Discriminative Deep Generalized Dependency Analysis for Multi-View Data

Debamita Kumar and Pradipta Maji

***Abstract*—In recent years, a surging interest is noted for combining the information of multiple views to obtain a joint representation of the given data. In multi-view data analysis, the joint representation should be learned from the given input views in such a way that the view-specific information as well as the cross-view dependency are preserved properly. In the context of cross-view dependency, it is expected that both view-consistency and view-discrepancy are addressed simultaneously. Discriminability of the joint representation is also an important aspect in classification problem. In this regard, a novel deep learning model is proposed to efficiently encapsulate the underlying data distribution over the space of input views. Considering both consensus and complementary principles, a loss function is introduced, based on the concept of Hilbert-Schmidt independence criterion, to capture the relevant cross-view information from the given multi-view data. Incorporating the supervised information of sample categories not only enhances the discriminative ability of the model, but also allows it to classify the given samples into different categories. An upper bound on the error probability of the proposed deep model is estimated in terms of the model architecture. It facilitates determining the optimal architecture of the proposed model for each database. The proficiency of the model is studied on numerous application domains with reference to several state-of-the-art multi-view classification algorithms.**

***Impact Statement*—This work contributes towards the development of a predictive model for the classification of multi-view data. In the proposed approach, the relationship between each pair of views is assumed to be unique. Hence, a loss function is proposed to efficiently capture the cross-view dependency across several views. It extracts the relevant cross-view information in terms of consensus and/or complementary knowledge from the input pairs of views. Instead of heuristically determining the architecture of the proposed deep model, an optimal architecture is estimated for each given database based on the Bayes error analysis of the network. While the number of layers is estimated from the total error probability of the model, the number of nodes at each layer is computed based on the Hilbert-Schmidt independence criterion. The proposed model outperforms state-of-the-art algorithms in 81.82% cases, considering five benchmark and three omics databases. In case of omics data, where the number of samples is significantly lower than that of features, the proposed model performs significantly better than the existing approaches; while the proposed model provides better results for large-scale benchmark databases in most cases.**

***Index Terms*—Boltzmann machine, deep learning, multi-view analysis, cross-view learning, dependency analysis.**

I. Introduction

THE primary objective of a predictive model is to identify and analyze the inherent structures of the data, which are relevant to categorize the given samples or observations into different groups. However, in multi-view scenario, different views may provide different representations of the underlying data distribution. So, information from various sources needs to be consolidated appropriately following either consensus or complementary principles. The input views, also referred to as modalities, are expected to agree upon the inherent latent distribution from where the views are assumed to be generated. The consensus principle focuses on maximizing the agreement on different views [1]. On the other hand, the complementary principle states that each view of the given data may provide some knowledge which is distinct from the rest of the views [2]. Therefore, the underlying coherent and complementary information of different views can be suitably exploited to enhance the proficiency of multimodal predictive models.

D. Kumar and P. Maji are with the Biomedical Imaging and Bioinformatics Lab, Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. E-mail: {debamita_r, pmaji}@isical.ac.in. Corresponding author: Pradipta Maji

A. Literature Review

1) Consensus: Following the consensus principle, various multiset canonical correlation analysis (MCCA) [1] based approaches have been developed to characterize the correlated structures across different views. In [3], graph-regularized MCCA (GMCCA) and graph-regularized kernel MCCA (GMKCCA) approaches have been developed, which utilize the graph based knowledge to embed the information of common sources into the framework. The deep CCA with view generation (DCCA-VG) [4] focuses on learning multi-view representation by fusing spatial and spectral information. The tensor CCA (TCCA) [5] network considers higher-order correlation to solve the deep optimization problem by decomposing a covariance tensor. In [6], multiset correlative covariation projection (MCCP) is developed, which employs a novel canonical $\mathcal{F}$-correlation framework for the construction of $\mathcal{F}$-intraset and $\mathcal{F}$-interset covariation matrices in order to capture the non-linear relationship between different modalities. Sun et al. [7] have introduced Tucker decomposition-based TCCA method with orthogonality and sparsity constraints (TCCA-OS) to reduce irrelevant and redundant information from the higher-order feature representation of multi-view data. However, these approaches are primarily unsupervised in nature and hence, the correlated subspaces lack discriminative information. Recently, supervised covariance-based multi-view CCA (SCMVCCA) [8] is proposed, which considers the class label information to obtain the discriminative non-linear mapping between different feature representations.

The multi-view graph restricted Boltzmann machine (mgRBM) [9] preserves the data manifold structure by performing local structural learning. The multimodal deep Boltzmann machine (MDBM) has been proposed in [10] for

learning a generative model to create a fused representation of the data, while the MDL-CW model [11] is developed based on autoencoder to maximize the mutual information between the given modalities. The multimodal graph neural network (MMGNN) [12] represents an input data as a graph and initial representations of the nodes in the graph are obtained from priors, learned from the deep convolution neural networks. In MDBM [10], the fused representation is learned without considering the shared knowledge of different views. Although the mgRBM [9] jointly learns view-consistent and view-specific graph representations, only the view-consistent representation is considered for classification purpose.

2) Complementary: Following the complementary principle, Kan et al. [13] have formulated multi-view discriminant analysis (MvDA) to jointly obtain the linear transforms for multiple views. The MvDA with view-consistency (MvDA-VC) [2] introduces a constraint to enforce the view-consistency of the linear transforms. In [14], multi-view common component discriminant analysis (MvCCDA) has been proposed to incorporate the supervised information and local geometry into the latent subspace. Hu et al. [15] have formulated multi-view linear discriminant analysis network (MvLDAN) to eliminate the discrepancy among multiple views by maximizing the between-class variations and inter-view covariances. The multi-view generative adversarial network (MVGAN) [16] expands the labelled multi-view samples to efficiently train the multistream convolutional neural network. The MvLDAN [15] and MVGAN [16] concentrate on the view-specific information, but do not take into consideration the shared knowledge of different views.

3) Both: In recent years, several deep learning models have been developed to efficiently capture both consensus and complementary information from the given multi-view data. Dorfer et al. [17] have presented towards deep and discriminative CCA (TDDCCA), which considers the consensus and complementary information by incorporating a discriminative regularizer into the existing deep CCA objective function. In [18], deep adversarial CCA (DACCA) has been proposed by incorporating the concept of adversarial learning into the theory of CCA. While TDDCCA [17] eventually disregards the underlying data distribution of the given observations, the performance of DACCA [18] is highly dependent on the input signal-to-noise ratio. Also, the effectiveness of adversarial models, such as DACCA [18], depends on the distance between a test point and the manifold of training data. Consequently, the networks are more likely to be vulnerable to the blind-spot attacks.

B. Motivation

The deep Boltzmann machine (DBM) [19] is a powerful paradigm of undirected generative models that precisely captures the non-linear dependencies between observed and latent variables by examining the energy landscape of the input observations. Hence, the latent non-linear data distribution of the given observations is expected to be encapsulated by the joint subspace learned from the DBM based multi-view model. However, the architecture of DBM is essentially unsupervised in nature. In multi-view classification problem, the joint subspace is required to embody the supervised information so that the similarity in the latent space implies the similarity in the corresponding concepts. Also, the joint representation should be learned from the given input views in such a way that the view-specific information as well as the cross-view dependency are preserved properly. In the context of cross-view dependency, it is expected that both view-consistency and view-discrepancy are addressed simultaneously. However, in existing literature, a unified approach, based on consensus and/or complementary principles, is considered to represent view-consistency or view-discrepancy in the joint subspace. In the current study, it is primarily assumed that each view corresponds to a completely different subspace and so, the relationship between each pair of views is assumed to be unique. Hence, a view-pair specific approach needs to be considered to quantify the relevant cross-view dependency among the given input views.

C. Contribution of the Current Study

In this regard, a novel deep learning model, termed as discriminative deep generalized dependency analysis (D2GDA), is introduced based on the framework of DBM in multi-view environment. In order to capture the cross-view dependency in the joint representation of the given input data, a loss function is formulated, based on the concept of Hilbert-Schmidt independence criterion (HSIC). A primary attribute of the proposed dependency analysis is that it provides a view-pair specific approach to capture the coherent structures or complementary information from the given multi-view data. The presence of class nodes in the proposed deep model enhances the discriminability of the latent subspaces, which in turn, allows the model to classify the given observations into multiple categories without employing any additional classifier. An upper bound on the error probability of the proposed model is estimated in terms of the model architecture, which enables the framework to obtain an optimal deep architecture for each experimental set-up. Analytical formulation demonstrates that the proposed model is the generalization of certain existing feature extraction techniques. The efficacy of the proposed model is demonstrated on several benchmark and real-life cancer data sets.

II. Hilbert-Schmidt Independence Criterion

A vector space with the inner-product $\langle \cdot, \cdot \rangle$ operation is referred to as inner-product space.

Property 1 *[20]: Any finite dimensional inner-product space is a Hilbert space.*

Let us consider, k represents a kernel in Hilbert space.

Theorem 1 *[20]: If k is a positive definite kernel, then there exists a unique reproducing kernel Hilbert space (RKHS) $\mathcal{F}$ whose kernel is k.*

The HSIC [21] efficiently measures the dependency between two random variables, by mapping the variables into RKHS such that the correlations measured in that space correspond to

higher-order joint moments between the original distributions. The advantage of the HSIC over state-of-the-art dependency measures is that it provides an empirical definition to compute the dependency between two random variables without estimating the joint distribution.

Consider two random variables X and Y, which are defined over two input spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively, with a joint distribution p_{xy}. Let us define a mapping $\phi(x)$ from $x \in \mathcal{X}$ to RKHS $\mathcal{F}$, such that the inner product between two vectors in $\mathcal{F}$ is given by a kernel function $k^1(x,x') = \langle \phi(x), \phi(x') \rangle$. Let $\mathcal{G}$ be another RKHS defined on input space $\mathcal{Y}$ with mapping $\varphi(y)$ and kernel function $k^2(y,y') = \langle \varphi(y), \varphi(y') \rangle$. The linear cross-covariance operator $C_{xy} : \mathcal{G} \rightarrow \mathcal{F}$ between two feature maps is defined as

$$C_{xy} = \mathbb{E}[(\phi(x) - \mu_x) \otimes (\varphi(y) - \mu_y)], \tag{1}$$

where $\mathbb{E}[\cdot]$ denotes the expectation operator, $\mu_x = \mathbb{E}[\phi(x)]$ and $\mu_y = \mathbb{E}[\varphi(y)]$ represent the mean values of $\phi(x)$ and $\varphi(y)$, respectively, and $\otimes$ signifies the tensor product. Using Riesz's representation theorem [22], it can be shown that if $\mathbb{E}[k^1(x,x')]$ and $\mathbb{E}[k^2(y,y')]$ are finite, then C_{xy} exists and is unique.

Given two separable RKHSs $\mathcal{F}$, $\mathcal{G}$, and the joint distribution p_{xy}, the HSIC between two variables X and Y is defined as the Hilbert-Schmidt norm of the cross-covariance operator:

$$\text{HSIC}(p_{xy}, \mathcal{F}, \mathcal{G}) = \|C_{xy}\|^2_{\text{HS}}. \tag{2}$$

Let $\mathcal{Z} = \{(x_1, y_1), \cdots, (x_N, y_N)\} \subseteq \mathcal{X} \times \mathcal{Y}$ be a set of N independent observations drawn from p_{xy}. The empirical estimate of the HSIC is given by

$$\text{HSIC}(\mathcal{Z}, \mathcal{F}, \mathcal{G}) = \frac{1}{(N-1)^2} \mathbf{tr}(K^1 D K^2 D), \tag{3}$$

where $\mathbf{tr}(\cdot)$ denotes the trace operator, $K^1, K^2, D \in \mathbb{R}^{N \times N}$, K^1 and K^2 are Gram matrices corresponding to the kernels k^1 and k^2, respectively, where $K^1_{i,j} = k^1(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$, $K^2_{i,j} = k^2(y_i, y_j) = \langle \varphi(y_i), \varphi(y_j) \rangle$, and $D_{i,j} = \delta_{i,j} - N^{-1}$ centers the Gram matrix to have zero mean in the feature space. In [21], it has been shown that $\text{HSIC}(\mathcal{Z}, \mathcal{F}, \mathcal{G})$ converges to $\text{HSIC}(p_{xy}, \mathcal{F}, \mathcal{G})$ at a rate of $\mathcal{O}(N^{-1/2})$ with a bias of $\mathcal{O}(N^{-1})$.

Theorem 2 *[21]: Denote by $\mathcal{F}$, $\mathcal{G}$ RKHSs with universal kernels k^1, k^2 on the compact domains $\mathcal{X}$ and $\mathcal{Y}$, respectively. Assume without loss of generality that $\|f\|_\infty \leq 1$ and $\|g\|_\infty \leq 1$ for all functional $f \in \mathcal{F}$ and $g \in \mathcal{G}$. Then, $\|C_{xy}\|_{HS} = 0$ if and only if X and Y are independent.*

In general, if the covariance between the variables X and Y is zero, then it does not imply that the variables are independent of each other. If X and Y are non-linearly related, then it will not be reflected in the corresponding covariance value. However, a zero HSIC value does imply independence of the associated variables. Hence, it can be said that the HSIC takes higher-order moments into account while measuring the dependency between two random variables.

III. Proposed Method

In this section, the proposed D2GDA model, along with its learning, is discussed in details. At first, the concept of generalized dependency analysis (GDA) is proposed in the current study to capture the cross-modal information from the given view-specific representations. Then, the architecture of the D2GDA model is described to encapsulate the underlying data distribution over the space of multimodal inputs. Finally, the learning of the D2GDA model is discussed by considering the objective of GDA in the proposed framework.

A. Proposed Generalized Dependency Analysis

Let us consider that the given input modalities $\{\mathbf{v}^m\}$ are transformed into respective modality-specific subspaces $\{\mathbf{h}^m\}$. Now, the joint subspace $\mathbf{h}$, learned from the $\{\mathbf{h}^m\}$, should be able to capture the view-specific characteristics as well as the cross-modal dependency across various modalities. The pictorial representation of the above concept is depicted in Fig. 1, where M denotes the total number of input modalities.

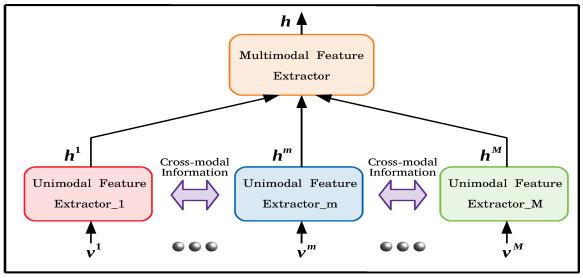

Fig. 1. Illustration of proposed multi-view data analysis framework.

Now, the relevant cross-modal information can be embedded in the correlated structures or complementary knowledge of different views. Since each view has a fundamentally distinct representation of the underlying data distribution, the relationship between each pair of views is assumed to be unique. Hence, instead of considering a unified approach among all the pairs, a view-pair specific method should be employed for efficient representation of cross-modal information in the joint subspace. In this regard, an objective function is proposed, based on the concept of HSIC, that not only quantifies dependency among multiple modalities, but also facilitates to identify relevant cross-modal information in terms of coherent structures or complementary knowledge from the given data.

In the proposed approach, it is assumed that the vector space spanned by each modality-specific representation is $\mathbb{R}^H$, where H is the dimension of $\mathbf{h}^m$, $\forall_{m=1}^M$. Hence, it is essentially a Hilbert space of dimension H since it holds Property 1.

Let, K^m be the Gram matrix corresponding to the kernel k^m associated with the Hilbert space spanned by $\mathbf{h}^m$. In the proposed method, K^m is defined as

$$K^m = (\mathbf{h}^m - \underline{\mathbf{h}}^m) \otimes (\mathbf{h}^m - \underline{\mathbf{h}}^m), \quad \forall m \in \{1, 2, \cdots, M\}, \tag{4}$$

where $\underline{\mathbf{h}}^m$ represents the mean vector of $\mathbf{h}^m$. So, K^m is defined to be a cross-covariance matrix.

Property 2 *: K^m is always positive semi-definite.*

The positive definiteness of K^m can be ensured by restricting the variance value equal to 1 using Lagrange multiplier. So, by Theorem 1, it can be said that the vector space spanned by the corresponding modality-specific representation is RKHS. Hence, based on the K^m considered in the current study, the value of HSIC between the representations $\mathbf{h}^m$ and $\mathbf{h}^r$ is computed as

$$\begin{aligned}\text{HSIC}(\mathbf{h}^m,\mathbf{h}^r) &= \frac{1}{(N-1)^2}\Big\{\mathbf{tr}(K^m K^r)\Big\}\\ &= \frac{1}{(N-1)^2}\Bigg\{\mathbf{tr}\Bigg(\Big[\big(\sum_{n=1}^{N}(h^m_{nj}-\underline{h}^m_j)(h^m_{nk}-\underline{h}^m_k)\big)_{j,k}\Big]_{H\times H}\\ &\qquad \Big[\big(\sum_{n=1}^{N}(h^r_{nj}-\underline{h}^r_j)(h^r_{nk}-\underline{h}^r_k)\big)_{j,k}\Big]_{H\times H}\Bigg)\Bigg\}\\ &= \frac{1}{(N-1)^2}\sum_{n=1}^{N}\left[\left\{\sum_{j=1}^{H}(h^m_{nj}-\underline{h}^m_j)(h^r_{nj}-\underline{h}^r_j)\right\}^2\right]. \quad (5)\end{aligned}$$

Thus, from the definition of K^m, presented in (4), the expression of $\text{HSIC}(\mathbf{h}^m,\mathbf{h}^r)$ is obtained in (5), which is to be considered for the rest of the current study. Here, the centering matrices of (3) are ignored since K^m in (4) is already defined to be mean centered. Based on the above analysis, the following theorem is introduced for $\text{HSIC}(\mathbf{h}^m,\mathbf{h}^r),\ \forall m,r$, corresponding to the Gram matrix K^m defined in (4).

Theorem 3 *: The value of HSIC($\boldsymbol{h}^m,\boldsymbol{h}^r$) between $\boldsymbol{h}^m$ and $\boldsymbol{h}^r$ lies within the range of [0,1], if $H \le \frac{4(N-1)}{\sqrt{N}}$ and $h^m_{nj} \in$ [0,1], $\forall n,j,m$.*

Proof: It is assumed that $h^m_{nj}\in[0,1],\ \ \forall n,j,m.$

$$\text{So, } (h^m_{nj}-\underline{h}^m_j)\in[-\frac{1}{2},\frac{1}{2}],\ \ \forall n,j,m$$

$$\Rightarrow (h^m_{nj}-\underline{h}^m_j)(h^r_{nj}-\underline{h}^r_j)\in[-\frac{1}{4},\frac{1}{4}],\ \forall n,j,m,r$$

$$\Rightarrow \left\{\sum_{j=1}^{H}(h^m_{nj}-\underline{h}^m_j)(h^r_{nj}-\underline{h}^r_j)\right\}^2\in[0,\frac{|H^2|}{16}],\ \forall n,m,r$$

$$\Rightarrow \text{HSIC}(\mathbf{h}^m,\mathbf{h}^r)\in[0,\frac{N|H^2|}{16(N-1)^2}],\ \forall m,r.$$

$$\text{So, if } H\le\frac{4(N-1)}{\sqrt{N}}, \text{HSIC}(\mathbf{h}^m,\mathbf{h}^r)\in[0,1],\forall m,r.\square$$

Thus, Theorem 3 provides an upper bound on the dimensionality of modality-specific representations $\mathbf{h}^m,\ \forall_{m=1}^M$ based on the number of given observations N. It is important to note here that the bound obtained in Theorem 3 becomes particularly effective in cases where the dimension or the number of features in a representation is heuristically determined. In the proposed approach, the definition of HSIC, obtained from (5), is employed to quantify the cross-modal information between each pair of modality-specific representations. As discussed in Section II, the higher value of HSIC indicates higher dependency between two random variables, while the zero value of HSIC implies independence between the associated variables. In order to capture the non-linear dependency among multiple modalities, a new measure, termed as balanced HSIC (BHSIC), is introduced by incorporating a balance parameter between each pair of views as follows:

$$\begin{aligned}&\text{BHSIC}(\mathbf{h}^1,\cdots,\mathbf{h}^m,\cdots,\mathbf{h}^M)\\ &= \frac{2}{M(M-1)}\sum_{m=1}^{M-1}\sum_{r=(m+1)}^{M}|\gamma_{mr}|\text{HSIC}(\mathbf{h}^m,\mathbf{h}^r). \quad (6)\end{aligned}$$

Here, γ_{mr} denotes the balance parameter between a pair of views $\mathbf{v}^m$ and $\mathbf{v}^r$. The value of γ_{mr} in (6) signifies the contribution of dependency between $\mathbf{h}^m$ and $\mathbf{h}^r$ in the overall cross-modal information of the given input data.

Property 3 *: Given HSIC($\boldsymbol{h}^m,\boldsymbol{h}^r$)$\in$ [0,1] and $\gamma_{mr}\in$ [-1,1], $\forall_{m,r=1}^M$, the value of BHSIC($\boldsymbol{h}^1,\cdots,\boldsymbol{h}^m,\cdots,\boldsymbol{h}^M$) lies within the range of [0,1]. A zero value of BHSIC denotes that all the modality-specific representations are completely independent of each other, while a higher value of BHSIC signifies higher dependency between the given representations.*

Based on the BHSIC measure, defined in (6), a loss function is proposed to learn view-consistency and view-discrepancy simultaneously across several modalities, which is as follows:

$$\begin{aligned}E_B(\mathbf{h}^1,\cdots,\mathbf{h}^m,\cdots,\mathbf{h}^M) &= -\sum_{m=1}^{M}\lambda_m\Big(1-\mathbf{tr}(K^m)\Big)\\ &\quad -\Big\{\sum_{m=1}^{M-1}\sum_{r=(m+1)}^{M}\gamma_{mr}\mathbf{tr}(K^m K^r)\Big\}\\ &= -\sum_{n=1}^{N}\Bigg[\sum_{m=1}^{M}\lambda_m\Big(1-\sum_{j=1}^{H}(h^m_{nj})^2\Big)\\ &\quad +\sum_{m=1}^{M-1}\sum_{r=(m+1)}^{M}\gamma_{mr}\left\{\sum_{j=1}^{H}(h^m_{nj}-\underline{h}^m_j)(h^r_{nj}-\underline{h}^r_j)\right\}^2\Bigg], \quad (7)\end{aligned}$$

where $\gamma_{mr}\in[-1,1]$ and λ_m represents the Lagrange multiplier. The first term in (7) ensures that the variance value of $\mathbf{h}^m$ is equal to 1, or equivalently, $K^m,\ \forall m$, is always positive definite. The second term computes the weighted dependency value between each pair of input modalities. Thus, for $\gamma_{mr}\in(0,1]$, minimizing E_B will ensure that the BHSIC value and correspondingly, the dependency between the modality-specific representations are maximized. As a consequence, the coherent knowledge between the views will be reflected in the joint subspace. However, if $\gamma_{mr}\in[-1,0)$, minimization of E_B corresponds to the minimization of the BHSIC value. Hence, the independence between the associated pair of views will be maximized, which in turn, enhances the complementary information of the individual views in the joint representation.

Property 4 *: If $\gamma_{mr}\in(0,1]$, then the dependency between $\boldsymbol{h}^m$ and $\boldsymbol{h}^r$ is maximized; if $\gamma_{mr}\in[-1,0)$, then the independence between $\boldsymbol{h}^m$ and $\boldsymbol{h}^r$ is maximized; and if $\gamma_{mr}=0$, then the dependency between the corresponding view-pair is not taken under consideration in order to minimize E_B.*

The loss function, defined in (7), and Property 4 ensure that a view-pair specific method can be developed, which may appropriately capture cross-modal information across multiple modalities in terms of correlated or complementary structures of the data, based on the appropriate values of balance parameters. In order to learn the optimal value of γ_{mr}, a deep model, based on DBM, is proposed next.

B. Architecture of Proposed Model

In multi-view classification problem, it is expected that the non-linear structures embedded in the given input views, along with the supervised information of sample categories, are suitably reflected in the joint subspace. In this regard, the D2GDA model is developed based on the framework of multimodal discriminative deep Boltzmann machine (MDDBM) [23]. It incorporates the class nodes into the proposed architecture which allows the corresponding latent subspaces to have better discriminative ability and also, enables the model to classify the given observations into multiple categories.

Let the input view corresponding to the m-th modality be represented by $\mathbf{v}^m = \{v_1^m, \cdots, v_i^m, \cdots\}$ and $\mathbf{y} = \{y_1, \cdots, y_c, \cdots\}$ denotes the supervised information. Let us consider that $L_1 > 0$ signifies the number of modality-specific hidden layers and $L_2 > 0$ refers to the number of joint hidden layers. The modality-specific hidden representation, corresponding to the l-th layer of the m-th modality, is denoted as $\mathbf{h}^{lm} = \{h_1^{lm}, \cdots, h_j^{lm}, \cdots\}$, whereas $\mathbf{h}^l = \{h_1^l, \cdots, h_j^l, \cdots\}$ indicates the l-th joint hidden representation. Here, the number of nodes in a representation is designated by the corresponding capital letter. For example, the number of nodes in $\mathbf{v}^m$ is denoted by V^m.

The bidirectional weight parameter w_{ij}^{1m} connects the i-th visible node to the j-th hidden node of first modality-specific hidden layer from the m-th modality. The j-th hidden node of l-th modality-specific hidden layer is connected to the k-th hidden node of $(l+1)$-th modality-specific hidden layer from m-th modality through $w_{jk}^{(l+1)m}$, the j-th hidden node of L_1-th modality-specific hidden layer from m-th modality is connected to the k-th hidden node of first joint hidden layer through $w_{jk}^{(L_1+1)m}$, and the j-th hidden node of l-th joint hidden layer is connected to the k-th hidden node of $(l+1)$-th joint hidden layer through $w_{jk}^{(l+1)}$. Similarly, the c-th class node is connected to the j-th hidden node of the l-th modality-specific hidden layer from m-th modality through u_{cj}^{lm} and the c-th class node is connected to j-th hidden node of l-th joint hidden layer through u_{cj}^{l}. The bias parameters a_i^m, b_j^{lm}, b_j^l, and d_c are associated with the i-th visible node of the m-th modality, j-th hidden node of l-th modality-specific hidden layer from m-th modality, j-th hidden node of l-th joint hidden layer, and c-th class node, respectively.

The energy function E_a and parameter space $\boldsymbol{\theta}_a$ of the MDDBM model are presented in (9) and (10), respectively.

$$\boldsymbol{\theta}_a = \{w^{1m}, \cdots, w^{(L_1+1)m}, w^1, \cdots, w^{(L_2-1)}, u^{1m}, \cdots, u^{L_1 m}, u^1, \cdots, u^{L_2}, a^m, b^{1m}, \cdots, b^{L_1 m}, b^1, \cdots, b^{L_2}, d\},\ \forall m. \quad (10)$$

Thus, the supervised information of sample categories can be appropriately incorporated at each layer of the architecture through proper learning of the set of parameters $\{u^{1m}, \cdots, u^{L_1 m}, u^1, \cdots, u^{L_2}, d\},\ \forall_{m=1}^M$, which in turn, enhances the proficiency of the proposed framework.

C. Generalized Dependency Analysis for Learning Proposed D2GDA Architecture

The objective function of GDA is judiciously integrated with the learning objective of MDDBM architecture to develop the proposed D2GDA model. The proposed model is not only able to learn the intrinsic characteristics associated with each of the given input modalities, but also identifies the relevant cross-modal information across different views.

1) Objective Function of Proposed Model: The proper learning of the MDDBM framework ensures that the joint subspace suitably represents the underlying inherent characteristics of the given modalities as well as supervised information of the sample categories. In order to encapsulate the cross-view dependency across several modalities in the shared subspace, the loss function, proposed in (7), is considered in the D2GDA model. The principle of GDA is integrated with the objective of the MDDBM model since the following properties hold.

- In MDDBM, $h_j^{L_1 m},\ \forall j, m$, represents the state of the j-th hidden node of modality-specific hidden layer L_1 from modality m, which is essentially a real value. Hence, $\mathbf{h}^{L_1 m}$ spans $\mathbb{R}^H$, where H denotes the dimension of $\mathbf{h}^{L_1 m},\ \forall m$, that is, $H^{L_1 1} = H^{L_1 2} = \cdots H^{L_1 M} = H$. So, Property 1 holds.
- The Gram matrix K^m is a variance-covariance matrix corresponding to $\mathbf{h}^{L_1 m}$ of the MDDBM architecture. In effect, it satisfies Property 2.
- In MDDBM framework, $h_j^{L_1 m} \in \{0, 1\},\ \forall j, m$. Considering $H \leq \frac{4(N-1)}{\sqrt{N}}$ and $\gamma_{mr} \in [-1, 1]$, it can be ensured that both HSIC($\mathbf{h}^{L_1 m}, \mathbf{h}^{L_1 r}$) and BHSIC($\mathbf{h}^{L_1 1}, \cdots, \mathbf{h}^{L_1 M}$), $\forall_{m,r=1}^M$, lie within the range of [0,1]. Thus, Theorem 3 and Property 3 are satisfied.

So, the loss function of GDA, presented in (7), corresponding to each given observation, can be efficiently combined with the energy function (9) of the MDDBM architecture. Hence, the overall objective of the D2GDA model turns out be

$$E(\mathbf{v}, \mathbf{h}, \mathbf{y}) = E_a(\mathbf{v}, \mathbf{h}, \mathbf{y}) + E_b(\mathbf{h}^{L_1 1}, \cdots, \mathbf{h}^{L_1 M}), \quad (11)$$

where E_b represents the loss function E_B of (7) for each given observation. The parameter space of the new model is defined by $\boldsymbol{\theta} = \boldsymbol{\theta}_a \cup \boldsymbol{\theta}_b$; where $\boldsymbol{\theta}_b = \{\lambda_m, \gamma_{mr}\},\ \forall_{m,r=1}^M$. The M number of λ_m and $\binom{M}{2}$ number of γ_{mr} parameters, along with the other parameters of D2GDA model, have to be learned. If γ_{mr} is learned to be positive, then the energy function in (11) decreases with increase in the dependency between $\mathbf{h}^{L_1 m}$ and $\mathbf{h}^{L_1 r}$. However, if γ_{mr} is learned to be negative, then $E(\mathbf{v}, \mathbf{h}, \mathbf{y})$ in (11) decreases as the corresponding modality-specific representations become more independent of each other.

The learning of D2GDA corresponds to estimating the model parameter set $\boldsymbol{\theta}$ that maximizes the probability of observing the given input data. So, the objective function of D2GDA is given by the corresponding log-likelihood function $\ln \mathrm{L}(\boldsymbol{\theta}|\mathbf{v}, \mathbf{y})$ and the partition function is defined as $Z = \sum_{\mathbf{v},\mathbf{h},\mathbf{y}} e^{-E(\mathbf{v},\mathbf{h},\mathbf{y})}$. Since the parameter space of the D2GDA model is quite large, the gradient ascent on the log-likelihood is commonly used to determine the optimal parameters of the model, which turns out to be the difference between the expectation of gradient of energy function under model distribution, referred to as data-independent expectation, and under the conditional distribution of hidden representation given the input views, termed as data-dependent expectation.

$$E_a(\mathbf{v},\mathbf{h},\mathbf{y}) = -\sum_{m=1}^{M}\sum_{i=1}^{V^m}\sum_{j=1}^{H^{1m}} v_i^m w_{ij}^{1m} h_j^{1m} - \sum_{l=1}^{L_1}\sum_{m=1}^{M}\sum_{c=1}^{Y}\sum_{j=1}^{H^{lm}} y_c u_{cj}^{lm} h_j^{lm} - \sum_{l=1}^{L_1-1}\sum_{m=1}^{M}\sum_{j=1}^{H^{lm}}\sum_{k=1}^{H^{(l+1)m}} h_j^{lm} w_{jk}^{(l+1)m} h_k^{(l+1)m}$$
$$- \sum_{m=1}^{M}\sum_{j=1}^{H^{L_1m}}\sum_{k=1}^{H^1} h_j^{L_1m} w_{jk}^{(L_1+1)m} h_k^1 - \sum_{l=1}^{L_2}\sum_{c=1}^{Y}\sum_{j=1}^{H^l} y_c u_{cj}^l h_j^l - \sum_{l=1}^{L_2-1}\sum_{j=1}^{H^l}\sum_{k=1}^{H^{(l+1)}} h_j^l w_{jk}^l h_k^{(l+1)} - \sum_{m=1}^{M}\sum_{i=1}^{V^m} a_i^m v_i^m$$
$$- \sum_{l=1}^{L_1}\sum_{m=1}^{M}\sum_{j=1}^{H^{lm}} b_j^{lm} h_j^{lm} - \sum_{l=1}^{L_2}\sum_{j=1}^{H^l} b_j^l h_j^l - \sum_{c=1}^{Y} d_c y_c. \tag{9}$$

2) Estimation of Data-Dependent Expectations: Now, the exact maximum likelihood learning is intractable, so the variational learning [24] is employed to estimate the data-dependent expectation. In variational inference, the posterior distribution $P(\mathbf{h}|\mathbf{v},\mathbf{y})$ is approximated with a tractable mean field distribution $Q(\mathbf{h}|\mathbf{v},\mathbf{y}) \approx P(\mathbf{h}|\mathbf{v},\mathbf{y})$. Now,

$$\ln P(\mathbf{v},\mathbf{y}) \geq \sum_{\mathbf{h}} Q(\mathbf{h}|\mathbf{v},\mathbf{y}) \ln \frac{P(\mathbf{v},\mathbf{h},\mathbf{y})}{Q(\mathbf{h}|\mathbf{v},\mathbf{y})} = \mathcal{L}_v, \tag{12}$$

where $P(\mathbf{v},\mathbf{h},\mathbf{y}) = (1/Z)e^{-E(\mathbf{v},\mathbf{h},\mathbf{y})}$ represents the probability associated with the joint configuration $\{\mathbf{v},\mathbf{h},\mathbf{y}\}$. Thus, the mean field approximation provides a lower bound $\mathcal{L}_v$ on the log-likelihood function. So, better approximation of $P(\mathbf{h}|\mathbf{v},\mathbf{y})$ implies tighter bound on $\ln P(\mathbf{v},\mathbf{y})$.

Here, let the mean field distribution be defined as

$$Q(\mathbf{h}|\mathbf{v},\mathbf{y}) = \prod_{l=1}^{L_1}\prod_{m=1}^{M}\prod_{j=1}^{H^{lm}} q(h_j^{lm}|\mathbf{v},\mathbf{y}) \prod_{l=1}^{L_2}\prod_{j=1}^{H^l} q(h_j^l|\mathbf{v},\mathbf{y}), \tag{13}$$

where the hidden units $\{h_j\}$ are considered to be Bernoulli variables with $q(h_j|\mathbf{v},\mathbf{y}) = \mu_j^{\{h_j=1\}}(1-\mu_j)^{\{h_j=0\}}$ and μ_j denotes the probability of being the state of h_j as 1. The definitions of $Q(\mathbf{h}|\mathbf{v},\mathbf{y})$, presented in (13), and $P(\mathbf{v},\mathbf{h},\mathbf{y})$, corresponding to the energy function obtained in (11), are substituted in (12) to obtain the final expression of $\mathcal{L}_v$, which is reported in (14). The detailed derivation of (14) is presented in Section SI of the supplementary material.

Since, the mean field parameters ($\boldsymbol{\mu}$) of $\mathcal{L}_v$, presented in (14), define the equilibrium state of the model, they need to be updated accordingly. In order to obtain the mean field parameters of the proposed model, $\mathcal{L}_v$ of (14) is maximized with respect to $\boldsymbol{\mu}$ for a fixed $\boldsymbol{\theta}$. The update rule for the hidden nodes of each layer of the proposed model is derived in details in Section SII of supplementary material. A representative update rule for the hidden nodes of layer L_1, corresponding to the modality m, is given by

$$\mu_j^{L_1m} = \sigma\Bigg(\sum_{k=1}^{H^{(L_1-1)m}} \mu_k^{(L_1-1)m} w_{kj}^{L_1m} + \sum_{k=1}^{H^1} w_{jk}^{(L_1+1)m} \mu_k^1$$
$$+ \sum_{r\neq m=1}^{M} \gamma_{mr}(1-2\underline{h}_j^{L_1m})(\mu_j^{L_1r} - 2\mu_j^{L_1r}\underline{h}_j^{L_1r} + \underline{h}_j^{L_1r})$$
$$+2\sum_{r\neq m=1}^{M} \gamma_{mr} \sum_{k\neq j=1}^{H^{L_1m}} (\mu_j^{L_1r} - \underline{h}_j^{L_1r})(\mu_k^{L_1m} - \underline{h}_k^{L_1m})$$
$$(\mu_k^{L_1r} - \underline{h}_k^{L_1r}) - \lambda_m + \sum_{c=1}^{Y} y_c u_{cj}^{L_1m} + b_j^{L_1m} \Bigg), \ \forall j,m, \tag{15}$$

where $\sigma(x) = \dfrac{1}{1+e^{-x}}$ is the sigmoid function. Given the equilibrium state of the model, the parameter set $\boldsymbol{\theta}$ of the proposed architecture, corresponding to the data-dependent expectation, can be learned by maximizing $\mathcal{L}_v$ with respect to $\boldsymbol{\theta}$ for the equilibrium mean field parameters $\boldsymbol{\mu}$. The expression for differentiation of $\mathcal{L}_v$ with respect to each of the parameters is presented in Section SIII of the supplementary material.

3) Estimation of Data-Independent Expectations: The Markov Chain Monte Carlo based stochastic approximation procedure [25] is considered to approximate the data-independent expectations. The idea behind this approach is to sample a new state of the model from the current state, based on the conditional distributions over visible and hidden nodes for a fixed parameter set $\boldsymbol{\theta}$. The detailed derivation of the conditional distributions, corresponding to the proposed model, is presented in Section SIV of supplementary material. Few representative conditional distributions are given by

$$P(h_j^{L_1m}|\mathbf{h}^{(L_1-1)m}, \mathbf{h}^1, \mathbf{h}_{-j}^{L_1m}, \mathbf{h}^{L_1r}, \mathbf{y})$$
$$= \sigma\Bigg(\sum_{k=1}^{H^{(L_1-1)m}} h_k^{(L_1-1)m} w_{kj}^{L_1m} + \sum_{k=1}^{H^1} w_{jk}^{(L_1+1)m} h_k^1 - \lambda_m$$
$$+ \sum_{r\neq m=1}^{M} \gamma_{mr}(1-2\underline{h}_j^{L_1m}) \Big(h_j^{L_1r} - \underline{h}_j^{L_1r} \Big)^2$$
$$+2\sum_{r\neq m=1}^{M} \sum_{k\neq j=1}^{H^{L_1m}} \gamma_{mr}(h_j^{L_1r} - \underline{h}_j^{L_1r})(h_k^{L_1m} - \underline{h}_k^{L_1m})$$
$$(h_k^{L_1r} - \underline{h}_k^{L_1r}) + \sum_{c=1}^{Y} y_c u_{cj}^{L_1m} + b_j^{L_1m} \Bigg), \ \forall j,m, \tag{16}$$

$$P(y_c|\mathbf{h}^{11},\cdots,\mathbf{h}^{L_1M},\mathbf{h}^1,\cdots,\mathbf{h}^{L_2}) = \frac{e^{X_c}}{\sum_{\bar{c}=1}^{Y} e^{X_{\bar{c}}}},$$

$$X_c = \sum_{l=1}^{L_c}\sum_{m=1}^{M}\sum_{j=1}^{H^{lm}} u_{cj}^{lm} h_j^{lm} + \sum_{l=1}^{L_2}\sum_{j=1}^{H^l} u_{cj}^l h_j^l + d_c, \tag{17}$$

where $\mathbf{h}_{-j}^{L_1m}$ represents modality-specific hidden representation, corresponding to L_1-th layer of the m-th modality, consisting values of all the nodes except $h_j^{L_1m}$.

Given that the convergence criteria, discussed in Section III-C5, are satisfied, if a Markov chain is run for sufficient number of steps, then it can be ensured that the chain will converge to a unique stationary distribution such that the subsequent states of the chain will be accordingly distributed. The gradient of the energy function under model distribution is estimated by drawing samples from the obtained stationary distribution. So, many persistent chains are run in parallel and states of the chains are sampled based on the conditional

$$\mathcal{L}_v = \sum_{\mathbf{h}} Q(\mathbf{h}|\mathbf{v},\mathbf{y})\{\ln P(\mathbf{v},\mathbf{h},\mathbf{y}) - \ln Q(\mathbf{h}|\mathbf{v},\mathbf{y})\} = \sum_{\mathbf{h}} Q(\mathbf{h}|\mathbf{v},\mathbf{y})\{-E(\mathbf{v},\mathbf{h},\mathbf{y}) - \ln Z\} - \sum_{\mathbf{h}} Q(\mathbf{h}|\mathbf{v},\mathbf{y}) \ln Q(\mathbf{h}|\mathbf{v},\mathbf{y})$$
$$= E_a(\mathbf{v},\boldsymbol{\mu},\mathbf{y}) + \sum_{m=1}^{M} \lambda_m \left(1 - \sum_{j=1}^{H^{L_1 m}} \mu_j^{L_1 m}\right) + \sum_{m=1}^{M-1} \sum_{r=(m+1)}^{M} \sum_{j=1}^{H^{L_1 m}} \gamma_{mr}(\mu_j^{L_1 m} - 2\mu_j^{L_1 m}\underline{h}_j^{L_1 m} + \underline{h}_j^{L_1 m})(\mu_j^{L_1 r} - 2\mu_j^{L_1 r}\underline{h}_j^{L_1 r}$$
$$+\underline{h}_j^{L_1 r}) + 2\sum_{m=1}^{M-1} \sum_{r=(m+1)}^{M} \sum_{j=1}^{H^{L_1 m}-1} \sum_{k=(j+1)}^{H^{L_1 m}} \gamma_{mr}(\mu_j^{L_1 m} - \underline{h}_j^{L_1 m})(\mu_j^{L_1 r} - \underline{h}_j^{L_1 r})(\mu_k^{L_1 m} - \underline{h}_k^{L_1 m})(\mu_k^{L_1 r} - \underline{h}_k^{L_1 r}) - \ln Z -$$
$$\sum_{l=1}^{L_1} \sum_{m=1}^{M} \sum_{j=1}^{H^{lm}} \left\{\mu_j^{lm} \ln \mu_j^{lm} + (1-\mu_j^{lm}) \ln(1-\mu_j^{lm})\right\} - \sum_{l=1}^{L_2} \sum_{j=1}^{H^l} \left\{\mu_j^{l} \ln \mu_j^{l} + (1-\mu_j^{l}) \ln(1-\mu_j^{l})\right\}. \quad (14)$$

distributions. The expression for data-independent expectation is presented in Section SV of the supplementary material.

4) Learning Rule of Proposed Model Parameters: Let us assume that t, η, N, S, ρ, and ζ represent the current epoch, learning rate, number of training observations, number of persistent Markov chains, weight decay, and momentum constant, respectively. Thus, the learning rule for the parameters of the proposed D2GDA model, required to perform gradient ascent on the log-likelihood function corresponding to the energy function of (11), can be defined as:

$$\boldsymbol{\theta}^{(t+1)} = F\big(\boldsymbol{\theta}^t + \Delta\boldsymbol{\theta}^t\big), \quad (18)$$

$$\text{where} \quad \Delta\boldsymbol{\theta}^t = \eta\left\{\frac{1}{N}\sum_{n=1}^{N}\left(\frac{\partial \mathcal{L}_v}{\partial \boldsymbol{\theta}}\right)_n - \frac{1}{S}\sum_{s=1}^{S}\left(\frac{\partial E}{\partial \boldsymbol{\theta}}\right)_s\right\} - \rho\boldsymbol{\theta}^t + \zeta\Delta\boldsymbol{\theta}^{(t-1)}, \quad (19)$$

and $F(\cdot)$ denotes hyperbolic tangent function in case of balance parameter γ_{mr}, and identity function for the rest of the parameters belonging to $\boldsymbol{\theta}$. Thus, it can be noted that in the proposed D2GDA model, the values of γ_{mr} are learned in such a way that $\gamma_{mr} \in [-1, 1]$, and hence, Property 4 is satisfied. It can also be observed that all the parameters of the model are learned simultaneously using (18), and the modification in the parameter values of $\boldsymbol{\theta}_a$ in (10) at any epoch depends only on the pre-synaptic and post-synaptic nodes of the parameter, which is in accordance with the Hebbian learning rule. Also, subtracting the data-independent expectations from the corresponding data-dependent terms in (18) basically stabilizes the distribution of parameters as well as allows the proposed model to propagate uncertainties associated with ambiguous inputs.

5) Convergence Analysis: In the D2GDA model, variational learning is employed to estimate the data-dependent expectations. It provides a lower bound $\mathcal{L}_v : \mathbb{R}^{|\boldsymbol{\theta}|} \mapsto \mathbb{R}$ on the log-likelihood function of the proposed model. From the definition of $\mathcal{L}_v$, presented in (14), corresponding to the D2GDA framework, it can be noted that $\mathcal{L}_v$ is a differential function having β-Lipschitz continuous gradient for some $\beta \geq 0$, that is, $\|\nabla\mathcal{L}_v(\theta_1) - \nabla\mathcal{L}_v(\theta_2)\|_2 \leq \beta\|\theta_1 - \theta_2\|_2$. It can also be observed that the gradient function $\nabla\mathcal{L}_v(\theta)$ is independent of θ, that is, $\nabla\mathcal{L}_v(\theta_1) = \nabla\mathcal{L}_v(\theta_2)$, $\forall\theta_1, \theta_2 \in \boldsymbol{\theta}$. Hence, following similar analysis presented in [23], the convergence of the proposed D2GDA model can be established.

IV. Different Aspects of Proposed Model

In this section, different aspects of the proposed D2GDA model, which include error analysis and generalization ability of the model, are discussed.

A. Error Analysis of Proposed Model

The proposed D2GDA model is developed to classify the observations of the given multi-view data into different categories. Now, the Bayes discriminant function provides the optimal solution to any classification problem. So, the mean-squared error between the prediction rule (17) of the D2GDA model and Bayes decision rule is studied, in order to analyze the discriminative ability of the proposed framework. Thus, reduction in the corresponding error will indicate better approximation of Bayes discriminant function by the proposed model, which in turn, will ensure better discriminative ability of the model.

Let, $\mathbf{v}$ be the input visible vector, ω_c represents the c-th input class, where C represents the total number of classes, that is, $C = Y$, Ω_c denotes the set of all possible visible vectors, and $\Omega = \bigcup_{c=1}^{C} \Omega_c$ signifies the input space. In the proposed D2GDA model, the class label for the input vector $\mathbf{v}$ is predicted as $\arg\max_c P(y_c = 1|\mathbf{v})$, where $P(y_c = 1|\mathbf{v})$ is obtained using (17). Now, the Bayes optimal discriminant functions are given by $g_c(\mathbf{v}) = P(\omega_c|\mathbf{v})$, $\forall c \in \{1, 2, \cdots, C\}$, and the corresponding decision rule is $\mathbf{v} \in \omega_c$ if $g_c(\mathbf{v}) \geq g_k(\mathbf{v})$, $\forall k \neq c$. This decision rule is termed as minimum error decision rule as it provides the minimum probability of error.

In order to establish that the prediction criterion (17) of the proposed architecture approximates the Bayes decision rule, it is to be demonstrated that the following error criterion is minimized through learning of the architecture:

$$\epsilon^2(\boldsymbol{\theta}) = \sum_{c=1}^{C} \int_{\Omega} \Big[\ln P(y_c = 1|\mathbf{v}) - g_c(\mathbf{v})\Big]^2 p(\mathbf{v}) d\mathbf{v}. \quad (20)$$

In the proposed framework, learning of the model (**??**) refers to estimating the parameter values for which the probability of observing the given samples is maximized, that is,

$$\max_{\boldsymbol{\theta}}\left\{\frac{1}{N}\sum_{\mathbf{v},\mathbf{y}} \ln P(\mathbf{v},\mathbf{y})\right\} = \max_{\boldsymbol{\theta}}\left\{\frac{N_1}{N}\frac{1}{N_1}\sum_{\mathbf{v}\in\Omega_1} \ln P(\mathbf{v}, y_1 = 1) + \cdots + \frac{N_C}{N}\frac{1}{N_C}\sum_{\mathbf{v}\in\Omega_C} \ln P(\mathbf{v}, y_C = 1)\right\}, \quad (21)$$

where N_c denotes the number of training observations corresponding to the class ω_c. Now, the number of feature vectors drawn from $p(\mathbf{v})$ for any given class is proportional to the a priori probability of that class. Considering a class with non-zero probability of occurrence, $N \to \infty$ will imply $N_c \to \infty$. So, by strong law of large numbers, (21) can be written as

$$\min_{\boldsymbol{\theta}} \left\{ \epsilon^2(\boldsymbol{\theta}) - \sum_{c=1}^{C} \int_{\Omega} \left\{ \ln P(y_c = 1|\mathbf{v}) \right\}^2 p(\mathbf{v}) d\mathbf{v} \right\} - \sum_{c=1}^{C} \int_{\Omega} \left\{ g_c(\mathbf{v}) \right\}^2 p(\mathbf{v}) d\mathbf{v} + \int_{\Omega} p(\mathbf{v}) \ln p(\mathbf{v}) d\mathbf{v}. \quad (22)$$

Hence, it can be observed that learning of the proposed architecture with prediction criterion (17) attempts to provide a classifier which is mean-squared error approximation to the Bayes optimal classifier. So, minimization of the mean-squared error depends on the efficient learning of the model parameters, which in turn, depends on the model architecture. Thus, by modifying the architecture of the model or parameter values, the discriminative ability of the model can be varied. Hence, there must exist a relation between the model architecture and the values of the parameters with the error probability of the proposed D2GDA model.

Because of the resemblance of prediction criterion between the proposed method and Bayes decision rule, the error probability of the D2GDA model can be defined in accordance with the Bayes multi-class classifier [26], which is given by

$$P_e = \mathbb{E}[P(e|\mathbf{v})] = \mathbb{E}[1 - \max_c P(y_c = 1|\mathbf{v})] \leq 2 \sum_{\mathbf{v} \in \Omega} p(\mathbf{v}) \sum_{c=1}^{C-1} \sum_{k=(c+1)}^{C} P(y_c = 1|\mathbf{v}) P(y_k = 1|\mathbf{v}). \quad (23)$$

In the current study, $P(y_c = 1|\mathbf{v})$ is defined as conditional distribution in (17), which can be replaced in (23) and the corresponding upper bound on the error probability of the D2GDA model can be obtained as

$$P_e \leq \sum_{\mathbf{v} \in \Omega} p(\mathbf{v}) \left\{ 2 \sum_{c=1}^{C-1} \sum_{k=(c+1)}^{C} \frac{e^{X_c}}{\sum_{\bar{c}=1}^{C} e^{X_{\bar{c}}}} \frac{e^{X_k}}{\sum_{\bar{k}=1}^{C} e^{X_{\bar{k}}}} \right\}$$

$$\Rightarrow P_e \leq \sum_{\mathbf{v} \in \Omega} p(\mathbf{v}) \left\{ 1 - \frac{\sum_{c=1}^{C} \left(e^{X_c} \right)^2}{\left(\sum_{\bar{c}=1}^{C} e^{X_{\bar{c}}} \right)^2} \right\}, \quad (24)$$

where $X_c = \sum_{l=1}^{L_1} \sum_{m=1}^{M} \sum_{j=1}^{H^{lm}} u_{cj}^{lm} h_j^{lm} + \sum_{l=1}^{L_2} \sum_{j=1}^{H^l} u_{cj}^{l} h_j^{l} + d_c$. (25)

So, an upper bound of the error probability P_e is achieved in terms of X_c, which depends on both architecture as well as parameters of the model. Through proper learning of the model parameters, the upper bound on the error probability can be minimized. Also, by suitably varying the model architecture, a tighter bound on P_e can be achieved.

Let, $u_c^1 = \min_{j,l,m} \{u_{cj}^{lm}\}$, $h^1 = \min_{j,l,m} \{h_j^{lm}\}$, $u_c^2 = \min_{j,l} \{u_{cj}^{l}\}$, $h^2 = \min_{j,l} \{h_j^{l}\}$, $H^1 = \min_{l,m} \{H^{lm}\}$, and $H^2 = \min_{l} \{H^{l}\}$. So,

$$X_c \leq L_1 M H^1 u_c^1 h^1 + L_2 H^2 u_c^2 h^2. \quad (26)$$

If the value of X_c in (25) is substituted with the formulation of (26), the inequality will still hold, which is given by

$$P_e \leq \sum_{\mathbf{v} \in \Omega} p(\mathbf{v}) \left\{ 1 - \frac{\sum_{c=1}^{C} e^{2L_1 M H^1 u_c^1 h^1 + 2L_2 H^2 u_c^2 h^2}}{\left(\sum_{c=1}^{C} e^{L_1 M H^1 u_c^1 h^1 + L_2 H^2 u_c^2 h^2} \right)^2} \right\}. \quad (27)$$

It can be observed from (27) that instead of heuristically determining the architecture of the proposed framework, an optimal deep architecture can be obtained for the analysis of the given multi-view data. Apart from the model parameters and architecture of the proposed D2GDA framework, the error probability also depends on the nature and complexity of the given classification problem.

B. Generalization Ability of Proposed Model

In this section, CCA [27], generalized multi-view principal component analysis (GMPCA) [28], and partial least squares (PLS) [29], are shown as special cases of proposed E_b.

1) CCA: In CCA [1], the main objective is to maximize the correlation between each pair of given input modalities. Let, $\gamma_{mr} = 1$ and $\underline{\mathbf{h}}^{L_1 m} = 0$, $\forall_{m,r=1}^{M}$. So, the energy function $E_b(\mathbf{h})$, presented in (11), reduces to

$$E_c(\mathbf{h}) = -\sum_{m=1}^{M-1} \sum_{r=(m+1)}^{M} \left\{ \sum_{j=1}^{H^{L_1 m}} h_j^{L_1 m} h_j^{L_1 r} \right\}^2 - \sum_{m=1}^{M} \lambda_m \left(1 - \sum_{j=1}^{H^{L_1 m}} (h_j^{L_1 m})^2 \right). \quad (28)$$

Here, the first term $\sum_{j=1}^{H^{L_1 m}} h_j^{L_1 m} h_j^{L_1 r}$ corresponds to the trace of covariance between $\mathbf{h}^{L_1 m}$ and $\mathbf{h}^{L_1 r}$, and the second term represents the constraint that the variance of $\mathbf{h}^{L_1 m}$ is equal to 1. So, in order to minimize the energy function $E_c(\mathbf{h})$ in (28), $\mathbf{tr}(\mathrm{cov}(\mathbf{h}^{L_1 m}, \mathbf{h}^{L_1 r}))$ is to be maximized subject to the constraint that $\mathrm{var}(\mathbf{h}^{L_1 m}) = 1$. Hence, the energy function $E_c(\mathbf{h})$ in (28) is essentially the Lagrangian of the CCA. Now, if $E_c(\mathbf{h})$, obtained in (28), is considered in $E(\mathbf{v}, \mathbf{h}, \mathbf{y})$ of (11), then the joint representation will be learned from maximally correlated modality-specific subspaces, and hence, the corresponding model is referred to as MDDBM_CCA.

2) GMPCA: The GMPCA [28] aims to determine the direction where the variance of each given modality as well as the covariance between every pair of modalities are maximized. Suppose, $\gamma_{mr} = 1$, $\lambda_m = -1$, and $\underline{\mathbf{h}}^{L_1 m} = 0$, $\forall m, r$. In this case, the energy function $E_b(\mathbf{h})$ of (11) is reduced to

$$E_g(\mathbf{h}) = -\sum_{m=1}^{M-1} \sum_{r=(m+1)}^{M} \left\{ \sum_{j=1}^{H^{L_1 m}} h_j^{L_1 m} h_j^{L_1 r} \right\}^2 - \sum_{m=1}^{M} \sum_{j=1}^{H^{L_1 m}} (h_j^{L_1 m})^2. \quad (29)$$

From (29), it can be observed that the first term represents the squared value of trace of covariance between every pair of modality-specific hidden representations, while the second term denotes the variance of $\mathbf{h}^{L_1 m}$, $\forall m$. So, in order to detect

the optimal minima in the given energy landscape, both the variance and covariance terms need to be maximized, which is primarily the objective of the GMPCA. The deep framework obtained by considering $E_g(\mathbf{h})$ of (29) in the energy function of (11) is termed as MDDBM_GMPCA.

3) PLS: The objective of PLS [29] is to maximize the covariance between each pair of input modalities. Assume, $\gamma_{mr}=1,\ \lambda_m=0$, and $\underline{\mathbf{h}}^{L_1 m}=0,\ \forall m,r$. In such a scenario, the energy function $E_b(\mathbf{h})$ of (11) turns out to be

$$E_p(\mathbf{h}) = -\sum_{m=1}^{M-1}\sum_{r=(m+1)}^{M}\left\{\sum_{j=1}^{H^{L_1 m}} h_j^{L_1 m} h_j^{L_1 r}\right\}^2. \tag{30}$$

It is evident from (30) that $E_p(\mathbf{h})$ is the negative of the squared value of $\mathbf{tr}(\text{cov}(\mathbf{h}^{L_1 m}, \mathbf{h}^{L_1 r}))$, $\forall m,r$. Hence, in order to minimize $E_p(\mathbf{h})$ in (30), the covariance between each pair of modality-specific hidden representations is required to be maximized, which is clearly the objective of PLS. Now, if the energy function $E_p(\mathbf{h})$ of (30) is employed in the overall energy function $E(\mathbf{v},\mathbf{h},\mathbf{y})$ of (11), then the joint subspace of the MDDBM architecture will be formed in such a way that the covariance between the modality-specific representations in the projected space is maximum. The deep framework, obtained using the energy function of (30), is termed as MDDBM_PLS.

Thus, the proposed loss function is the generalization of the three acknowledged feature extraction techniques, namely, CCA, GMPCA, and PLS. In this context, it is to be mentioned here that, if $\gamma_{mr}=0,\ \lambda_m=0$, and $\underline{\mathbf{h}}^{L_1 m}=0,\ \forall m,r$, then the D2GDA model boils down to the MDDBM model. The parameter values for the MDDBM, MDDBM_CCA, MDDBM_GMPCA, and MDDBM_PLS models can be efficiently learned by suitably replacing the values of γ_{mr} and $\underline{\mathbf{h}}^{L_1 m}$ in (14), (15), and (16).

V. Experimental Results and Discussions

The proficiency of the proposed D2GDA model is analyzed extensively and the corresponding results are presented in this section. The performance of several state-of-the-art approaches is studied to validate the efficacy of the proposed architecture for training-testing as well as 10-fold cross-validation (CV). For training-testing, overall classification accuracy is employed, whereas in case of 10-fold CV, mean, median, standard deviation, and p-values evaluated using paired-*t* (one-tailed) and Wilcoxon signed-rank (one-tailed) tests, with 95% confidence level, are considered. The executable code of the proposed model is available at https://www.isical.ac.in/~bibl/results/d2gda/d2gda.html. Additional results, including scatter plots of the proposed D2GDA model as well as the existing approaches, are reported in the supplementary document.

A. Description of Data Sets

In order to evaluate the performance of different algorithms, five benchmark databases, namely, Digits [30], Cora [31], NUS-WIDE-OBJECT (NW-OBJECT) [32], Reuters [33], and Animals with Attributes (AwA) [34], and three cancer data sets are considered. The Digits, NW-OBJECT, and AwA are image based databases, Cora consists of scientific publications with annotated labels, whereas Reuters is a multilingual categorization data set. Three real-life omics data sets, corresponding to cervical carcinoma (CESC), lower grade glioma (LGG), and lung carcinoma (LUNG), are obtained from The Cancer Genome Atlas [35]. The number of samples, number of classes, number of views, and number of features in each view of the databases are reported in Table I. Each of the databases are randomly partitioned into two sets and ten separate folds for training-testing and 10-fold CV, respectively. In both the cases, the samples are equally distributed with reference to the given classes. A description of the databases is presented in Appendix document, available at https://www.isical.ac.in/~bibl/results/d2gda/d2gda.html

TABLE I
Brief Description of Data Sets

	Data Sets	Sample	Class	View	V^1	V^2	V^3	V^4	V^5	V^6
Benchmark	Digits	2000	10	6	240	76	216	47	64	6
	Cora	2708	7	4	1433	2708	2708	2708	-	-
	NW-OBJECT	30000	31	5	64	225	144	73	128	-
	Reuters	18758	6	5	21531	24893	34279	15506	11547	-
	AwA	30475	50	6	2688	2000	252	2000	2000	2000
Omics	CESC	104	3	4	291368	192	174	12028	-	-
	LGG	374	3	5	293965	181	139	11973	6261	-
	LUNG	546	2	5	294668	180	216	20502	49230	-

B. Model Architecture Based on Error Bound

In the proposed method, an upper bound on the error probability (27) 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 order to determine the optimal number of layers in the proposed D2GDA model, extensive experiments are carried out on both benchmark and omics data sets. During the pretraining, the number of modality-specific hidden layers (L_1) is varied from 1 to 5 for each of the data sets, 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 the analysis of the particular data set. The variation of error bound with respect to L_1 and L_2, corresponding to each of the databases, is presented in Section SVI of the supplementary document for both training-testing and 10-fold CV. The optimal values of L_1 and L_2, obtained from the corresponding error plots, are tabulated in Table II. It 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.

TABLE II
Optimal Number of Layers for Proposed D2GDA Model Based on Estimated Error Bound

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

In the current study, greedy layer-wise pretraining [19] is performed to initialize the model parameters sensibly. From the given set of training samples, the mini-batches are formed to update the parameters of the model. The number of hidden nodes is upper bounded by Theorem 3, presented in Section III-A. Each of the L_1 layers consists of 25 hidden nodes, whereas the L_2 joint layers have 10 hidden nodes each. The considered values for the momentum, weight decay, and number of epochs are 0.5, 0.0005, and 100, respectively. Initialized at 0.01, the value of learning rate is gradually decreased with the increase in number of epochs. In order to estimate of the data-independent expectations, 20 distinct Markov chains and 100 Gibbs steps are considered.

C. Effectiveness of Proposed D2GDA Model

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. The corresponding results are presented in Tables III and IV. From the results presented in Table III for benchmark databases, it can be observed that although MDDBM_PLS performs better than the proposed model in Cora database, the D2GDA model performs significantly better on Digits, Reuters, and AwA data sets and achieves comparable result on NW-OBJECT data set. The results tabulated in Table IV on Omics data sets signify that the proposed model outperforms all the variants for both training-testing and 10-fold CV. Statistical significance analysis reveals that out of the total 24 cases, the proposed model obtains significantly better p-values in 20 cases and better but not significant p-values in the rest 4 cases.

TABLE III
COMPARATIVE PERFORMANCE ANALYSIS OF DIFFERENT VARIANTS OF PROPOSED MODEL ON BENCHMARK DATA

Data Sets	MDDBM	MDDBM _CCA	MDDBM _GMPCA	MDDBM _PLS	D2GDA
Digits	85.60	91.60	88.80	90.00	**97.30**
Cora	64.37	82.13	81.69	**83.02**	82.46
NW-OBJECT	43.75	46.97	46.44	49.38	**55.39**
Reuters	61.25	67.45	62.74	74.26	**86.70**
AwA	56.86	59.76	65.40	60.04	**68.57**

D. Comparative Performance Analysis

Finally, the performance of the proposed D2GDA model is compared with that of several existing approaches. The corresponding results are reported in Tables V and VI.

1) Performance of Consensus Principle Based Methods: In this section, several state-of-the-art consensus principle based methods, namely, regularized generalized CCA (RGCCA) [36], MCCA [1], GMCCA [3], GMKCCA [3], large-scale generalized CCA (LasCCA) [37], distributed CCA (DisCCA) [37], MCCP [6], and TCCA-OS [7], are considered for performance evaluation of the proposed D2GDA architecture. Each of the existing algorithms considers 25 features to represent the joint subspace. From the results reported in Table V, it can be noted that although RGCCA and TCCA-OS achieve considerable results on Digits data, the proposed model exhibits significantly better performance with respect to the existing methods on all the benchmark databases. The results reported in Table VI corresponding to the omics data sets demonstrate that the proposed model outperforms all the eight existing methods for both training-testing and 10-fold CV. Statistical significance analysis reveals that the proposed model achieves significantly better p-values for all the 48 cases.

TABLE IV
PERFORMANCE ANALYSIS OF DIFFERENT VARIANTS OF PROPOSED MODEL ON OMICS DATA SETS

Data Sets	Different Metrics		MDDBM	MDDBM _CCA	MDDBM _GMPCA	MDDBM _PLS	D2GDA
CESC	Train-Test		67.31	61.54	59.62	57.69	**78.85**
	10-fold CV	Mean	64.17	70.00	72.50	70.00	**84.17**
		Median	66.67	66.67	75.00	70.83	**83.33**
		StdDev	**5.62**	8.96	13.64	11.92	6.15
		Paired-*t*:p	1.62E-04	2.14E-03	1.24E-02	4.73E-03	-
		Wilcoxon:p	3.48E-03	7.23E-03	2.11E-02	1.20E-02	-
LGG	Train-Test		79.57	75.00	71.81	77.13	**90.32**
	10-fold CV	Mean	63.95	86.32	84.47	81.58	**98.68**
		Median	63.16	85.53	82.89	81.58	**98.68**
		StdDev	6.08	7.42	6.50	3.51	**1.39**
		Paired-*t*:p	1.28E-08	2.90E-04	2.13E-05	1.51E-08	-
		Wilcoxon:p	2.50E-03	3.42E-03	2.50E-03	2.45E-03	-
LUNG	Train-Test		94.51	95.24	94.87	94.51	**96.34**
	10-fold CV	Mean	90.18	96.96	96.61	96.79	**97.32**
		Median	91.96	97.32	96.43	**98.21**	**98.21**
		StdDev	7.49	**2.92**	3.09	3.24	2.95
		Paired-*t*:p	3.24E-03	2.22E-01	1.84E-02	9.67E-02	-
		Wilcoxon:p	3.79E-03	2.98E-01	3.17E-02	2.46E-01	-

2) Performance of Complementary Principle Based Approaches: Here, the performance of the proposed model is analyzed with reference to that of several complementary principle based approaches, namely, MvDA [13], MvDA-VC [2], and MvCCDA [14]. Each of the existing algorithms considers 25 features to represent the shared subspace. The results corresponding to Table V demonstrate that the existing methods achieve considerable accuracy on most of the benchmark databases. However, the highest classification accuracy is attained by the proposed model in all the cases. From the results reported in Table VI, it can be observed that the proposed model performs considerably better than the existing methods on all the cancer data sets for both training-testing and 10-fold CV. The p-values obtained from the two statistical significance tests indicate that out of total 18 cases, the proposed model attains significantly better p-values for 16 cases and better but not significant p-values in the rest 2 cases.

3) Performance of Deep Learning Models: Finally, the proficiency of the proposed framework is compared with that of several state-of-the-art multi-view deep learning models. These include MDBM [10], mgRBM [9], MvLDAN [15], DACCA [18], DCCA-VG [4], TDDCCA [17], TCCA [5], MDL-CW [11], MMGNN [12], and MVGAN [16], where the shared subspace is represented with 2048, 50, 20, 80, 20, 50, 64$|V|B$, 600, 50, and 50 features, respectively, and $|V|$, B denote the number of input views and total number of batches, considered for a particular data set. The architecture for each of these models follows the same as described in the corresponding papers. From the results reported in Table V, it is evident that significant improvement in classification accuracy is achieved by the proposed architecture in comparison to the

TABLE V
COMPARATIVE PERFORMANCE ANALYSIS ON BENCHMARK DATABASES

Data Sets	Consensus and Complementary Principles Based Classical Approaches											
	RGCCA (2011)	MCCA (1971)	GMCCA (2019)	GMKCCA (2019)	LasCCA (2016)	DisCCA (2016)	MCCP (2022)	TCCA-OS (2023)	MvDA (2012)	MvDA-VC (2016)	MvCCDA (2019)	D2GDA
Digits	90.30	87.00	11.20	6.60	10.20	5.60	89.90	**97.40**	92.40	93.50	92.80	97.20
Cora	52.16	32.85	30.97	30.19	31.63	30.19	68.92	62.04	53.94	55.72	58.71	**82.46**
NW-OBJECT	18.91	30.34	4.56	6.43	7.40	10.93	51.07	49.35	29.03	28.62	37.33	**55.39**
Reuters	55.26	57.50	24.78	28.69	28.67	23.27	79.19	74.03	56.01	55.15	55.36	**86.70**
AwA	6.90	15.08	1.58	3.09	1.59	1.94	57.61	53.48	16.55	15.37	62.67	**68.57**

Data Sets	Deep Learning Models										
	MDBM (2014)	mgRBM (2022)	MvLDAN (2019)	DACCA (2020)	DCCA-VG (2019)	TCCA (2022)	TDDCCA (2016)	MDL-CW (2016)	MMGNN (2020)	MVGAN (2018)	D2GDA
Digits	10.00	88.60	90.70	84.60	89.00	**97.80**	85.90	86.40	88.90	82.70	97.20
Cora	10.99	58.16	63.26	50.06	44.62	56.94	53.05	41.40	42.06	44.95	**82.46**
NW-OBJECT	26.07	32.16	36.27	38.42	19.23	33.61	17.80	18.20	37.87	27.61	**55.39**
Reuters	46.84	58.13	53.36	56.38	64.38	75.97	57.27	62.69	59.16	57.60	**86.70**
AwA	27.16	53.50	47.41	**69.20**	46.59	64.63	53.32	59.58	44.96	69.06	68.57

TABLE VI
COMPARATIVE PERFORMANCE ANALYSIS ON OMICS DATA SETS

Data Sets	Different Metrics		Consensus and Complementary Principles Based Classical Approaches											
			RGCCA (2011)	MCCA (1971)	GMCCA (2019)	GMKCCA (2019)	LasCCA (2016)	DisCCA (2016)	MCCP (2022)	TCCA-OS (2023)	MvDA (2012)	MvDA-VC (2016)	MvCCDA (2019)	D2GDA
CESC	Train-Test		61.54	38.46	42.31	44.23	42.31	36.54	53.85	63.46	42.31	40.38	59.62	**78.85**
	10-fold CV	Mean	75.00	45.83	49.17	38.33	35.00	39.17	65.83	63.33	46.67	50.00	61.67	**84.17**
		Median	79.17	50.00	50.00	41.67	33.33	33.33	66.67	62.50	41.67	50.00	58.33	**83.33**
		StdDev	13.03	13.75	14.41	11.92	15.61	10.43	12.70	9.78	15.32	14.16	12.55	**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	4.39E-05	1.85E-05	7.22E-04	-
		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	2.52E-03	2.47E-03	3.61E-03	-
LGG	Train-Test		41.40	39.78	33.33	38.71	44.09	29.03	76.88	83.87	75.81	73.12	77.96	**90.32**
	10-fold CV	Mean	45.00	35.53	40.53	33.16	38.68	38.68	80.79	72.02	75.79	81.05	77.63	**98.68**
		Median	47.37	34.21	40.79	31.58	36.84	38.16	81.58	71.78	76.32	78.95	77.63	**98.68**
		StdDev	10.99	7.67	8.70	4.99	7.95	7.24	6.80	5.71	8.02	7.83	6.11	**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	3.87E-06	2.56E-05	4.74E-07	-
		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	2.50E-03	2.47E-03	2.52E-03	-
LUNG	Train-Test		87.91	46.52	68.86	86.08	82.42	47.62	92.31	94.87	92.31	91.58	93.77	**96.34**
	10-fold CV	Mean	87.68	51.43	68.39	86.07	85.18	50.71	92.32	93.71	94.82	95.54	96.96	**97.32**
		Median	86.61	50.89	69.64	87.50	85.71	48.21	93.75	93.95	96.43	95.54	**98.21**	**98.21**
		StdDev	4.16	3.13	7.48	8.32	6.68	8.84	6.19	**1.32**	4.16	3.29	3.37	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	2.76E-02	7.48E-03	2.22E-01	-
		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	4.63E-02	1.21E-02	2.97E-01	-

Data Sets	Different Metrics		Deep Learning Models										
			MDBM (2014)	mgRBM (2022)	MvLDAN (2019)	DACCA (2020)	DCCA-VG (2019)	TCCA (2022)	TDDCCA (2016)	MDL-CW (2016)	MMGNN (2020)	MVGAN (2018)	D2GDA
CESC	Train-Test		48.08	63.46	65.38	47.12	65.38	61.54	53.98	65.38	55.77	57.69	**78.85**
	10-fold CV	Mean	52.50	63.33	65.83	39.81	67.50	60.19	78.20	69.17	69.17	61.67	**84.17**
		Median	54.17	62.50	66.67	39.42	70.83	61.54	78.20	66.67	66.67	58.33	**83.33**
		StdDev	17.59	9.78	10.72	5.37	16.87	5.74	**0.06**	14.72	13.64	12.55	6.15
		Paired-*t*:p	3.15E-04	7.68E-05	1.00E-03	7.72E-08	2.94E-03	1.38E-06	6.80E-03	5.00E-03	9.36E-03	7.23E-04	-
		Wilcoxon:p	3.98E-03	3.58E-03	5.76E-03	2.52E-03	8.81E-03	2.53E-03	6.23E-03	1.24E-02	1.30E-02	3.53E-03	-
LGG	Train-Test		65.05	72.04	75.81	65.59	77.96	81.18	66.85	73.12	71.51	74.73	**90.32**
	10-fold CV	Mean	27.63	70.26	76.84	59.35	51.32	77.89	57.14	76.84	72.11	76.84	**98.68**
		Median	18.42	68.42	77.63	58.87	51.32	77.63	57.00	77.63	72.37	77.63	**98.68**
		StdDev	14.90	12.09	7.63	3.42	3.34	4.51	**0.39**	7.63	3.96	7.63	1.39
		Paired-*t*:p	4.59E-08	1.89E-05	2.97E-06	2.00E-11	1.43E-11	2.61E-08	2.31E-15	2.97E-06	1.92E-09	2.97E-06	-
		Wilcoxon:p	2.34E-03	2.52E-03	2.50E-03	2.53E-03	2.49E-03	2.46E-03	2.53E-03	2.50E-03	2.40E-03	2.50E-03	-
LUNG	Train-Test		87.91	87.55	90.48	95.05	89.74	93.41	67.42	93.77	90.84	92.67	**96.34**
	10-fold CV	Mean	61.25	86.96	95.18	95.71	94.11	95.42	67.72	93.93	82.14	94.11	**97.32**
		Median	42.86	90.18	94.64	95.60	95.54	95.24	67.65	95.54	93.75	94.64	**98.21**
		StdDev	23.81	11.67	3.77	1.17	4.69	1.00	**0.38**	4.31	19.32	3.95	2.95
		Paired-*t*:p	3.24E-04	2.95E-03	9.00E-03	9.29E-02	5.99E-03	5.71E-02	1.44E-10	2.00E-04	1.59E-02	2.56E-03	-
		Wilcoxon:p	2.49E-03	2.52E-03	8.24E-03	1.01E-01	5.81E-03	5.71E-02	2.53E-03	3.30E-03	1.03E-02	8.88E-03	-

existing models on all the benchmark databases, except for TCCA and DACCA models on Digits and AwA databases, respectively. For omics data sets, the results are presented in Table VI, which signify that the proposed model outperforms all the existing deep learning models considered. Statistical significance analysis reveals that out of total 60 cases, the proposed model achieves significantly better p-values for 56 cases and better but not significant p-values in the rest 4 cases.

VI. CONCLUSION

The primary contributions of the current study include (a) formulation of a loss function, based on the concept of HSIC,

to capture the relevant cross-view dependency between the given modalities; (b) integrating the principle of cross-view dependency learning with the objective of MDDBM architecture; (c) determining the database specific architecture of D2GDA model based on the estimated error bound; and finally, (d) illustrating the efficacy of the D2GDA model on different domains of application, namely, object recognition, document classification, multilingual categorization, and cancer subtype identification.

In this study, a loss function is developed to efficiently represent the cross-modal dependency across several modalities in terms of coherent as well as complementary structures of the given multi-view data. The MDDBM architecture includes the modality-specific characteristics as well as supervised information of sample categories into the joint subspace. Incorporating the loss function, corresponding to the proposed cross-view dependency analysis, into the learning objective of MDDBM architecture enables the D2GDA model to encapsulate the latent probability distribution of the given multimodal data as well as predict class labels of the given observations. The error analysis and generalization ability establish its effectiveness. The comparative performance analysis demonstrates the proficiency of the proposed model on several multi-view data sets, considering both training-testing and 10-fold CV.

REFERENCES

[1] J. R. Kettenring, "Canonical Analysis of Several Sets of Variables," *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.

[2] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-View Discriminant Analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 188–194, 2016.

[3] J. Chen, G. Wang, and G. B. Giannakis, "Graph Multiview Canonical Correlation Analysis," *IEEE Trans. on Signal Processing*, vol. 67, no. 11, pp. 2826–2838, 2019.

[4] 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*. International Society for Optics and Photonics, 2019, p. 1115009.

[5] 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*, vol. 34, no. 6, pp. 2948–2961, 2022.

[6] 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*, 2022, pp. 19 260–19 269.

[7] 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.

[8] X. Bao, Y.-H. Yuan, Y. Li, J. Qiang, and Y. Zhu, "Learning Supervised Covariation Projection Through General Covariance," in *Proc. of the IEEE International Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[9] N. Zhang and S. Sun, "Multiview Graph Restricted Boltzmann Machines," *IEEE Trans. on Cybernetics*, vol. 52, no. 11, pp. 12 414–12 428, 2022.

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

[11] 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*, 2016, pp. 2601–2609.

[12] 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*, 2020, pp. 12 746–12 756.

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

[14] 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*, vol. 92, pp. 37–51, 2019.

[15] P. Hu, D. Peng, Y. Sang, and Y. Xiang, "Multi-View Linear Discriminant Analysis Network," *IEEE Trans. on Image Processing*, vol. 28, no. 11, pp. 5352–5365, 2019.

[16] 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*, vol. 66, no. 10, pp. 8244–8252, 2018.

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

[18] 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*. SIAM, 2020, pp. 352–360.

[19] R. Salakhutdinov and G. Hinton, "Deep Boltzmann Machines," in *Proc. of Artificial Intelligence and Statistics*, 2009, pp. 448–455.

[20] M. Hein and O. Bousquet, "Kernels, Associated Structures and Generalizations," *Technical Report*, no. 127, 2004.

[21] A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf, "Kernel Methods for Measuring Independence," *Journal of Machine Learning Research*, vol. 6, pp. 2075–2129, 2005.

[22] M. Reed and B. Simon, *Methods of Modern Mathematical Physics: Functional Analysis*. Academic Press, 1980.

[23] D. Kumar and P. Maji, "Discriminative Deep Canonical Correlation Analysis for Multi-View Data," *IEEE Trans. on Neural Networks and Learning Systems*, pp. 1–13, 2023. [Online]. Available: 10.1109/TNNLS.2023.3277633

[24] R. M. Neal and G. E. Hinton, "A View of the EM Algorithm That Justifies Incremental, Sparse, and Other Variants," in *Learning in Graphical Models*, M. I. Jordan, Ed. Springer Netherlands, 1998, pp. 355–368.

[25] T. Tieleman, "Training Restricted Boltzmann Machines Using Approximations to the Likelihood Gradient," in *Proc. of the 25th International Conf. on Machine Learning*, 2008, pp. 1064–1071.

[26] S. Y. Sekeh, B. Oselio, and A. O. Hero, "Learning to Bound the Multi-Class Bayes Error," *IEEE Trans. on Signal Processing*, vol. 68, pp. 3793–3807, 2020.

[27] H. Hotelling, "Relations Between Two Sets of Variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[28] 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*. IEEE, 2012, pp. 2160–2167.

[29] H. Wold, "Path Models with Latent Variables: The NIPALS Approach," in *Quantitative Sociology*, H. M. Blalock, A. Aganbegian, F. M. Borodkin, R. Boudon, and V. Capecchi, Eds. Academic Press, 1975, pp. 307–357.

[30] M. Gnen, "Bayesian Supervised Dimensionality Reduction," *IEEE Trans. on Cybernetics*, vol. 43, no. 6, pp. 2179–2189, 2013.

[31] R. Rossi and N. Ahmed, "The Network Data Repository With Interactive Graph Analytics and Visualization," in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 29, no. 1, 2015.

[32] 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.

[33] 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*, vol. 22, pp. 28–36, 2009.

[34] 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*, vol. 41, no. 9, pp. 2251–2265, 2019.

[35] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The Cancer Genome Atlas (TCGA): An Immeasurable Source of Knowledge," *Contemporary Oncology*, vol. 19, no. 1A, p. A68, 2015.

[36] A. Tenenhaus and M. Tenenhaus, "Regularized Generalized Canonical Correlation Analysis," *Psychometrika*, vol. 76, no. 2, pp. 257–284, 2011.

[37] 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*, 2016, pp. 871–876.