to initialize the parameters of the architecture sensibly by learning a stack of modified restricted Boltzmann machines. During the pretraining, different architectures are considered for each of the data sets and the corresponding values of error bound are computed. The architecture for which the error bound attains minimum value is employed for the analysis of the particular database. Hence, a data specific deep framework is obtained from the pretraining stage only.



(a) Digits (b) Caltech (c) CiteSeer (d) Cora (e) NW-OBJECT (f) Reuters (g) AwA

Figure SI: Variation of error bound with respect to the architecture of the proposed model on benchmark data.

In order to determine the optimal number of layers in the proposed D2GDA model, extensive experiments are carried out on seven benchmark databases, namely, Digits [10], Caltech [14], CiteSeer [17], Cora [17], NW-OBJECT [3], Reuters [1], and Animals with Attributes (AwA) [25], and five cancer data sets, which include cervical carcinoma (CESC), colorectal carcinoma (CRC), kidney carcinoma (KIDNEY), lower grade glioma (LGG), and lung carcinoma (LUNG). The corresponding results are plotted in Fig. SI for benchmark databases and in Fig. SII for each of the omics data sets. At first, the number



(a) CESC (b) CRC (c) KIDNEY (d) LGG (e) LUNG

Figure SII: Variation of error bound with respect to the architecture of the proposed framework (top row: training-testing and bottom row: 10-fold CV on omics data).

Table SI: Optimal Number of Layers for Proposed D2GDA Model Based on Estimated Error Bound

Number of Layers	Training-Testing												10-fold CV				
	Digits	Caltech	CiteSeer	Cora	NW-OBJECT	Reuters	AwA	CESC	CRC	KIDNEY	LGG	LUNG	CESC	CRC	KIDNEY	LGG	LUNG
L_1, L_2	4, 2	4, 2	3, 2	5, 4	4, 3	5, 5	4, 4	1, 3	1, 2	2, 2	2, 4	4, 2	5, 4	1, 2	5, 2	1, 4	5, 2

of modality specific hidden layers (L_1) is varied from 1 to 5, keeping the number of joint hidden layers (L_2) fixed at 1 and the corresponding values of error bound are noted. Then, L_1 is fixed at the value for which the error bound has achieved the minimum value, while L_2 is varied from 1 to 5 and the variation in the error bound is observed. The value of L_2 for which error bound attains the minimum value is considered for analysis of the particular data set. The optimal values of L_1 and L_2, obtained from Fig. SI and Fig. SII, are tabulated in Table SI which are considered to develop the proposed deep architecture for each of the given input databases. It also establishes the fact that different number of layers is required by the proposed deep architecture to address the challenges offered by each of the data sets for various experimental set-up.

SVII Effectiveness of Proposed Model

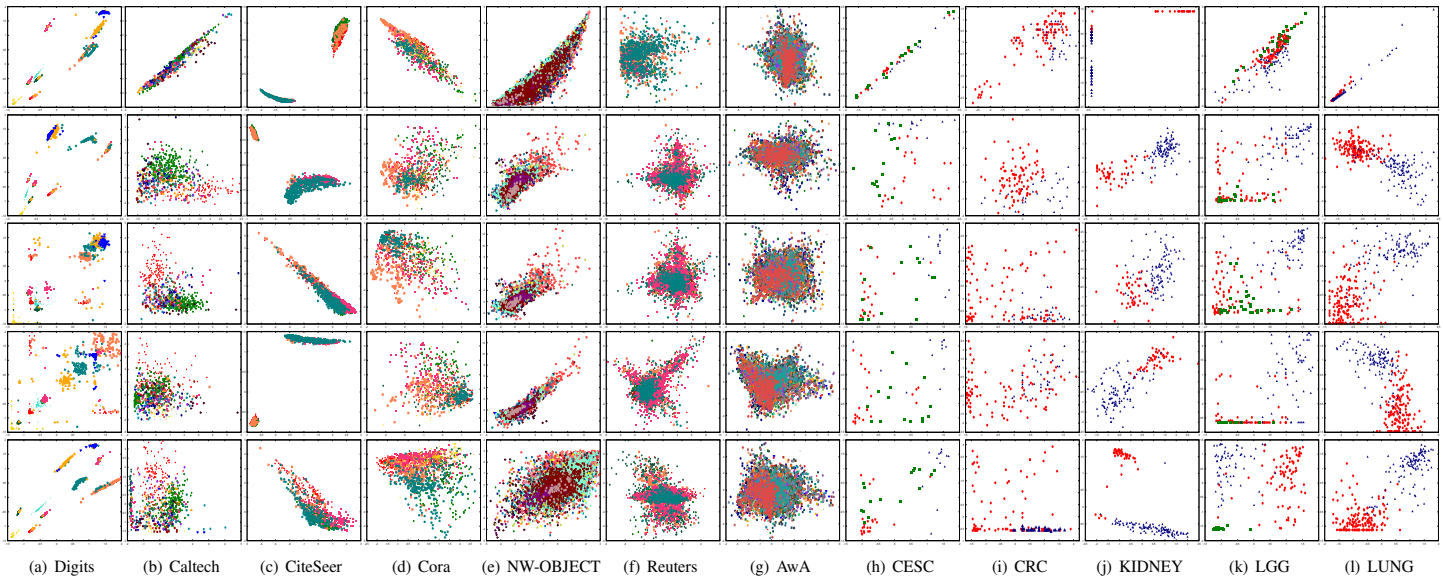
Various state-of-the-art feature extraction methods, namely CCA [8], GMPCA [19] and PLS [24], can be expressed as special cases of the proposed loss function employed in the D2GDA model. So, for different values of $\gamma_{mr}, \lambda_m,$ and $\overline{\mathbf{h}}^{L_1 m}$, the energy function of the proposed D2GDA framework reduces to the objective function of MDDBM, MDDBM_CCA, MDDBM_GMPCA, and MDDBM_PLS models. In order to establish the effectiveness of the proposed model, the performance of the D2GDA framework is extensively studied in comparison to its different variants on seven benchmark databases and five omics data sets, considering both training-testing and 10-fold CV. The corresponding results for Digits, Cora, NW-OBJECT, Reuters, AwA, CESC, LGG, and LUNG data sets are reported in Tables III and IV of main paper, and Caltech, CiteSeer, CRC, and KIDNEY databases are presented in Tables SII and SIII. The scatter plots for all the benchmark and omics data sets are presented in Fig. SIII. It is to be mentioned here that the architecture remains the same for the proposed model as well as the variants corresponding to each of the data sets. Only the learning objectives are changed based on the values of $\gamma_{mr}, \lambda_m,$ and $\overline{\mathbf{h}}^{L_1 m}$, in (15), (16), and (17) of the main paper, respectively. The scatter plots of Fig. SIII are depicted by considering the most relevant feature at x-axis and the corresponding most significant feature at y-axis, obtained using the concept of rough hypercuboid approach [15]. The first row of Fig. SIII corresponds to MDDBM framework, the second, third, and fourth rows refer to MDDBM_CCA, MDDBM_GMPCA, and MDDBM_PLS models, respectively, while the plots of last row are obtained from the proposed D2GDA framework.

Table SII: Performance Analysis of Different Variants of Proposed Model on Benchmark Data

Data	MDDBM	MDDBM _CCA	MDDBM _GMPCA	MDDBM _PLS	D2GDA
Caltech	74.14	83.78	82.51	84.28	**92.52**
CiteSeer	71.93	73.02	**73.48**	71.30	73.12

Table SIII: Comparative Performance Analysis of Different Variants of Proposed Model on Omics Data Sets

Data	Different Metrics		MDDBM	MDDBM_CCA	MDDBM_GMPCA	MDDBM_PLS	D2GDA
CRC	Train-Test		78.46	80.00	77.69	79.23	**85.50**
	10-fold CV	Mean	71.11	83.33	84.07	82.59	**87.04**
		Median	74.07	81.48	85.19	81.48	**88.89**
		StdDev	15.20	5.59	6.31	5.25	**4.00**
		Paired-t:p	6.26E-03	1.15E-02	7.63E-02	6.50E-03	-
		Wilcoxon:p	2.34E-03	7.97E-03	8.59E-02	1.36E-02	-
KIDNEY	Train-Test		98.03	98.03	96.05	94.74	**98.68**
	10-fold CV	Mean	90.32	99.35	99.35	98.10	**99.68**
		Median	91.94	**100.00**	**100.00**	98.21	**100.00**
		StdDev	8.60	1.36	1.36	**0.42**	1.02
		Paired-t:p	3.84E-03	2.96E-01	1.72E-01	1.53E-03	-
		Wilcoxon:p	3.68E-03	2.82E-01	1.59E-01	6.06E-03	-



(a) Digits (b) Caltech (c) CiteSeer (d) Cora (e) NW-OBJECT (f) Reuters (g) AwA (h) CESC (i) CRC (j) KIDNEY (k) LGG (l) LUNG

Figure SIII: Scatter plots of different variants of proposed architecture on benchmark and omics data sets (1st row: MDDBM; 2nd row: MDDBM_CCA; 3rd row: MDDBM_GMPCA; 4th row: MDDBM_PLS; and 5th row: D2GDA).

From the results reported in Tables SII and SIII, it can be observed that expect for MDDBM_GMPCA model on CiteSeer database, the D2GDA model performs better than the different variants of the model on benchmark as well as omics databases for both training-testing and 10-fold cross-validation (CV). Considering the scatter plots of Fig. SIII corresponding to the benchmark databases, it can be observed that all the classes of Digits data set can be properly identified by the different variants of the proposed model. For Caltech and Cora databases, although the samples from different classes tend to overlap for MDDBM architecture, the separation between the classes has improved for rest of the models. However, in case of CiteSeer data set, almost similar plots are obtained for all the models. The scatter plots of Fig. SIII corresponding to the omics data sets demonstrate that all the given classes can be efficiently identified by the proposed architecture. While the classes are well separated for all the variants on KIDNEY and LUNG data, the separation between the samples of different categories has detoriated in case of rest of the data sets.

SVIII Comparative Performance Analysis

In this section, the performance of the proposed D2GDA model is studied with reference to several existing view-pair specific approaches and multi-view joint representation based approaches on both benchmark as well as omics data sets. The corresponding results are reported in Tables SIV and SVI, and Tables SV and SVII on benchmark and omics data, respectively, and the scatter plots of existing approaches as well as proposed model are presented in Fig. SIV and Fig. SV for all the databases considered. The existing view-pair specific approaches include RGCCA [22], MCCA [13], GMCCA [2], GMKCCA [2], LasCCA [6], DisCCA [6], MCCP [29], TCCA-OS [21], DCCA-VG [23], and MDL-CW [16], and multi-view joint

representation based approaches include MvDA [11], MvDA-VC [12], MvCCDA [28], MDBM [20], mgRBM [30], MvLDAN [9], DACCA [5], TDDCCA [4], TCCA [27], MMGNN [7], and MVGAN [26]. The scatter plots of Fig. SIV and Fig. SV are depicted by considering the most relevant feature at x-axis and the corresponding most significant feature at y-axis, obtained using the concept of rough hypercuboid approach [15]. The plots from top to bottom row of Fig. SIV correspond to non-deep approaches, which include RGCCA, MCCA, GMCCA, GMKCCA, LasCCA, DisCCA, MvDA, MvDA-VC, MvCCDA, and D2GDA, respectively, whereas the plots of Fig. SV from top to bottom refer to MDBM, mgRBM, MvLDAN, DACCA, DCCA-VG, TDDCCA, TCCA, MDL-CW, MMGNN, MVGAN, and D2GDA framework, respectively.

Table SIV: Comparative Performance Analysis of View-Pair Specific Approaches on Benchmark Data Sets

Data	RGCCA (2011)	MCCA (1971)	GMCCA (2019)	GMKCCA (2019)	LasCCA (2016)	DisCCA (2016)	MCCP (2022)	TCCA-OS (2023)	DCCA-VG (2019)	MDL-CW (2016)	D2GDA
Digits	90.30	87.00	11.20	6.60	10.20	5.60	89.90	**97.40**	89.00	86.40	97.20
Caltech	33.71	41.83	4.82	7.48	4.06	3.17	83.02	82.51	49.68	54.25	**92.52**
CiteSeer	27.61	58.13	23.43	24.98	22.25	20.71	64.67	60.31	37.33	42.05	**73.12**
Cora	52.16	32.85	30.97	30.19	31.63	30.19	68.92	62.04	44.62	41.40	**82.46**
NW-OBJECT	18.91	30.34	4.56	6.43	7.40	10.93	51.07	49.35	19.23	18.20	**55.39**
Reuters	55.26	57.50	24.78	28.69	28.67	23.27	79.19	74.03	64.38	62.69	**86.70**
AwA	6.90	15.08	1.58	3.09	1.59	1.94	57.61	53.48	46.59	59.58	**68.57**

Table SV: Comparative Performance Analysis of View-Pair Specific Approaches on Omics Data Sets

Data	Different Metrics		RGCCA (2011)	MCCA (1971)	GMCCA (2019)	GMKCCA (2019)	LasCCA (2016)	DisCCA (2016)	MCCP (2022)	TCCA-OS (2023)	DCCA-VG (2019)	MDL-CW (2016)	D2GDA
CESC	Train-Test		61.54	38.46	42.31	44.23	42.31	36.54	53.85	63.46	65.38	65.38	**78.85**
	10-fold CV	Mean	75.00	45.83	49.17	38.33	35.00	39.17	65.83	63.33	67.50	69.17	**84.17**
		Median	79.17	50.00	50.00	41.67	33.33	33.33	66.67	62.50	70.83	66.67	**83.33**
		StdDev	13.03	13.75	14.41	11.92	15.61	10.43	12.70	9.78	16.87	14.72	**6.15**
		Paired-t:p	1.21E-02	1.55E-05	6.79E-05	1.23E-07	9.46E-07	2.45E-07	3.82E-04	7.68E-05	2.94E-03	5.00E-03	-
		Wilcoxon:p	1.55E-02	2.49E-03	2.49E-03	2.47E-03	2.50E-03	2.47E-03	2.52E-03	3.82E-03	8.81E-03	1.24E-02	-
CRC	Train-Test		83.85	73.85	83.85	50.77	78.46	73.08	82.44	84.73	77.69	76.15	**85.50**
	10-fold CV	Mean	81.85	60.74	78.52	54.07	78.52	62.59	70.33	81.77	76.30	77.04	**87.04**
		Median	83.33	62.96	81.48	55.56	77.78	68.52	73.05	81.89	75.93	77.78	**88.89**
		StdDev	5.64	10.36	7.57	8.59	4.20	13.23	10.45	**1.66**	4.68	3.83	4.00
		Paired-t:p	8.24E-03	5.10E-05	2.74E-03	9.26E-07	3.10E-04	6.43E-05	1.10E-03	8.29E-04	1.87E-04	1.04E-05	-
		Wilcoxon:p	1.41E-02	2.53E-03	6.14E-03	2.52E-03	2.50E-03	2.50E-03	2.53E-03	3.44E-03	3.79E-03	2.38E-03	-
KIDNEY	Train-Test		91.45	55.92	85.53	82.89	74.34	58.55	93.42	94.08	93.42	92.76	**98.68**
	10-fold CV	Mean	92.90	60.97	75.81	81.61	79.03	60.97	92.90	97.42	96.45	95.48	**99.68**
		Median	91.94	61.29	77.42	83.87	77.42	64.52	95.16	98.39	96.77	95.16	**100.00**
		StdDev	4.76	6.17	6.32	7.91	10.99	8.93	7.72	2.96	2.82	3.12	**1.02**
		Paired-t:p	9.15E-04	7.22E-09	3.05E-07	2.92E-05	1.42E-04	1.48E-07	1.14E-02	1.24E-02	4.23E-03	1.86E-03	-
		Wilcoxon:p	5.61E-03	2.34E-03	2.45E-03	2.50E-03	2.39E-03	2.46E-03	5.61E-03	2.06E-02	1.16E-02	7.88E-03	-
LGG	Train-Test		41.40	39.78	33.33	38.71	44.09	29.03	76.88	83.87	77.96	73.12	**90.32**
	10-fold CV	Mean	45.00	35.53	40.53	33.16	38.68	38.68	80.79	72.02	51.32	76.84	**98.68**
		Median	47.37	34.21	40.79	31.58	36.84	38.16	81.58	71.78	51.32	77.63	**98.68**
		StdDev	10.99	7.67	8.70	4.99	7.95	7.24	6.80	5.71	3.34	7.63	**1.39**
		Paired-t:p	6.57E-08	1.10E-09	2.85E-09	2.97E-12	7.39E-10	7.39E-10	2.62E-06	4.75E-08	1.43E-11	2.97E-06	-
		Wilcoxon:p	2.47E-03	2.46E-03	2.50E-03	2.50E-03	2.47E-03	2.50E-03	2.52E-03	2.53E-03	2.49E-03	2.50E-03	-
LUNG	Train-Test		87.91	46.52	68.86	86.08	82.42	47.62	92.31	94.87	89.74	93.77	**96.34**
	10-fold CV	Mean	87.68	51.43	68.39	86.07	85.18	50.71	92.32	93.71	94.11	93.93	**97.32**
		Median	86.61	50.89	69.64	87.50	85.71	48.21	93.75	93.95	95.54	95.54	**98.21**
		StdDev	4.16	3.13	7.48	8.32	6.68	8.84	6.19	**1.32**	4.69	4.31	2.95
		Paired-t:p	8.00E-05	2.72E-12	2.45E-07	8.54E-04	1.01E-04	1.92E-08	1.72E-03	9.01E-03	5.99E-03	2.00E-04	-
		Wilcoxon:p	2.50E-03	2.53E-03	2.53E-03	2.53E-03	2.52E-03	2.53E-03	5.40E-03	1.09E-02	5.81E-03	3.30E-03	-

The results corresponding to Table SIV for benchmark databases demonstrate that although RGCCA and MCCP achieve considerable classification accuracy on Digits data set, the proposed model exhibits significantly better performance with respect to all the existing methods on the seven benchmark databases, except for TCCA-OS method on Digits database. From the results reported in Table SV, it can be observed that the proposed model outperforms all the existing approaches for both training-testing and 10-fold CV, considering the five omics data sets. The statistical significance test reveals that the proposed model achieves significantly better p-values for all the 100 cases.

Table SVI: Performance Analysis of Multi-View Joint Representation Based Approaches on Benchmark Data Sets

Data	MvDA (2012)	MvDA-VC (2016)	MvCCDA (2019)	MDBM (2014)	mgRBM (2022)	MvLDAN (2019)	DACCA (2020)	TCCA (2022)	TDDCCA (2016)	MMGNN (2020)	MVGAN (2018)	D2GDA
Digits	92.40	93.50	92.80	10.00	88.60	90.70	84.60	**97.80**	85.90	88.90	82.70	97.20
Caltech	76.30	75.29	76.68	2.53	78.75	79.47	73.89	87.58	74.52	74.27	77.31	**92.52**
CiteSeer	37.69	43.51	45.69	17.80	63.23	49.05	55.22	56.40	40.42	41.78	46.32	**73.12**
Cora	53.94	55.72	58.71	10.99	58.16	63.26	50.06	56.94	53.05	42.06	44.95	**82.46**
NW-OBJECT	29.03	28.62	37.33	26.07	32.16	36.27	38.42	33.61	17.80	37.87	27.61	**55.39**
Reuters	56.01	55.15	55.36	46.84	58.13	53.36	56.38	75.97	57.27	59.16	57.60	**86.70**
AwA	16.55	15.37	62.67	27.16	53.50	47.41	**69.20**	64.63	53.32	44.96	69.06	68.57

Considering the results presented in Table SVI, it can be noted that the multi-view joint representation based existing methods achieve considerable accuracy on most of the benchmark databases. However, the proposed model exhibits significantly better performance with respect to the existing approaches, except for TCCA and DACCA on Digits and AwA databases, respectively. The results corresponding to the Table SVII for real-life cancer data sets signify that the existing methods achieve considerable classification accuracy on most of the omics databases. However, the highest classification accuracy is attained by the proposed model for both training-testing and 10-fold CV, except for TCCA method on CRC data set. The p-values obtained from the two statistical significance tests indicate that out of total 110 cases, the proposed model attains significantly better p-values for 100 cases and better but not significant p-values in 9 cases.

References

[1] M. R. Amini, N. Usunier, and C. Goutte. Learning from Multiple Partially Observed Views-An Application to Multilingual Text Categorization. *Advances in Neural Information Processing Systems*, 22:28–36, 2009.

[2] J. Chen, G. Wang, and G. B. Giannakis. Graph Multiview Canonical Correlation Analysis. *IEEE Trans. on Signal Processing*, 67(11):2826–2838, 2019.

[3] T-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y-T. Zheng. NUS-WIDE: A Real-World Web Image Database from National University of Singapore. In *Proc. of the ACM Conf. on Image and Video Retrieval*, 2009.

[4] M. Dorfer and G. Widmer. Towards Deep and Discriminative Canonical Correlation Analysis. In *Proc. of the ICML Workshop on Multi-View Representaiton Learning*, pages 1–5, 2016.

[5] W. Fan, Y. Ma, H. Xu, X. Liu, J. Wang, Q. Li, and J. Tang. Deep Adversarial Canonical Correlation Analysis. In *Proc. of the International Conf. on Data Mining*, pages 352–360. SIAM, 2020.

[6] X. Fu, K. Huang, E. E. Papalexakis, H. Song, P. P. Talukdar, N. D. Sidiropoulos, C. Faloutsos, and T. Mitchell. Efficient and Distributed Algorithms for Large-Scale Generalized Canonical Correlations Analysis. In *Proc. of the IEEE Conf. on Data Mining*, pages 871–876, 2016.

[7] D. Gao, K. Li, R. Wang, S. Shan, and X. Chen. Multi-Modal Graph Neural Network for Joint Reasoning on Vision and Scene Text. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 12746–12756, 2020.

[8] H. Hotelling. Relations Between Two Sets of Variates. *Biometrika*, 28(3/4):321–377, 1936.

[9] P. Hu, D. Peng, Y. Sang, and Y. Xiang. Multi-View Linear Discriminant Analysis Network. *IEEE Trans. on Image Processing*, 28(11):5352–5365, 2019.

[10] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical Pattern Recognition: A Review. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.

[11] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen. Multi-View Discriminant Analysis. In *Proc. of the European Conf. on Computer Vision*, pages 808–821, 2012.

[12] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen. Multi-View Discriminant Analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 38(1):188–194, 2016.

[13] J. R. Kettenring. Canonical Analysis of Several Sets of Variables. *Biometrika*, 58(3):433–451, 1971.

[14] F. Li, R. Fergus, and P. Perona. One-Shot Learning of Object Categories. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.

Table SVII: Comparative Performance Analysis of Deep Architectures on Omics Data Sets

Data	Different Metrics		MvDA (2012)	MvDA-VC (2016)	MvCCDA (2019)	MDBM (2014)	mgRBM (2022)	MvLDAN (2019)	DACCA (2020)	TCCA (2022)	TDDCCA (2016)	MMGNN (2020)	MVGAN (2018)	D2GDA
CESC	Train-Test		42.31	40.38	59.62	48.08	63.46	65.38	47.12	61.54	53.98	55.77	57.69	**78.85**
	10-fold CV	Mean	46.67	50.00	61.67	52.50	63.33	65.83	39.81	60.19	78.20	69.17	61.67	**84.17**
		Median	41.67	50.00	58.33	54.17	62.50	66.67	39.42	61.54	78.20	66.67	58.33	**83.33**
		StdDev	15.32	14.16	12.55	17.59	9.78	10.72	5.37	5.74	**0.06**	13.64	12.55	6.15
		Paired-t:p	4.39E-05	1.85E-05	7.22E-04	3.15E-04	7.68E-05	1.00E-03	7.72E-08	1.38E-06	6.80E-03	9.36E-03	7.23E-04	-
		Wilcoxon:p	2.52E-03	2.47E-03	3.61E-03	3.98E-03	3.58E-03	5.76E-03	2.52E-03	2.53E-03	6.23E-03	1.30E-02	3.53E-03	-
CRC	Train-Test		80.77	83.08	82.31	26.15	76.15	80.77	85.77	**86.92**	72.50	74.62	79.23	86.15
	10-fold CV	Mean	83.70	86.67	82.59	54.07	70.37	80.00	81.69	82.92	48.80	80.00	81.85	**87.04**
		Median	85.19	**88.89**	81.48	70.37	74.07	81.48	81.54	83.85	48.76	81.48	81.48	**88.89**
		StdDev	5.00	6.34	3.92	24.33	8.37	3.98	3.12	3.55	**0.30**	4.35	3.68	4.00
		Paired-t:p	2.07E-02	4.36E-01	2.56E-03	1.25E-03	2.41E-04	3.59E-04	7.28E-03	3.39E-02	1.60E-10	9.58E-05	2.63E-04	-
		Wilcoxon:p	1.01E-01	8.07E-01	8.47E-03	2.38E-03	2.47E-03	2.45E-03	1.09E-02	2.97E-02	2.53E-03	3.56E-03	3.76E-03	-
KIDNEY	Train-Test		92.76	94.74	95.39	69.08	86.84	96.71	96.71	96.71	51.59	89.47	95.39	**98.68**
	10-fold CV	Mean	93.23	94.19	95.16	70.97	91.94	87.74	96.64	96.58	42.48	95.16	81.61	**99.68**
		Median	93.55	93.55	93.55	67.74	91.94	85.48	96.38	96.71	42.43	93.55	77.42	**100.00**
		StdDev	2.38	2.54	3.80	10.20	7.17	8.30	1.09	1.19	**0.22**	3.80	12.64	1.02
		Paired-t:p	1.43E-05	1.54E-04	1.30E-03	4.45E-06	4.82E-03	9.70E-04	2.79E-05	1.62E-04	2.20E-16	1.30E-03	6.94E-04	-
		Wilcoxon:p	2.32E-03	3.19E-03	7.03E-03	1.95E-03	8.57E-03	8.68E-03	2.49E-03	3.31E-03	2.52E-03	7.03E-03	5.71E-03	-
LGG	Train-Test		75.81	73.12	77.96	65.05	72.04	75.81	65.59	81.18	66.85	71.51	74.73	**90.32**
	10-fold CV	Mean	75.79	81.05	77.63	27.63	70.26	76.84	59.35	77.89	57.14	72.11	76.84	**98.68**
		Median	76.32	78.95	77.63	18.42	68.42	77.63	58.87	77.63	57.00	72.37	77.63	**98.68**
		StdDev	8.02	7.83	6.11	14.90	12.09	7.63	3.42	4.51	**0.39**	3.96	7.63	1.39
		Paired-t:p	3.87E-06	2.56E-05	4.74E-07	4.59E-08	1.89E-05	2.97E-06	2.00E-11	2.61E-08	2.31E-15	1.92E-09	2.97E-06	-
		Wilcoxon:p	2.50E-03	2.47E-03	2.52E-03	2.34E-03	2.52E-03	2.50E-03	2.53E-03	2.46E-03	2.53E-03	2.40E-03	2.50E-03	-
LUNG	Train-Test		92.31	91.58	93.77	87.91	87.55	90.48	95.05	93.41	67.42	90.84	92.67	**96.34**
	10-fold CV	Mean	94.82	95.54	96.96	61.25	86.96	95.18	95.71	95.42	67.72	82.14	94.11	**97.32**
		Median	96.43	95.54	**98.21**	42.86	90.18	94.64	95.60	95.24	67.65	93.75	94.64	**98.21**
		StdDev	4.16	3.29	3.37	23.81	11.67	3.77	1.17	1.00	**0.38**	19.32	3.95	2.95
		Paired-t:p	2.76E-02	7.48E-03	2.22E-01	3.24E-04	2.95E-03	9.00E-03	9.29E-02	5.71E-02	1.44E-10	1.59E-02	2.56E-03	-
		Wilcoxon:p	4.63E-02	1.21E-02	2.97E-01	2.49E-03	2.52E-03	8.24E-03	1.01E-01	5.71E-02	2.53E-03	1.03E-02	8.88E-03	-

[15] P. Maji. A Rough Hypercuboid Approach for Feature Selection in Approximation Spaces. *IEEE Trans. on Knowledge and Data Engineering*, 26(1):16–29, 2014.

[16] S. Rastegar, M. Soleymani, H. R. Rabiee, and S. M. Shojaee. Mdl-cw: A Multimodal Deep Learning Framework With Cross Weights. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2601–2609, 2016.

[17] R. Rossi and N. Ahmed. The Network Data Repository With Interactive Graph Analytics and Visualization. In *Proc. of the AAAI Conf. on Artificial Intelligence*, volume 29, 2015.

[18] R. Salakhutdinov and G. Hinton. Deep Boltzmann Machines. In *Proc. of Artificial Intelligence and Statistics*, pages 448–455, 2009.

[19] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs. Generalized Multiview Analysis: A Discriminative Latent Space. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2160–2167. IEEE, 2012.

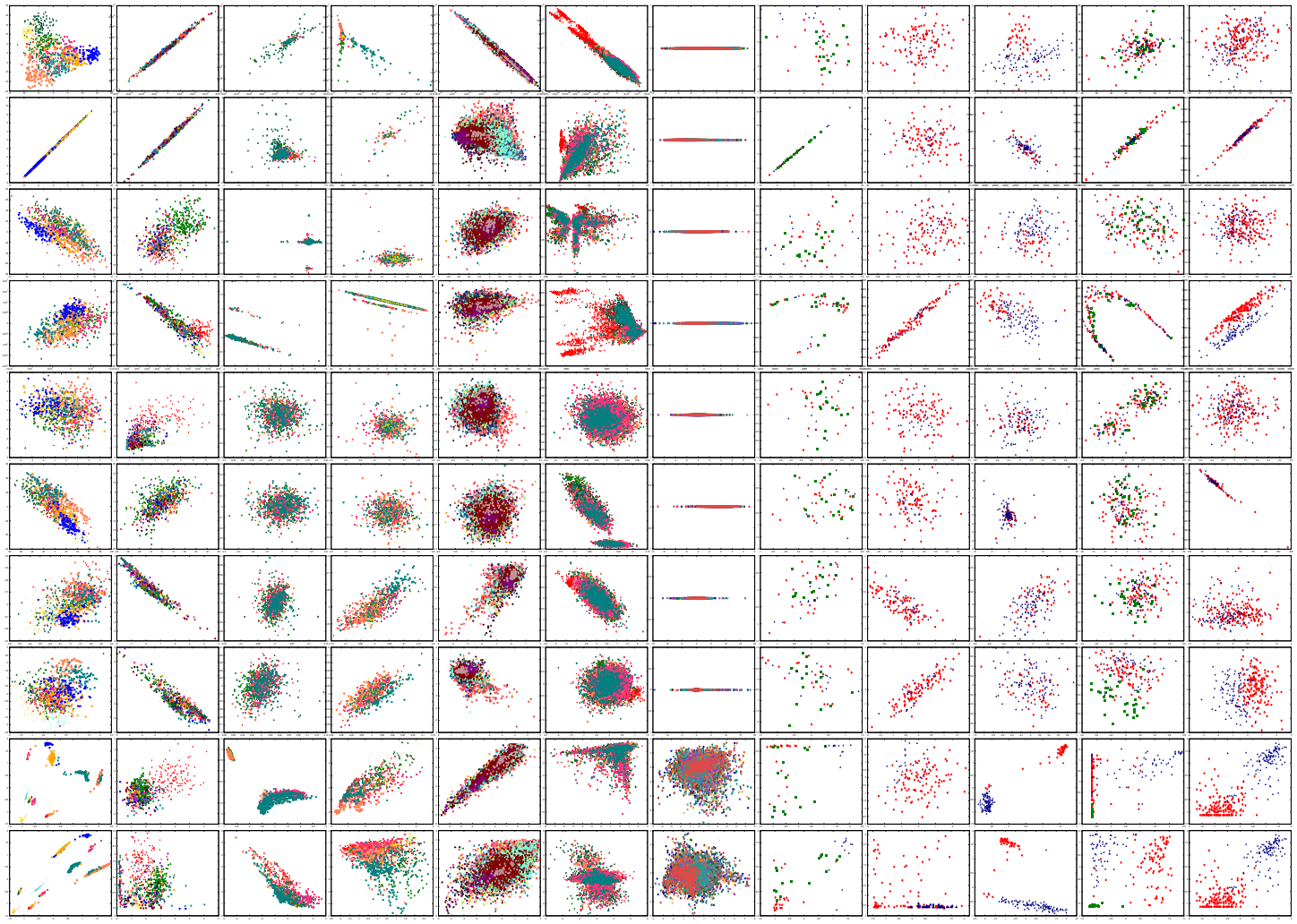
[20] N. Srivastava and R. Salakhutdinov. Multimodal Learning with Deep Boltzmann Machines. *Journal of Machine Learning Research*, 15:2949–2980, 2014.

[21] J. Sun, X. Xiu, Z. Luo, and W. Liu. Learning High-Order Multi-View Representation by New Tensor Canonical Correlation Analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[22] A. Tenenhaus and M. Tenenhaus. Regularized Generalized Canonical Correlation Analysis. *Psychometrika*, 76(2):257–284, 2011.

[23] K. G. Toker and S. E. Yüksel. Deep Canonical Correlation Analysis for Hyperspectral Image Classification. In *Proc. of the Remote Sensing of the Ocean, Sea Ice, Coastal Waters, and Large Water Regions*, page 1115009. International Society for Optics and Photonics, 2019.

[24] H. Wold. Path Models with Latent Variables: The NIPALS Approach. In H. M. Blalock, A. Aganbegian, F. M. Borodkin, R. Boudon, and V. Capecchi, editors, *Quantitative Sociology*, pages 307–357. Academic Press, 1975.



(a) Digits (b) Caltech (c) CiteSeer (d) Cora (e) NW-OBJECT (f) Reuters (g) AwA (h) CESC (i) CRC (j) KIDNEY (k) LGG (l) LUNG

Figure SIV: Scatter plots of existing non-deep approaches on benchmark and omics data sets (from top to bottom row: RGCCA, MCCA, GMCCA, GMKCCA, LasCCA, DisCCA, MvDA, MvDA-VC, MvCCDA, and D2GDA, respectively).

[25] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata. Zero-Shot Learning A Comprehensive Evaluation of the Good, the Bad and the Ugly. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 41(9):2251–2265, 2019.

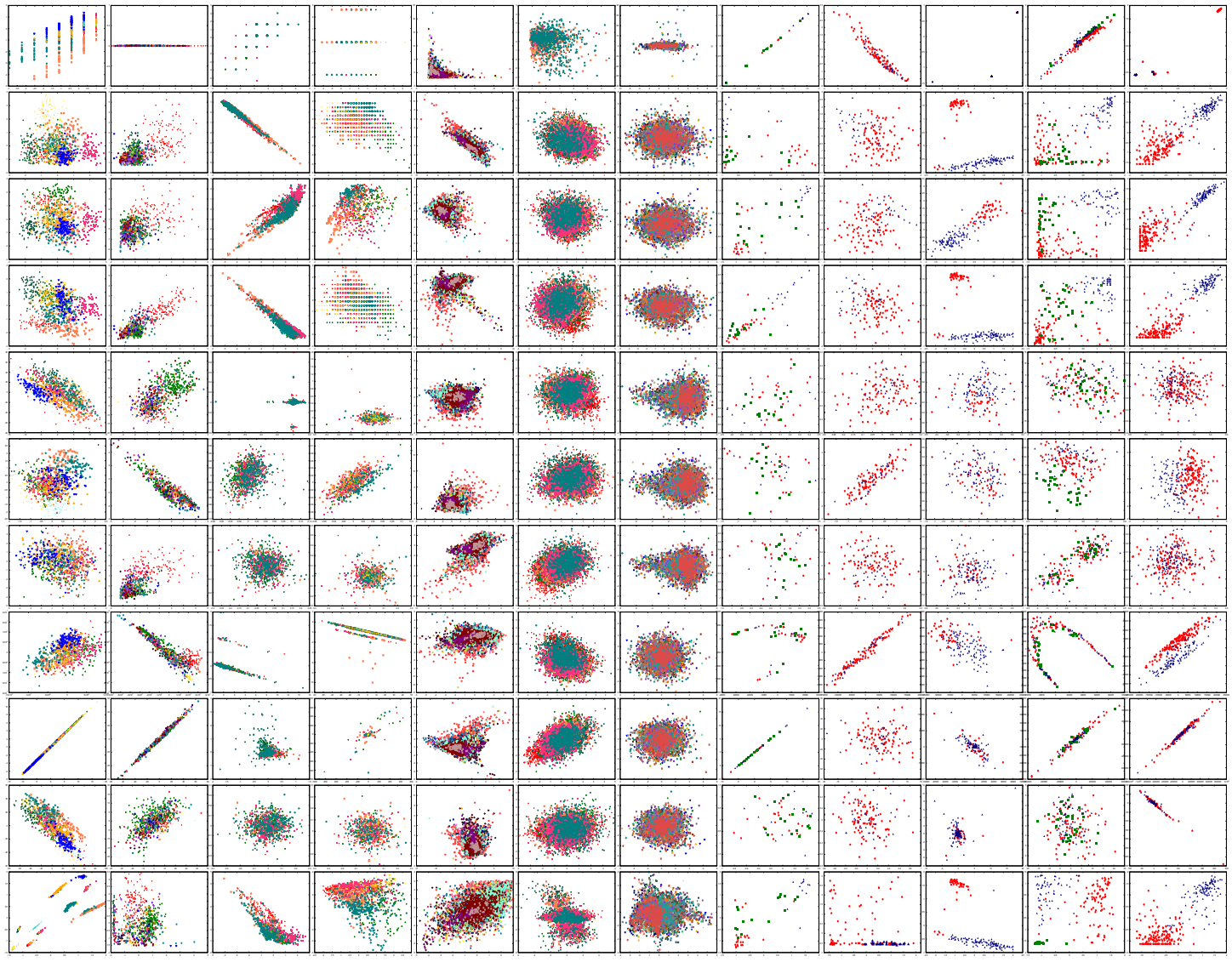
[26] Q. Xuan, Z. Chen, Y. Liu, H. Huang, G. Bao, and D. Zhang. Multiview Generative Adversarial Network and its Application in Pearl Classification. *IEEE Trans. on Industrial Electronics*, 66(10):8244–8252, 2018.

[27] X. Yang, W. Liu, and W. Liu. Tensor Canonical Correlation Analysis Networks for Multi-View Remote Sensing Scene Recognition. *IEEE Trans. on Knowledge and Data Engineering*, 34(6):2948–2961, 2022.

[28] X. You, J. Xu, W. Yuan, X.-Y. Jing, D. Tao, and T. Zhang. Multi-View Common Component Discriminant Analysis for Cross-View Classification. *Pattern Recognition*, 92:37–51, 2019.

[29] Y.-H. Yuan, J. Li, Y. Li, J. Qiang, Y. Zhu, X. Shen, and J. Gou. Learning Canonical F-Correlation Projection for Compact Multiview Representation. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 19260–19269, 2022.

[30] N. Zhang and S. Sun. Multiview Graph Restricted Boltzmann Machines. *IEEE Trans. on Cybernetics*, 52(11):12414–12428, 2022.



(a) Digits (b) Caltech (c) CiteSeer (d) Cora (e) NW-OBJECT (f) Reuters (g) AwA (h) CESC (i) CRC (j) KIDNEY (k) LGG (l) LUNG

Figure SV: Scatter plots of existing deep models on benchmark and omics data sets (from top to bottom row: MDBM, mgRBM, MvLDAN, DACCA, DCCA-VG, TDDCCA, TCCA, MDL-CW, MMGNN, MVGAN, and D2GDA, respectively).