Multi-View Correlation and Discriminant Analysis:
Structure Preservation, Sparsity to Multi-Task Learning

A thesis submitted to Indian Statistical Institute
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computer Science

by

Sankar Mondal
Senior Research Fellow

Under the supervision of
Dr. Pradipta Maji, Professor

Machine Intelligence Unit
Indian Statistical Institute, Kolkata

December 2024

Patience, perseverance, and belief:
possiblities

Acknowledgements

The journey of a PhD is somewhere similar to that of life, where new ideas are born, grow up with uncertainty, and ultimately end with knowledge and experience. I would like to convey my profound appreciation to all individuals who have provided support and guidance during this journey.

Above all, I express my sincere gratitude to my supervisor, Professor Pradipta Maji, for his important mentorship, perceptive input, and unwavering assistance. I am immensely thankful for his mentorship, since his skills and support have played a crucial role in molding my thesis. The usual expressions of gratitude are insufficient to convey the immeasurable dedication he has shown in enhancing my technical expertise and motivating me to become an autonomous and innovative researcher.

I wish to convey my gratitude to Subhas sir, my beloved childhood teacher. The basic mathematics I learned from him have helped me get this far. A special thanks to Sukriti Roy for always being a good friend and standing beside me during my challenging times. I express my sincere thanks to my friends, Bibhuti, Mainak, Rathindra, Abhinav, Sucheta, Anjan, Joginder, Susanta, and Priyanka, for their support and companionship. I am grateful to have such friends, and I have countless unforgettable memories with them. I wish to express my gratitude to my hostel seniors, Aparajita di, Sanchayan da, Avisek da, and Manjari di, for their valuable suggestions and for treating me like their younger brother. During my tenure, I have come across some excellent fellow hostelmates: Anup, Subha, Gourab Chandra, Gourab Kumar, Suman, Faizan, Koushik, Debanjan, Meghna, Sudipa, and Monisha. The love and respect I got from them was immeasurable.

I am profoundly grateful to my Biomedical Imaging and Bioinformatics Lab members, Ankita Di, Suman Da, Debamita Di, and Ekta Di, for their companionship and cooperation. They are all like elder brothers and sisters to me. They have shown great love and affection to their younger brother throughout this journey.

I would like to express my gratitude to the Dean of Studies and the Director of the Indian Statistical Institute for providing me with exceptional research facilities and a conducive atmosphere. I express my sincere gratitude to the faculty members of the Machine Intelligence Unit at the Indian Statistical Institute for their enlightening lectures, valuable feedback, and motivation. The cumulative knowledge from all of you have significantly enhanced my research experience. I would like to express my gratitude for the consistent and timely

assistance provided by the administrative personnels of our institute during my Ph.D. program.

Finally, and of utmost importance, I would like to convey my deep appreciation to my family: my mother Mrs. Suchitra Mondal, father Mr. Ratan Mondal, elder brother Dipak Mondal, sister-in-law Tanushree Mondal, and my little nephew Anantajit Mondal. Words cannot express the depth of their love and support.

Sankar Mondal

Sankar Mondal

Abstract

Advancements in data acquisition make multiple data sources available to explain different perspectives of an object. In order to enhance the performance of a single-task learning such as classification, the multi-view learning (MVL) leverages the complementary and consistent information across multiple views. However, MVL has its own set of challenges. The major issues associated with MVL include selecting relevant and informative views while discarding the noisy and redundant views, integrating heterogeneous views while constructing discriminant subspaces, handling "high-dimension low-sample size" nature of different views, and finding the intrinsic non-linear class-geometry of the data across all the views. Moreover, applying MVL under a multi-task learning (MTL) framework, for learning multiple related tasks simultaneously to improve the performance of single-task MVL, is a major challenge.

In this regard, the thesis introduces some supervised MVL algorithms, based on the theories of canonical correlation analysis (CCA). In order to construct the discriminative subspaces while preserving the non-linear class-geometry of the data, a novel supervised MVL method, termed as class-structure preserving multi-view correlated discriminant analysis (CSP-MVCDA), is proposed, which judiciously integrates the merits of multiset CCA (MCCA), linear discriminant analysis (LDA), and a locality preserving norm. The proposed method jointly optimizes the inter-set correlation across all the views and intra-set discrimination in each view to obtain a common discriminative latent space, where the shared and complementary information across multiple views is exploited. The locality preserving norm with prior class labels helps to preserve the local class-structure of the data, while both MCCA and LDA take care of its global class-structure across multiple views. A closed form solution, based on the generalized eigenvalue problem, makes the proposed method applicable for high-dimensional multi-omics data integration. In order to compute view relevance and inter-view dependency for a desired task, and to address the problem of "high-dimension low-sample size" nature of different views, a novel supervised MVL method, termed as supervised graph regularized multi-view canonical correlation and discrimination analysis (SGR-MCCDA), is next introduced based on the maximum variance formulation of MCCA. Incorporating the known geometry of source vectors encoded by the within-class and between-class graphs, the proposed method preserves the class-structure of the data, which facilitates multi-omics cancer stratification.

In imaging genetics study, sparse models are effective to select diagnosis- or task-specific features for a comprehensive understanding of the underlying disease, and to find the genetic basis for the brain function and structure associated with the disease. In this regard, a new sparse multi-task two-view algorithm, termed as multi-task learning and sparse discriminant canonical correlation analysis (MTL-SDCCA),

is proposed, judiciously integrating the theories of CCA and LDA under the MTL framework to find the association between an imaging and a genetic modality. It uses lasso and group lasso penalties to select the diagnosis-specific and diagnosis-consistent features from the large number of features to identify group-wise imaging genetic associations. In order to reduce the high complexity of existing algorithms, under multiple imaging and genetic modalities, a multi-task multi-view algorithm, termed as multi-view multi-task sparse canonical correlation analysis (MvMt-SCCA), is proposed, which learns multiple sparse CCA tasks together for identifying the group-wise imaging genetic association. Incorporating the lasso and fused lasso penalties, the proposed method is able to select the modality-wise, class-specific, and class-consistent features for large-scale imaging genetics studies.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the world of digital enhancement, data has evolved into a vital resource that reshapes the way we live, work, and make choices. Data, in its raw form, consists of facts, figures, and information collected from various sources and events. It can include everything from the numbers to complex multimedia files, including text, photos, videos, and sensor readings. The usefulness of data lies in its ability to provide insights, make decisions, and solve problems across a multitude of domains, like, business, health care, education, and scientific research. There has been an explosive growth in the amount of data generated worldwide over the past few decades. For example, in the early 2010s, the world generated approximately 1.2 zettabytes (1.2 trillion gigabytes) of data annually, whereas in recent times, the world has produced over 120 zettabytes of data annually, and the prediction says that the amount will reach 175 zettabytes per year in 2025 [212]. Apart from the increment in volume of data, the diversity of the data has also been increasing drastically. From one perspective, it is useful to understand the digital world, but on the contrary, it brings in the challenges of storing and analyzing of such huge data as well.

Data analysis acts as a link between raw data and insights that may be used to make decisions. It involves techniques such as data cleaning, visualization, statistical analysis, and more. Through this process, the data can be transformed into knowledge. A pattern in data, which gives knowledge about the data, refers to a repeating, significant structure or feature that may be observed within a data set. Patterns can take various forms and may include trends, seasonality, anomalies, clusters, correlations and many more. In the realm of relentless stream of data, pattern recognition, an automated process of identifying meaningful patterns or structures within data sets, becomes a necessary technique for the data analysis [248]. Pattern recognition is the fundamental aspect of both machine learning and artificial intelligence (AI) [26]. Machine learning, a branch of AI, employs algorithms to help computers learn from data and get better over time. It is essentially the practical use of pattern recognition. AI is a more general term, which strives to develop robots or systems that can mimic human intellect including abilities to comprehend natural language, reason, solve problems, and make decisions. AI systems leverage pattern recognition and machine learning to make sense of complex data, adapt to changing environments, and perform tasks that require intelligence.

Pattern recognition and machine learning algorithms may be roughly categorized into the following four groups depending on the learning technique. The difference between

these four groups, namely, supervised, unsupervised, semi-supervised and reinforcement learning, is represented pictorially in Fig. 1.1.

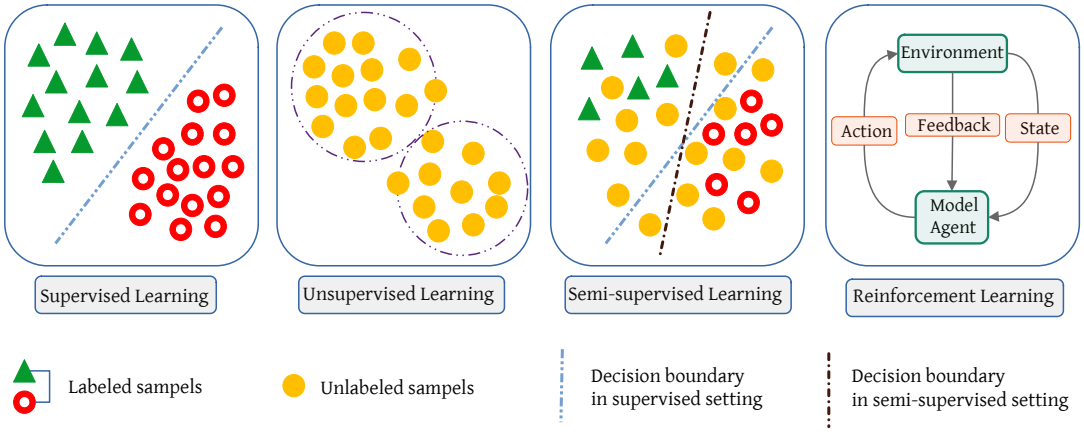


Figure 1.1: Difference between supervised, unsupervised, semi-supervised, and reinforcement learning.

- The primary objective of **supervised learning** is to learn a mapping from a set of given inputs or objects to a set of given labels or target variables, so that it can accurately classify or predict the output against unseen inputs or objects. It can be broadly viewed as two category, namely, classification and regression. The classification algorithms are applied when the target variables are categorical. These algorithms predict the category in which the unseen objects belong to. The examples include classifying emails as spam or not spam, classifying images of animals into different species, and so on. The regression algorithm involves a continuous target variable that predicts a numeric value, like predicting house prices depending on features such as area, number of bedrooms, transportation cost, etc.

- **Unsupervised Learning** does not require the prior labeled data, rather focuses on discovering hidden patterns, structures, or relationships within unlabeled data. It is very helpful for exploratory data analysis and finding insights when the expected results are not known in advance. Clustering and anomaly detection are some of the common tasks in unsupervised learning. Clustering algorithms group similar data points based on their intrinsic characteristics or similarities. For example, retailers use clustering to group their clients based on demographics, purchasing habits, and past purchases. Anomaly detection recognizes unusual or unexpected data items, which deviate from the norm in an unsupervised way. Anomaly detection is beneficial for fraud detection, network security, and quality control.

- **Semi-supervised Learning** combines aspects of both supervised and unsupervised learning. In this approach, the algorithm is trained on a data set that includes large amount unlabeled samples and a small amount labeled samples. In many real-world applications, collecting and annotating large amount of labeled data can be a

major barrier. In such cases, semi-supervised learning is very useful. Applications of semi-supervised learning include text classification, image recognition and speech recognition, where large number of instances are unlabeled and very few are labeled.

- **Reinforcement Learning**, a machine learning technique, aims to learn an agent-environment system, where the agent is trained to make sequential decisions in the environment in such a way that the cumulative reward based on each action by the agent is maximized. It is influenced by behavioral psychology, where an agent learns through trial and error by interacting with its surroundings. Due to its capacity to handle complicated decision-making tasks, such as game playing, robotics, and autonomous systems, reinforcement learning has drawn a lot of interest in machine learning.

The machine learning algorithms usually deal with a data set represented by a two dimensional matrix, $X^{n\times d}$, where n denotes the number of instances and d denotes the number of features for each instance. This type of representation is known as feature vector based representation of a data set. It can include numerical, categorical, textual, or binary feature sets for n samples in a d-dimensional measurable space. For example, in an image data set, an image is considered as a sample and intensity values of the pixels of the image can be considered as features. A "view" or "modality" of a data set is a collection of n samples that are represented by d-dimensional feature vectors.

The single-view learning is a machine learning paradigm in which a model is trained and used to operate on just one data source or view. It is appropriate for many machine learning applications when a single set of features is enough to generate reliable predictions or inferences. Traditional supervised or unsupervised algorithms, such as k-means clustering, support vector machine, and linear regression, often belongs to this category. Though single-view learning is popular and efficient in many applications, it may not completely use the potential of data when several viewpoints or perspectives are available.

1.1 Multi-View Learning

In many real-world applications, data can be represented by many ways or can be obtained from different perspectives. Multi-view learning (MVL) deals with the multiple views together to obtain insightful decisions from the data [237]. In comparison to single-view learning, MVL makes use of the strengths and complementary information of several data representations to produce models that are more reliable, accurate, and interpretable. It has gain enormous success in a wide range of practical applications, such as, multi-omics data integration, imaging genetics association, face recognition, biomedical imaging, natural language processing, and so on [151, 186, 209, 310]. For instance, in imaging genetics applications, genetic data, such as, single nucleotide polymorphism (SNP) or copy number variation, is integrated with different imaging modalities, such as, magnetic resonance imaging (MRI), diffusion tensor imaging, and positron emission tomography (PET), to find reliable biomarkers for improving the disease diagnosis. Figure 1.2 illustrates a few of the numerous applications of MVL.

Some of the key characteristics and concepts of MVL are discussed below, which are the main reasons behind the success of MVL in machine learning applications.

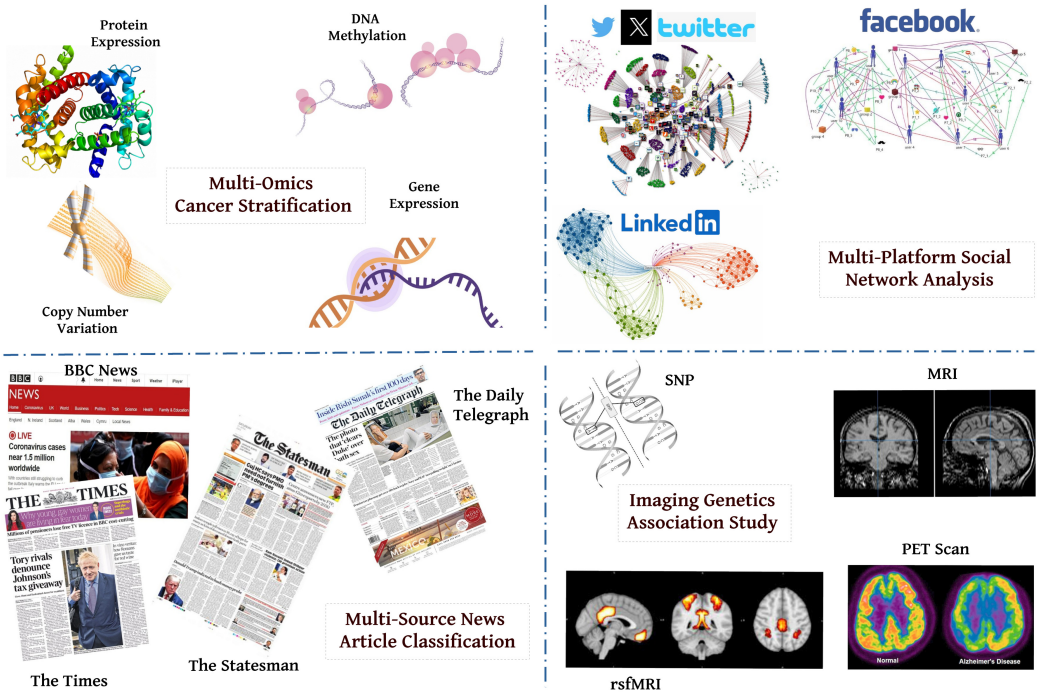


Figure 1.2: Diverse applications of multi-view data analysis.

- **Consistency and Agreement:** Many MVL techniques place a strong emphasis on the premise that data from multiple viewpoints should be coherent or agreeable with one another. Inconsistent information might hinder learning and inhibit the model's ability to utilize the multiple viewpoints. For instance, if two interpretations of the same piece of data are consistent, then their predictions or depictions ought to be comparable or congruent.

- **Complementary Information:** Even though all of the views of a system capture the same aspect of it, each view may provide complementary information that is not available in the other views, which potentially improve the performance and robustness of the learning models. For example, MRI and PET scans are two modalities to express the brain abnormalities of an individual. The MRI scan gives the tissue level mapping, which helps to identify the tumor region or decrease in brain volume, whereas the PET scan gives functional level information of the brain to identify the abnormal behavior of the brain.

- **Comprehensive View of the System:** Multiple distinct views of a system can give comprehensive knowledge about the system and upon integration, it has strong influence over the learning process. For instance, multiple cameras, capturing a person from different angles and perspectives, can recognize the person more conveniently than a single-view of the camera.

- **Resilience to Noise and Data Variability:** The real data is often noisy, incom-

plete, or subject to variations. The multiple distinct views help in minimizing the impact of noise or missing data in individual views. The model becomes more resilient to data variability by relying on complementary sources.

1.2 Challenges in Multi-View Learning

The MVL has great potential in dealing with multiple perspectives of data over single-view learning in various applications. But, it also has its own set of challenges due to the complexities in multi-view data [310]. The major challenges in MVL are as follows:

- **High-dimension Low-sample Size:** Data sets from real-world applications typically contain a large number of observed variables. An image may have 10^6 number of pixels, a DNA microarray contains almost 20,000 genes, a text file contains millions of words, and so forth. Conversely, there are often very few observed samples. The learning models tend to overfit the data as a result of the small number of training samples, which lowers the generalization performance. In the domain of multi-omics integration, large number of features and low-sample size nature of different views make the sample covariance matrix non-invertible, which lead to poor performance of the model. In case of imaging genetics study, the large number of features ($\sim 10^6$) compared to samples ($\sim 10^2$ or 10^3) requires sparse models to select the important feature from such huge number of features.

- **Non-linear Geometry:** The majority of the views in several real-world data sets have large number of attributes. Although, these views seem like a point cloud in a high-dimensional feature space, their important structures are often embedded in a lower-dimensional manifold or subspace that is part of the high-dimensional space. Hence, most of the variables in these views are collinear and should be taken care of while learning from the data. Moreover, capturing intrinsic non-linear class-geometry of the data embedded into high-dimensional space could also be a challenging task while fusing multiple views.

- **View Heterogeneity:** The most straightforward method for handling multi-view data is to create a single view by concatenating the feature sets of multiple views, and apply traditional machine learning techniques. But, the concatenation of features is not always effective, because each data view has some unique statistical properties and also differs in unit, scale and variance that are not usually compatible across the multiple views. For example, in imaging genetics study, the imaging views like functional MRI (fMRI) measures the frequency fluctuations in the brain, which are continuous data, whereas in genetic view like SNP data are categorical depending on minor allele count. In multi-omics study, the features of DNA methylation data represent the $\beta-$values, which range from 0 to 1, whereas the features in gene expression data are measured in RPM (reads per million), which may have size of 10^5. So, the concatenation can be biased towards the views, which have high scale and variance. Therefore, in order to preserve the intrinsic features of each view throughout the learning process, the data integration needs to remain impartial.

- **View Selection:** In real-world applications, measurement errors might lead to noise in the observations across different views. The noise in one view can be propagated in other views or even can be inflated during the data integration. The majority of machine learning algorithms assume that all the views are informative, and can produce uniform and consistent knowledge about the data set. However, in reality, there might exist views which contain repetitive, unimportant or even worse information. The learning process may not perform well when all available views are integrated, due to the presence of redundant, irrelevant, and noisy views. Therefore, it is important to decide which viewpoints to employ and how much weight to give them during the data fusion process.

- **Scalability:** In MVL, scalability describes an algorithm's capacity to manage large-scale data across several views in an effective manner. Each view can contain large number of features, which may increase the total feature size extremely high, while integrating the multiple views. For instance, in imaging genetics studies, genetic data like SNP has very high feature size of order 10^6, neuroimaging data like fMRI contains 10^4-10^7 number of voxel level measurements, which makes the data integration process challenging. Therefore, designing a machine learning model, which is scalable to the dimension of the views, is a difficult task.

- **Incomplete Views:** The conventional MVL methods assume that the multi-view data is complete, which means all the views have the same set of samples. However, in real-world scenario, this is not the case, since different artificial factors like equipment failure during the data collection procedure make the views incomplete. According to the missing information, incomplete data are mainly of two kind: (1) the information about a sample is completely missing in some of the views (view missing) and (2) the information about the sample is partially observed in some of the views (variable missing). Removing the samples having missing information from all the views could reduce the sample size, which leads to the overfitting problem. Another strategy is to impute the missing values by some special values such as 0 or mean of the corresponding attributes. But, this approach may introduce additional noise into the data and degrade the performance. Therefore, designing a machine learning model, which is able to find the underlying data distribution or a latent representation from the incomplete views, would be of great interest.

- **Overfitting:** In MVL, there is a limited number of training data while there exists different kinds of features to work with, which makes traditional machine learning model to be overfitted. The cross-validation and regularization strategies are frequently employed to lessen this issue.

While certain challenges, such as incomplete views and heterogeneous data, are exclusive to multi-view data, other challenges, such as capturing data geometry and high-dimension low-sample size nature, are also present in single-view data.

1.3 Multi-Task Multi-View Learning

The MVL methods are often used for dimensionality reduction, semi-supervised learning, supervised learning, transfer learning, and clustering. However, the MVL methods usually extract or select feature sets for single-task models. A good learning algorithm for single-task models needs a large amount of labeled training data and may have a large number of parameters to estimate. On the other hand, multi-task learning (MTL) [39] has the ability to improve the performance of single-task model by the knowledge-transfer approach. It can particularly be helpful when there is limited data available for the targeted task, and a lot of data from related tasks is easily available. Figure 1.3 explains the difference between single-task learning and MTL. The multi-task extension of mulit-view learning, termed as multi-task multi-view learning [98, 303], aims to use several MVL tasks simultaneously, to enhance the performance of each task by using relevant information from the related tasks.

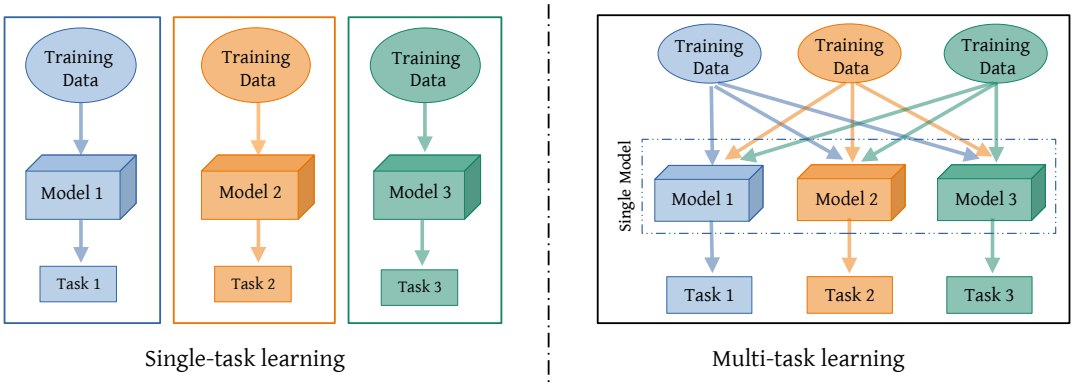


Figure 1.3: Difference between single-task and multi-task learning.

The main objective of multi-task learning (MTL) [39] is to train models that perform multiple related tasks simultaneously, instead of training separate models for each task. The model learns generalized representations of the data, which are applicable in multiple context by using all of the available data across different tasks. This idea is rooted in the notion that knowledge acquired from one task can leverage the performance of another task. Primarily, there are two factors in MTL. The first factor is task relatedness and the second one is the definition of task. The task relatedness relies on the knowledge of the relationships between various tasks, which can be incorporated into the MTL model for learning. The definition of the task is the second factor. The supervised tasks such as classification and regression, unsupervised tasks such as clustering, semi-supervised tasks, active learning, reinforcement learning, online learning, and multi-view learning are the primary types of learning tasks in machine learning. As a result, different learning challenges result in various MTL settings. In the context of classification problem, MTL tries to improve multiple classification tasks by learning them concurrently. An example of MTL in the domain of computer vision is facial recognition and emotion detection. AI systems can efficiently learn to identify faces while also identifying and deciphering facial expressions and emotions through MTL. By fostering a more comprehensive knowledge of

human face characteristics, this simultaneous learning technique improves the accuracy of emotion detection and facial recognition systems.

Some of the key aspects of MTL are given below, which make it a widespread machine learning approach across multiple domains, like imaging genetics, natural language processing (NLP), recommendation system and so on.

- **Shared and Complementary Knowledge:** The multiple tasks in MTL share some commonalities or underlying structure, and also have differences across the tasks, which are explored to come up with a more generalized and improved model. A real-life example is a text-based task in NLP, where multiple language related tasks are performed simultaneously, benefiting from shared and complementary knowledge across different tasks. Google's BERT (Bidirectional Encoder Representations from Transformers) is a prominent example of MTL in NLP.
- **Regularization Effect:** The MTL also functions as a regularizer. Training on multiple tasks helps to prevent overfitting problem because the model must balance its parameters to perform well on all tasks. This can result in better generalization to new data.
- **Data Efficiency:** The MTL can be useful when it comes to deal with limited data available for each individual task. It allows the model to leverage information from multiple tasks to boost performance.
- **Reduced Model Complexity:** By combining several tasks into one model, MTL can lower the overall complexity of a machine learning system.

As mentioned above, the design of MTL models is influenced by the comprehension of task relatedness. There are two primary aspects to define the task relatedness: feature-based and parameter-based. The feature-based MTL models assume that tasks with variable features have comparable or equivalent feature representations, which may consist of a subset or a transformation of the original features. The parameter-based MTL models embed task relatedness via regularization or prior knowledge on model parameters. In the domain of imaging genetics, feature-based MTL models are more popular [63, 74].

1.4 Challenges in Multi-Task Learning

Although the MTL framework has drawn a lot of interest due to its ability to increase efficiency and generalization by simultaneously training models on several related tasks, it comes with the following challenges:

- **Task Interference:** It is possible that different tasks in MTL require different representations or have competing aims. Task interference can occur when learning to perform well on one task has a detrimental effect on performance on another. For example, named entity recognition (NER) and sentiment analysis are two tasks in NLP. While sentiment analysis categorizes a text's sentiment (positive, negative, or neutral), NER recognizes and categorizes named entities, such as individuals, groups, and places. Due to potential conflicts between features relevant for NER (proper

nouns) and features beneficial for sentiment analysis (sentiment-related terms), these tasks may generate interference.

- **Transferability:** Improved performance can result from exchanging information across tasks, but there is no guarantee that knowledge gained from one task will be applied to others. Effective regularization strategies and the identification of tasks with complementary information are critical components of successful transfer.
- **Data Heterogeneity:** Different types of data, such as text, images, and structured data, may be involved in tasks inside an MTL framework. Therefore, it is necessary to build architectures that can handle a variety of data modalities.
- **Hyperparameter Tuning:** When it comes to tuning hyperparameters, MTL models frequently have more hyperparameters than single-task models. Determining the optimal set of hyperparameters that performs well for every task can be computationally expensive and time-taking.
- **Task Imbalance:** The tasks can differ with the relevance, complexity, and availability of data. To keep dominant tasks from overshadowing others throughout training, it is imperative to balance each task's contribution. For instance, finding anomalies (like tumours) and segmenting organs are two aspects of medical image analysis that aid in diagnosis. However, compared to organ segmentation, acquiring labeled data for anomalies is sometimes more difficult and costly. It is important to balance these tasks' influence during training so as to prevent the more accessible task from taking precedence.
- **Architecture Design:** It is difficult to design a shared architecture that can efficiently collect features relevant to every task. It is not easy to balance the model's complexity to prevent overfitting and make sure the model has enough capacity for each task.

In order to handle effectively the challenges corresponding to MVL and extract significant patterns hidden in multi-view data sets with respect to both single and multiple tasks, some more sophisticated algorithms need to be developed.

1.5 Scope and Organization of Thesis

One of the major challenges in multi-view data analysis is to capture the intrinsic data geometry with respect to each class across multiple distinct, high-dimensional, and heterogeneous views. Another significant issue with multi-view data analysis is the high-dimensional, low sample size nature of individual views, which makes the sample covariance matrix ill-conditioned. It is also possible that some of the views offer contradictory, redundant, or even worse information. Moreover, most of the MVL algorithms are designed to handle the single-task of extracting a set of features for either classification or clustering. But, task-specific feature selection under the MTL framework has gained interest recently in many real-world applications [39] [98]. For example, in the field of imaging genetics, diagnosis-specific feature selection facilitates the diagnosis and prognosis of many neurodegenerative disorders, such as Alzheimer's disease and Parkinson's disease. The principal

contribution of the thesis is to develop some novel algorithms for extracting or selecting meaningful and pertinent features from multi-view data sets with respect to single as well as multiple tasks.

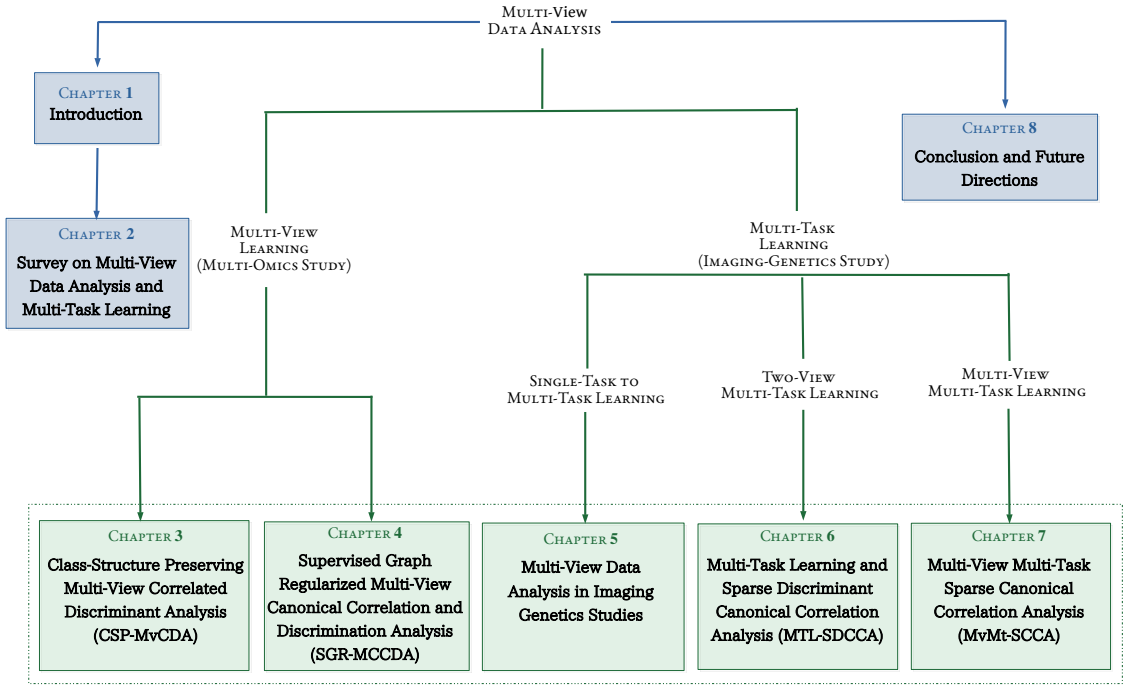


Figure 1.4: Outline of the thesis.

The outline of the thesis is given in Fig. 1.4. There are eight chapters in the thesis. In Chapter 1, the significance of multi-view data analysis and MTL is discussed. This chapter also covers some of the challenges of multi-view data analysis and MTL. Chapter 2 presents a brief review on existing multi-view data integration and MTL strategies.

The multiset canonical correlation analysis (MCCA) is a subspace learning method, which seeks a low dimensional latent space, where the pair-wise correlation among the multiple views is maximized. However, most of the MCCA based approaches do not exploit the prior knowledge of the data in terms of the labels or the geometry of the source vectors. In this regard, a supervised subspace learning method, termed as class-structure preserving multi-view correlated discriminant analysis (CSP-MvCDA), is introduced Chapter 3 to capture the class-geometry present in the data, exploring the shared and complementary information across multiple views. It judiciously integrates the merits of multiset canonical correlation analysis (MCCA), linear discriminant analysis (LDA), and a locality preserving norm. The proposed method jointly optimizes the inter-set correlation across all the views and intra-set discrimination in each view to obtain a common discriminative latent space. The locality preserving norm with prior class labels helps to preserve the local class-structure of the data, while the LDA maintains its global class-structure. In effect, both the local and global class-geometry of data can be preserved. Moreover, the proposed formulation of CSP-MvCDA has a closed-form solution to the problem, which most of the existing models lack. It is shown to be the simplest and most effective among the popular

existing supervised MVL techniques, namely, GMA, MULDA, and MvDA, since it has only two parameters to be optimized. Moreover, a strategy based on geometric separability index has also been introduced to find out the optimal parameters of the proposed method. The effectiveness of the proposed method has been shown by evaluating its performance on several multi-omics cancer and benchmark data sets. The experimental results establish that the proposed CSP-MvCDA method is superior to several state-of-the-art algorithms in terms of classification performance and preserving class-geometry of the data.

In MVL, the proper use of consistent and complementary information from different views is necessary to handle view dependency and view discrepancy. Most of the MVL algorithms ignore the view dependency problem and assume that each view is equally relevant to a desired task, which may not be the case. In this regard, a supervised MVL algorithm, based on maximum variance (MAXVAR) formulation of MCCA, is proposed in Chapter 4. The proposed formulation is termed as supervised graph regularized multi-view canonical correlation and discrimination analysis (SGR-MCCDA). The proposed model utilizes the label information, incorporates the known geometry of source vectors encoded by the within-class and between-class graphs, and learns a proper weight for each view according to their relevance in a unified way. An iterative rule is introduced for the optimization of SGR-MCCDA. The computational complexity of the proposed model and the convergence analysis of the iterative rule are also presented to justify the performance of the proposed model. To show the impact of the proposed method, extensive experiments over several multi-omics cancer and benchmark data sets are conducted. The results confirm that the proposed model surpasses the existing algorithms, in terms of the classification accuracy and has comparatively lower execution time.

In imaging genetics research, the main objective is to investigate the complex genotype-phenotype association for the disease under study. To understand the impact of genetic variations over the brain functions and structure, the genotypic data such as SNP is integrated with the phenotypic data such as imaging quantitative traits. In this regard, two new MVL algorithms, namely, CSP-MvCDA and SGR-MCCDA, proposed in Chapter 3 and Chapter 4, respectively, are successfully used to extract features for imaging genetics study in Chapter 5. A comparative performance analysis with existing MVL algorithms is also presented in this chapter. A real imaging genetics data set, obtained from Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort, is used for the analysis by using the Alzheimer's disease (AD) related genetic variants (SNPs) and brain imaging (fMRI) modalities. The results establish that the classification performance of proposed CSP-MvCDA and SGR-MCCDA on the ADNI data set is better than that of the existing MVL models. Moreover, these two approaches perform well with respect to capturing the correlation between imaging and genetic data, although they are not capable of identifying the disease-specific imaging genetics association, which necessitates multi-task framework in this field.

The sparse models, based on canonical correlation analysis (CCA), are popular in this area to find the complex bi-multivariate genotype-phenotype association, as the number of features in genotypic and/or phenotypic data is significantly higher as compared to the number of samples. However, the sparse CCA based methods are, in general, unsupervised in nature, and fail to identify the diagnose-specific features those play an important role for the diagnosis and prognosis of the disease under study. In this regard, a new supervised multi-task model is proposed in Chapter 6 to study the complex disease-specific

genotype-phenotype association. It judiciously integrates the merits of CCA and LDA under the MTL framework. The proposed model can identify the diagnose-specific as well as the diagnose-consistent features with significantly lower computational complexity. The performance of the proposed method, along with a comparison with the state-of-the-art methods, is evaluated on several synthetic data sets and one real imaging genetics data collected from ADNI cohort. In the current study, the SNP as genetic data and resting state functional MRI (fMRI) as imaging data are integrated to find the complex genotype-phenotype association. An important finding is that the proposed method has better correlation value, improved noise resistance and stability, and also has better feature selection ability. All the results illustrate the power and capability of the proposed method to find the diagnostic group-specific imaging genetic association, which may help to understand the neurodegenerative disorder in a more comprehensive way.

Imaging genetics primarily focuses on the study of neurodegenerative disorders to identify the complex connections between genetic variations and brain activity for the disease under consideration. The MTL and sparse CCA (SCCA) are the two effective approaches to integrate the high-dimensional genetic and imaging data. However, conventional MTL based SCCA approaches focus either on modality-wise or class-wise feature selection to find the association between genetic and imaging data. Moreover, these approaches require computing the inverse of a very high-dimensional covariance matrix, which has large time and space complexity. It makes the models inefficient for large-scale imaging genetics studies. In this regard, a multi-view multi-task sparse canonical correlation analysis (MvMt-SCCA) is proposed in Chapter 7 to identify the complex relationships between genetic and multimodal imaging data. It uses the lasso and fused lasso penalties simultaneously to select the modality-wise, class-specific and class-consistent features. An iterative algorithm is introduced using the block coordinate descent technique to solve the optimization problem efficiently. The proposed algorithm has very low time and space complexity, which makes the model suitable for large-scale imaging genetics studies. The performance of the proposed method is evaluated on both simulated and real-world data sets. The results on simulated data show that the proposed method provides a higher correlation value, has a better feature selection capability, and is robust with respect to noise. The real data is collected from the ADNI cohort. The study using ADNI data finds a group of risk genes and modality-wise abnormal brain regions corresponding to each diagnosis group, which may contribute to a more comprehensive understanding of the neurodegenerative disease.

The thesis is finally ended in Chapter 8, which discusses the improvements and future directions of the proposed research work.

Chapter 2

Survey on Multi-View Learning

This chapter provides an introduction to the fundamental concepts of multi-view learning and multi-task learning, along with a concise review of existing literature on the topics.

2.1 Multi-View Data Analysis

Suppose there are M number of views of n data samples $\{x_1, x_2, \ldots, x_n\}$, where $M \geqslant 2$. This creates multi-view data with M number of views and n samples. The thesis uses the terms "view" and "modality" interchangeably. The modalities or views may have either feature vector based or relation based representations. In feature vector based representation, multi-view data having M number of views can be represented by the M data matrices X_1, X_2, ..., X_M. Each data matrix, X_m, has p_m number of features for the n samples observed in a p_m dimensional measurement space. The measurement space is often considered the Euclidean space, that is, $X_m \in \mathbb{R}^{n \times p_m}$. However, the views can consist of other types of measurements, such as textual, categorical, binary, and so on. The dimensions of each data matrix need not be equal, and the measurement space can be different across multiple modalities. The data matrices, X_1, X_2, ..., X_M, might differ in terms of the variance, scale, and data distribution. Multiple views, in the case of relational data, can be represented by the similarity matrices, S_1, S_2, ..., S_M, where each S_m is a $(n \times n)$ matrix, defined by $S_m = [s_{ij}^m]_{n \times n}$, s_{ij}^m being the similarity or distance between the samples x_i and x_j of the m-th view.

An example of multi-view data in the domain of imaging genetics is given in Fig. 2.1. Genetic data such as SNP contains information about the change of a single nucleotide base in the genome sequence, whereas different neuroimaging modalities such as fMRI, PET-FDG, and PET-AV45 capture different quantitative traits (QTs) of the brain in terms of its function, glucose metabolism, and amyloid load, respectively. The objective is to analyze multimodal imaging and genetic data to find the genetic basis for brain function and structure. In an imaging genetics data set, the genetic data and multiple neuroimaging data are collected for n number of patients, which results in generating multi-view data with M number of modalities: X_1, X_2, ..., X_M. Each X_m contains different types of measurements, such as nucleotide changes in DNA, frequency fluctuations in fMRI, glucose metabolism in the brain tissue, and so on. Whole genome sequencing technologies can

obtain different types of omics data, such as epigenomic, proteomic, transcriptomic, and genomic data, forming another example of multi-view data. These multi-omics data can be integrated to gain comprehensive knowledge about a human's cancer grades.

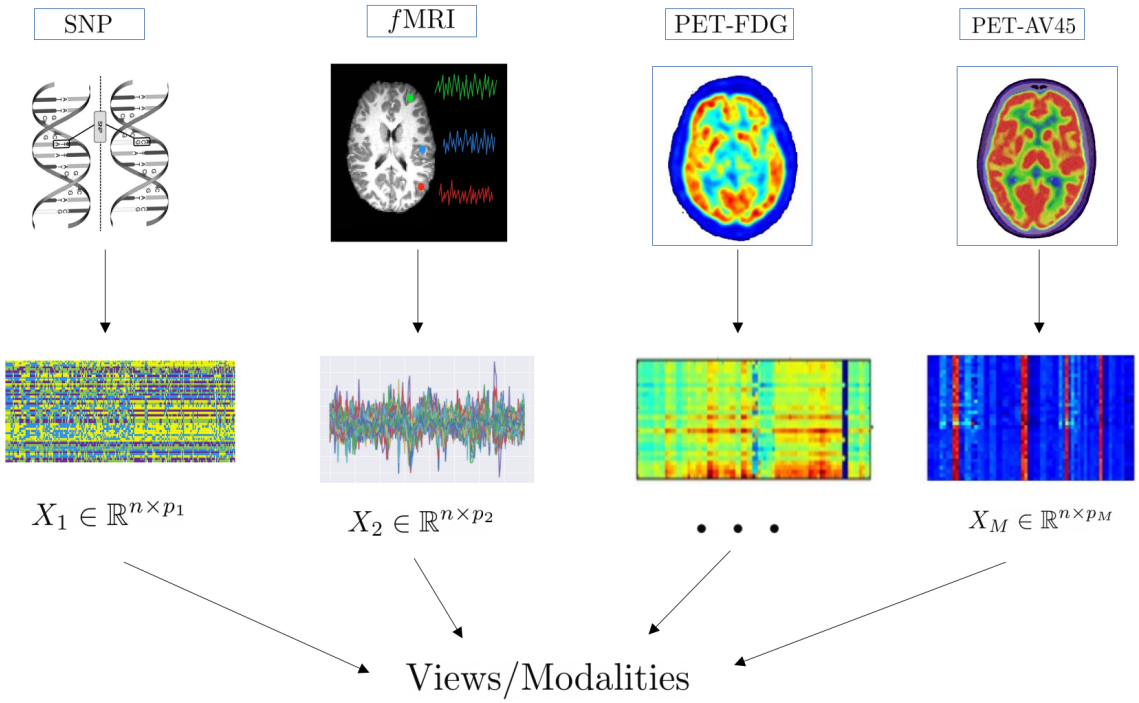


Figure 2.1: Different modalities of imaging genetics data

2.2 Multi-View Learning Approaches

Traditional machine learning methods, such as support vector machines, discriminant analysis, kernel machines, and eigendecomposition based methods, are designed to learn on a single view data. Thus, one simple way, to analyze multiple views, is to concatenate all the views into a single view and feed into the traditional machine learning algorithms. However, this trivial approach intensifies the problem of "curse of dimensionality", since it makes the feature size of the concatenated view very large while the sample size remains the same. Moreover, this approach lacks physical significance due to the distinct statistical properties of each view. Unlike single view learning, multi-view learning is a novel approach which learns one function to model each specific view and simultaneously optimizes all the functions to explore the different statistical properties of the same input data and enhance the learning performance. While the ways of integrating multiple views to enhance learning performance vary widely, they generally rely on either the complementary or the consensus principle to guarantee the success of multi-view learning. Depending on the learning strategy current multi-view learning algorithms can be categorized into four groups, namely, co-training style algorithms, multiple kernel learning, subspace learning, and deep multi-view learning.

2.2.1 Co-training Style Algorithms

Co-training [29], one of the earliest multi-view learning models, was initially developed for the issues of semi-supervised learning, when both labeled and unlabeled data are available. The scenario assumes that each example may be divided into two different points of view and relies on three key premises: (a) Sufficiency: Each individual view is capable of classifying on its own, (b) Compatibility: The target functions in both views accurately anticipate the same labels for features that occur together with a high probability, and (c) Conditional independence: Given the class labels, the views are conditionally independent. To address a wider range of multi-view learning problems, researchers have used the concept of co-training and have created many extended algorithms based on this approach. These include co-expectation-maximization (co-EM) [192], co-testing [185], and co-clustering [287]. Furthermore, significant and essential study was conducted on co-training style algorithms, which contributes to the advancement of co-training methodologies.

The study conducted by Nigam et al. implemented a generalized EM approach, in which unlabeled data was allocated changeable probabilistic labels [192]. Brefeld and Scheffer [32] have effectively devised a co-EM variant of support vector machines. In [183–185], Muslea introduced robust semi-supervised learning algorithms that combine active learning with co-training. The authors Yu and Yu have proposed Bayesian undirected graphical models for co-training [295]. Additionally, they have introduced a new co-training kernel specifically designed for Gaussian process classifiers. In the study of Wang (2010) [267], a method combining graph-based and disagreement-based semi-supervised learning was developed. The co-training process was seen as a combined label propagation over two different perspectives. The paper [226] presents a co-regularization system that uses multi-view regularization to learn classifiers in each view. Several co-training based multi-view clustering techniques have been presented in the papers [24, 140, 141]. Zhao et al. [311] introduced an approach that integrates the simple nature of k-means clustering and linear discriminant analysis in a co-training framework. This framework leverages automatically learned labels from one perspective to acquire discriminative subspaces from another perspectives. Although the success of co-training algorithms can be attributed to three primary factors as discussed above, in practical situations, it is very challenging to fulfill the requirement of conditional independence across perspectives. Therefore, a number of weaker alternatives have been suggested in [4, 18, 268].

2.2.2 Multiple Kernel Learning

The multiple kernel learning (MKL) was first designed to regulate the capacity of the search space for potential kernel matrices in order to achieve optimal performance. It provides a way of integrating various properties of objects, such as genes, proteins, metabolites, etc., by utilizing distinct kernel matrices [230]. The kernels in MKL represent distinct perspectives, and the incorporation of many kernels has the potential to enhance the learning efficacy. In recent years, MKL has emerged as a significant method for analyzing multi-view data sets. The reason for MKL being a popular technique is a result of its use of many optimization strategies [6, 143, 230] and its capacity to recognize patterns by exploring different combinations of basic kernels [134, 257, 281]. Several models have been developed to expand MKL approaches, including localized MKL [90], sample-adaptive MKL [158], Bayesian MKL [62], multiple empirical kernel learning [72, 269], two-stage MKL [187, 266],

and function approximation MKL [139, 223]. The formulation of MKL as a semi-definite programming problem is presented in [143]. In [15], MKL is utilized to create a second-order cone programme issue by formulating a dual version of the quadratically-constrained quadratic programme. To efficiently get the best solution, a sequential minimum optimization technique has been devised. Several effective semi-infinite linear programmes have been suggested in references [230, 231], with MKL being utilized to tackle large-scale data analysis.

2.2.3 Subspace Learning Approaches

The primary aim of subspace learning-based methods is to get a common latent subspace shared by multiple modalities, where each input view can be generated from this latent subspace. The "curse of dimensionality" issue can be overcome by the subspace learning techniques as the common latent space may have very small number of features compared to each input view. The canonical correlation analysis (CCA) [107] is a subspace learning technique, which obtains two linear transformation for two data views such that the correlation between these two data views is maximized in the latent space. In many of the significant scientific domains, including face recognition, text mining [57, 148, 314], imaging genomics [109, 131], integration of omics data [80, 162], facial expression analysis, and brain MRI data analysis [174, 202], CCA has extensively been used. The literature has a number of CCA variations, some of which are covered here.

- **Regularized CCA:** Since real-world data sets may contain very few samples and a high number of attributes, CCA finds it problematic to construct the latent space. Data sets with large number of feature usually have multicollinearity problem, which results in ill-conditioned covariance matrices of different views. This leads to the unreliability of their inverses, which in turn causes an erroneous CCA computation and an unreliable meta-space. The regularized CCA (RCCA) [259] is the extension of CCA, which makes the singular covariance matrix invertible by adding some small positive quantity to the diagonal of the matrix and prevents the overfitting problem. Moreover, RCCA is able to handle many real-life data sets that are often plagued with noise. The supervised RCCA [81] is the supervised extension of RCCA, which learns the optimal regularizer by some statistical test for better classification performance.

- **Penalized CCA:** In penalized CCA, certain penalties are appended to the weight vectors. This will lead to a constrained optimization problem involving CCA. This constraint optimization problem can be solved numerically using iterative optimization approaches because analytical solutions are not available. There are several optimization approaches that may be used, including the reduced gradient method, sequential quadratic programming, Broyden-Fletcher-Goldfarb-Shanno algorithm, and augmented-Lagrangian algorithm. In [31, 194], ℓ_1-norm penalty is added to canonical weight vectors to form sparse version of CCA, which makes the weight vectors sparse to handle "large feature small sample size" problem. To incorporate the prior knowledge of the samples or features such as natural ordering in the features or grouped according to property of the features, sparse CCA can be extended to group sparse CCA or structured sparse CCA [74, 155]. The relationship between genetic, neuropsychological, behavioral or clinical data, and brain imaging data has been established

using different penalized version of CCA in [65, 67, 93, 95, 110, 155]. In [175], a multimodal data fusion model employing sparse and structured CCA has been created to maintain the spatial structure of images.

- **Discriminant CCA:** The learning method of CCA is typically unsupervised and does not rely on class labels of the data. Linear discriminant analysis (LDA) [76] is a supervised subspace learning method for single view learning. It finds an optimal linear transformation to project the data onto a subspace, where the within-class scatter is minimized and between-class scatter is maximized. The multi-view LDA (MLDA) [286], which is based on CCA and LDA, optimizes the sum of correlation between two views and discrimination of each view. It is further extended to uncorrelated LDA (ULDA) [121, 286], where the optimum feature vectors are uncorrelated to each other. The multi-view extension of ULDA, called MULDA [239], extracts uncorrelated features by imposing more constraints on the solutions.

  The discriminative CCA (DCCA) has been introduced in [243] to include the supervised information into CCA. The DCCA finds the subspace by maximizing the within-class similarity and minimizing the between-class similarity. The multi-view common component discriminant analysis (MvCCDA) [292] considers supervised information and local geometry to find the discriminative common space for cross-view classification. However, the MvCCDA ignores the global geometry of the data, and uses an iterative approach to solve the problem.

  The multi-view discriminant analysis (MvDA) [124] forms a common discriminative latent space by maximizing the between-class variance and minimizing the within-class variance, considering both the intra-view and inter-view information across all the views. The MvDA-VC, an extension of MvDA, takes care of the view consistency property. The major drawback of MvDA is that it considers only the intra-view and inter-view discrimination information, which actually gives more emphasize on the complementary information and ignores the consensus information across multiple views.

- **Kernel CCA:** The CCA can only capture the linear relationships between two data views. Thus, if two input views have non-linear relationships, the CCA cannot capture their true relationships. To capture non-linear relationship among multiple views kernelization is a useful technique. Using kernelization technique, the CCA is extended to kernel CCA (KCCA) [7], which maps the input views to a higher dimensional Hilbert space, where the data views are assumed to have linear relationships. Using the kernel trick, LDA has been extended to generalized discriminant analysis (GDA) [20], which is a kernel based non-linear LDA. The MLDA [286], DCCA [243], ULDA [121, 286] and MULDA [239] are extended to their non-linear versions, namely, KMLDA [222], KDCCA [241], KULDA [154] and KMULDA [239], respectively. In [25], a KCCA-based method for computing a multivariate temporal filter that connects the connection between brain activity and functional magnetic resonance imaging is presented as a solution to the dynamic time-delay problem.

- **Locality Preserving CCA:** A locality-based or locality-preserving strategy is a very successful way for addressing data nonlinearity other than kernel tricks. The Locality Preserving Canonical Correlation Analysis (LPCCA) was introduced in [240].

LPCCA incorporates similarity matrices into the CCA algorithm to detect the local structure of the manifold. The fundamental concept behind LPCCA is that data points are considered to be in close proximity in the low-dimensional projected subspace if they are sufficiently close in the high-dimensional input space [125, 240, 252, 262, 285]. Based on the data, LPCCA may be categorized into two distinct categories. One group generates a local neighborhood graph by utilizing cross-correlation information among neighbors [252, 262], while the other group disregards the insignificant correlation between non-neighbors and focuses on providing a local manifold structure [125, 240, 285]. The authors of [285] have built a supervised LPCCA model with the aim of enhancing the classification performance. The inclusion of neighborhood property in [262] enhances the resilience of the model.

- **Multiset CCA:** The generalized CCA, also known as multiset CCA (MCCA) [105, 106], is the multi-view extension of the CCA. It maximizes the sum of pair-wise correlation for multiple views and finds one linear transformation for each view along which the views can be projected to form a common latent space. Based on different optimization criteria, several formulations have been made for MCCA [129,191]. The MCCA models may be categorized into two groups: pairwise-correlation or zero-order-correlation based models and high-order-correlation based models, depending on the concept of cross-view correlation for multi-view learning.

    1. **Pairwise-Correlation or Zero-Order-Correlation Based Models:** There are multiple ways to capture zero-order correlation or pair-wise correlation among multiple views, which creates multiple forms of MCCA. In SUMCOR-MCCA, the sum of all elements in the correlation matrix is maximized, whereas the sum of squares of all elements is maximized in SSQCOR-MCCA. In MAXVAR and MINVAR criterion of MCCA, the largest and smallest eigenvalue of the correlation matrix are maximized and minimized, respectively. The determinant of the correlation matrix is minimized in GENVAR formulation of MCCA. All these criteria and their significance can be found in [129, 191]. Hanafi et al. additionally examined the sum of absolute value correlations (SABSCOR) as an additional criteria [91]. In [53], the sum of covariance (SUMCOV) criteria was put out. The sum of squared covariance (SSQCOV) criteria has also been presented in [92]. In [246], various variants of SUMCOR, SSQCOR, and SABSCOR have been presented, which consider some conjectures regarding the relationships between sets of variables. This article has additionally taken the sum of absolute value covariances (SABSCOV) into consideration. The analyses of SUMCOV, SSQCOV, and SABSCOV are based on covariance between canonical variates, whereas the SUMCOR, MAXVAR, SSQCOR, MINVAR, GENVAR, and SABSCOR are based on maximizing a function of the correlation between canonical variates [247].
    2. **High-Order-Correlation Based Models:** In multi-view data analysis, tensors can be considered as generalizations of matrices, vectors, and scalars. It is employed in order to record higher-order correlations between different points of view [47,275,279]. In [161], the tensor-based generalization of CCA (TCCA) for more than two views was presented. TCCA builds a covariance tensor to esti-

mate the correlation of all views rather than computing the pairwise correlation matrix.

Like CCA, MCCA is unsupervised and cannot handle non-linear relationships among multiple views. Moreover, it has the same issue of ill-conditioned covariance matrix. In [298], Laplacian MCCA (LapMCCA) has been introduced to uncover the non-linear correlation hidden in multiple views, by considering the nearest neighbor graphs of local intra-view correlation and inter-view correlation structure. Although LapMCCA can capture the non-linear correlation, it is unable to use the class label information. To cope with the singularity problem of the covariance matrices regularized multiple-sets canonical correlation analysis (RGCANO) [245] have been proposed, which prevents the over fitting issue by adding some positive numbers to the diagonals of singular covariance matrices. Additionally, MCCA has been expanded into probabilistic models [51, 133, 260], deep learning frameworks [21], and kernel approaches [14, 215].

Apart from the CCA based methods, there are several subspace learning methods based on non-negative matrix factorization (NMF). Recently, a semi-supervised label-driven auto-weighted strategy (LACK) has been proposed in [296], which uses the label of the data to select the important views from a discriminative perspective. However, it only focuses on identifying the redundant views that have negative impact on the performance of the model, without considering the other aspects of multi-view data analysis such as, view consistency and data geometry. Some of the semi-supervised MVL algorithms, based on NMF, have been developed to simultaneously learn feature and sample representations. The multi-view semantic learning (MvSL) [196] focuses on capturing the semantic structure in the multi-view data by exploring the ideas of NMF and graph embedding. It constructs an affinity graph and a penalty graph for characterizing the intra-class compactness and inter-class separability in the common subspace. However, the MvSL only explores the complementary discrimination information of each view and ignores the consistent information across multiple views. In [153], a graph-regularized partially shared non-negative matrix factorization (GPSNMF) is introduced, which obtains a latent representation of each view by preserving the intrinsic geometry of the view via affinity graph construction. The co-consensus orthogonal non-negative matrix factorization (CONMF) has been proposed in [152] to learn not only the consensus information between the samples, but also between the samples and their sub-cluster centroid. To explore the relationship between the samples and their sub-cluster centroid, orthogonal NMF is used. However, all these methods restrict the data decomposition into non-negative weight matrices, as they are primarily based on the NMF of the data.

2.2.4 Deep Multi-View Learning

Deep learning algorithms have gained interest in recent years due to their great feature extraction capacity. Deep learning methods utilize several hierarchical layers to acquire intricate, non-linear, and abstract representations of the target data from various perspectives. There are various deep multi-view learning algorithms documented in the literature, including the multi-view convolutional neural network [165, 289], multi-view auto-encoder [103, 307], multi-view generative adversarial network [60, 251], multi-view graph

neural network [97, 130, 282], multi-view deep belief net [10, 304], and multi-view recurrent neural network [3, 218]. In addition to CCA, additional traditional learning methods are also included into the deep framework, including deep multi-view matrix factorization [308], deep multi-view spectral learning network [114], and deep multi-view information bottleneck [8].

The deep adversarial CCA [73] has been proposed by integrating the CCA with the framework of generative adversarial network (GAN). Although it learns the non-linear data representation by generating realistic multi-view samples through GAN, it does not incorporate the class label information of the samples. The tensor canonical correlation analysis network (TCCANet) [288] maximizes the high-order correlation among multiple views and handles the optimization problem by decomposing a covariance tensor. In [305], Zhang and Sun proposed a multi-view graph restricted Boltzmann machine (mgRBM), which simultaneously learns the view-specific and view-consistent representation while preserving the local manifold structure of the data. In [278], Xie et al. proposed a deep multi-view twin support vector machines (DMvTSVM) for multi-class classification problem. It uses deep neural network and auto-encoder to learn the data non-linearity from multiple views and twin support vector machines to construct the class separating hyperplanes. A semi-supervised multi-view deep discriminant learning (SMDDRL) method has been proposed in [116], which uses two representation learning networks to learn both the shared and complementary properties from different views. In [300], a multi-view graph learning approach has been introduced for semi-supervised classification. It proposes a diversity promoting regularized term to capture the diversity of multiple views, while minimizing the mutual redundancy among them. In effect, it does not consider the consistency present across the multiple views.

2.3 Muti-Task Learning

Suppose there are T number of tasks $\{\mathcal{T}_t\}_{t=1}^{T}$. For each task $\mathcal{T}_t$, there are N input/output examples $\{(x_1^{\mathcal{T}_t}, y_1^{\mathcal{T}_t}),(x_2^{\mathcal{T}_t}, y_2^{\mathcal{T}_t}),\ldots, (x_N^{\mathcal{T}_t}, y_N^{\mathcal{T}_t})\} \in \mathbb{R}^d \times \mathbb{R}$. The objective of multi-task learning (MTL) is to generate T functions $f_t : \mathbb{R}^d \rightarrow \mathbb{R}$, where $t = 1, 2, \ldots, T$, which accurately find the relationship between the input and output variables and have strong statistical prediction power.

Multi-task learning [39] has become increasingly popular in the last ten years because of its strong predictive capabilities. There are two primary methods for understanding the connection between related tasks: utilizing a shared feature representation and utilizing shared model parameters. The approaches for sharing a similar feature representation involve sharing common underlying structures [11, 13, 45], whereas the approaches for sharing model parameters include utilization of the shared prior distributions in hierarchical Bayesian models [17], employing kernel-based techniques with regularization [71, 169], and sharing parameters of a Gaussian process [30].

Multi-task learning has its influence over the domain of imaging genetics, where multiple task-specific features are learned to have a comprehensive understanding of underlying neurodegenerative disease. In [66], a multi-task SCCA (MTSCCA) has been introduced, which builds several SCCA tasks concurrently and associates SNPs with imaging QTs of each modality, in order to study the multimodal imaging genetics problem. The dirty

MTSCCA (DMTSCCA) [64] makes use of parameter decomposition and MTL simultaneously in order to identify not only the shared imaging QTs and genetic loci across multiple modalities but also the modality-specific imaging QTs and genetic loci. However, both MTSCCA and DMTSCCA do not use the diagnostic status of the samples, which may give important insight in finding the complex relationships among SNPs and multimodal imaging QTs. The label-guided MTSCCA (LGMTSCCA) [94] applies parameter decomposition and sparse regression analysis, along with the label information of the data, to obtain modality-consistent and modality-specific weight matrices. The label information is used in sparse regression analysis to retain the relevant and noise-free features. In [63], Du et al. proposed a multi-task bi-multivariate approach, termed as multi-task SCCA and logistic regression (MT-SCCALR), which selects diagnose-specific features to find genotype-phenotype patterns particular to each diagnosis. It uses MTSCCA and multi-task logistic regression simultaneously to select a set of relevant features for each diagnostic group. In [128], multi-task SCCA and regression (MT-SCCAR) has been introduced, which uses genetic and multimodal neuroimaging data, along with cognitive measures, to identify the high-risk brain regions and genetic risk factors associated with Alzheimer disease.

Apart from the imaging genetics study multi-task learning has also been used in other fields. According to the different tasks in machine learning, there exist multi-task unsupervised learning [13, 306], multi-task semi-supervised learning [157], multi-task reinforcement learning [147], multi-task active learning [5, 210] and multi-task online learning [55].

2.4 Multi-Task Multi-View Learning

Multi-task multi-view learning is an extension of multi-view learning that aims to enhance the performance of each multi-view learning problem by using important information from related tasks. Multi-task multi-view learning [98] addresses the challenge of learning numerous interconnected activities that share one or more common perspectives. The graph-based approach described in [98] effectively utilizes the diverse features and activities by mapping them to a Reproducing Kernel Hilbert Space using shared perspectives. The generic inductive learning framework described in [303] employs co-regularization and task connection learning to enhance the applicability of multi-task multi-view learning. The approach proposed in [119] enables the learning of shared predictive structures from numerous related activities. It leverages the consistency among diverse viewpoints to enhance performance. The system described in [104] utilizes many views and various types of visual characteristics to perform multi-task multi-view sparse learning. It also takes into account the link between tasks across different views and particles. The approach, called multi-task multi-view discriminant analysis, proposed in [120], addresses the issue of multi-task multi-view learning for heterogeneous tasks.

2.5 Conclusion

This chapter provides an overview of the fundamental concepts and principles of multi-view data analysis under both single-task and multi-task frameworks. Because of its good performance across several application domains, it quickly gained attention in machine learning research and has generated a substantial amount of literature over the last decade.

In the current chapter, a brief literature survey, primarily focused on the numerous feature extraction techniques of multi-view learning, has been presented. It also covers some of the literature on multi-task learning in the domain of imaging genetics study.

One major challenge in analyzing multi-view data is to determine the common discriminative latent space by productively using the inter-view correlation and intra-view discrimination information, where the local as well as global class-structure of the data can be preserved. In this context, the upcoming chapter presents a new supervised multi-view learning framework by judiciously combining the concepts of MCCA, LDA, and a locality preserving norm.

Chapter 3

Class-Structure Preserving Multi-View Correlated Discriminant Analysis

3.1 Introduction

Technological advancement made it possible to analyze an object using multiple views or sets of features. Each such set captures a different aspect of the objects under consideration. Instead of analyzing the objects using a single view, multiple views together may give complementary as well as consensus information across all the views. Multi-view learning (MVL), a machine learning framework for data integration from multiple views, is growing rapidly over the last decade [238, 310]. Presently, many MVL models, for instance, multi-view representation learning [116, 277], multi-view clustering [150, 168], multi-view classification [118, 305], multi-view graph learning [301], multi-view manifold learning [113, 190], and multi-view active learning [294, 309], have been introduced to deal with the particular machine learning tasks. The aim of MVL is to build a function for each view and then optimize all the functions jointly to enhance the generalization performance. A primary solution to MVL is to concatenate all the views to form a single view and then apply the single view algorithms. However, this would intensify the problem of "curse of dimensionality" [213].

To get rid of the dimensionality issue, the subspace learning approaches can be used. Subspace learning [58, 107], a popular technique in MVL, assumes that the multiple views can be generated from a single common latent space where the shared and complementary information from all the views can be explored. Canonical correlation analysis (CCA) [107], a popular statistical multivariate method, is a typical subspace learning approach. It finds a common latent space where two views are maximally correlated. The generalized CCA, also known as multiset CCA (MCCA) [105, 106], is the multi-view extension of the CCA. It maximizes the sum of pair-wise correlation for multiple views and finds one linear transformation for each view along which the views can be projected to form a common latent space. Based on different optimization criteria, several formulations have been made for MCCA [129, 191]. The CCA or MCCA has been used extensively in MVL, including multi-view classification [89, 239], multi-view regression [123], multi-view clustering [27, 42] and so on. However, classical CCA or MCCA only considers the linear relationships among

multiple views, and does not perform well for cross-view classification as it does not consider the class label information present in the data.

To deal with these limitations, many extensions of CCA and MCCA have been proposed in past few decades. Two unsupervised methods, namely, kernel CCA (KCCA) [7] and locality preserving CCA (LPCCA) [242], have been proposed to overcome the limitation of non-linearity. In [298], Laplacian MCCA (LapMCCA) has been introduced to uncover the non-linear correlation hidden in multiple views, by considering the nearest neighbor graphs of local intra-view correlation and inter-view correlation structure. Although LapMCCA can capture the non-linear correlation, it is unable to use the class label information. Another unsupervised method, called correlative covariation projection (CCP), has been proposed in [297], which projects the features rather than the samples on a new space by introducing a new canonical $\mathcal{F}$-correlation framework to explore the non-linearity in features. One of the important issues with the CCA or MCCA is the singularity problem of the covariance matrices. The real world data having small number of samples and large number of features makes the covariance matrices ill-conditional. To cope with the singularity problem of the covariance matrices, regularized CCA (RCCA) [259] and regularized multiple-set canonical correlation analysis (RGCANO) [245] have been proposed, which prevent the over fitting issue by adding some positive numbers to the diagonal of singular covariance matrices.

For cross-view classification and to improve the classification performance of RCCA, several supervised extensions of CCA and MCCA have also been introduced in [81, 163]. However, most of them do not use the supervised information to learn the subspace; rather they learn the subspace by CCA and then use some statistical tests to include supervised information [81], which is not much effective towards classification. The supervised feature extraction algorithm, termed as ReDMiCA, has been introduced based on MCCA [164]. It uses the supervised information to find optimal regularization parameters. Unlike these methods, the generalized multi-view analysis (GMA) [222] framework has been proposed, where the class labels are utilized to form a common discriminative latent space. In GMA, the intra-view discrimination information and pair-wise correlation among multiple views have been considered. The multi-view discriminant analysis (MvDA) [124] forms a common discriminative latent space by maximizing the between-class variance and minimizing the within-class variance, considering both the intra-view and inter-view information across all the views. The MvDA-VC, an extension of MvDA, takes care of the view consistency property. The major drawback of MvDA is that it considers only the intra-view and inter-view discrimination information, which actually gives more emphasize on the complementary information and ignores the consensus information across multiple views. Another supervised extension of MCCA is discriminative MCCA (DMCCA) [78], which incorporates the supervised information by maximizing the within-class correlation and minimizing the between-class correlation. The convex discriminant semantic correlation analysis [250] analyzes the sample representation space by considering both the correlation and cross-view semantic information. In [19], a model has been introduced, based on a supervised general covariance matrix. The model explores the class labels to construct a discriminating space and captures feature non-linearity as well. However, the MVL methods described above are either unsupervised in nature or capture linear relationships among multiple views. Most of them do not consider the geometry of the classes present in the data to learn the discriminative subspace. They focus either only on the complementary and discriminatory

information in each view or on the shared correlation information across multiple views, which make them vulnerable to the data having both the information.

In this regard, a new subspace learning approach, termed as class-structure preserving multi-view correlated discriminant analysis (CSP-MvCDA), is proposed in this chapter by judiciously integrating the merits of MCCA and linear discriminant analysis (LDA). The CSP-MvCDA considers equal importance to the consensus and complementary information present in different views. The consensus information across all the views is captured through the correlation structure, while the complementary information is captured from each view using within-class scatter and between-class scatter. Both the scatter matrices help to incorporate the discrimination information into the projection vectors. Thus, the projected space constructed by the proposed approach has optimum correlation and discrimination, which best represents the multiple views. Moreover, the CSP-MvCDA uses a locality preserving norm guided by the labels of the data, which can handle the non-linear structures of the classes present in the data. The idea, to include the label guided locality preserving norms, is to bring the samples belonging to the same class closer, while the samples belonging to different classes are farther apart in the projected space. In this way, the proposed model extracts the best possible set of features, which are discriminative and consistent across all the views and preserve the class geometry present in the data. The proposed method has only two parameters, which make the model computationally less complex with respect to parameter tuning. Based on geometrical separability index, a method is introduced to find out the optimum values of parameters of the proposed method. Moreover, a computational complexity analysis of the proposed as well as several existing MVL algorithms is presented.

Some of the novel characteristics of the proposed approach with respect to existing MVL approaches are highlighted below:

- The proposed method is built on the concept of singular value decomposition (SVD) of a block matrix. The block matrix consists of the cross-covariance matrices normalized by the within-class scatter of the data, which brings the supervised information into the model to preserve the global class geometry. However, most of the existing models are based on either non-negative matrix factorization (NMF) of the data matrices or SVD of the cross-covariance matrices, which constructs an unsupervised subspace. Hence, different regularizers are used to make these models supervised.

- A label guided regularizer term, based on the locality preserving projection, is introduced to the proposed model to preserve the local class geometry of the data in each view. On the other hand, the global class geometry of the data is preserved by jointly optimizing the ISC and ISD ratio. In effect, both local and global class geometry of the data are preserved.

- It is shown that the formulation of CSP-MvCDA has a closed-form solution to the problem, which most of the existing models lack. This formulation can be solved easily by solving a generalized eigenvalue problem.

- The proposed CSP-MvCDA model is shown to be simplest and most effective among the popular existing supervised MVL techniques, namely, GMA, MULDA, and MvDA, since it has only two parameters to be optimized.

The extensive experimental results on several cancer and benchmark data sets establish that the proposed CSP-MvCDA model performs significantly better than the state-of-the-art approaches, as it has the capability in extracting class discriminative features more accurately than the existing methods. The results on large data sets signify the scalability of the proposed method with respect to number of features and classes. Some of the results of this chapter can be found in [176].

The remaining sections in this chapter are arranged as follows: Section 3.2 gives the preliminary ideas in MCCA and LDA. Section 3.3 elaborates the proposed algorithm. An analytical solution is given to obtained a closed form solution based on the generalized eigenvalue problem. The effectiveness of the proposed algorithm is shown by comparing the performance of it with the other state-of-the-art algorithms over several omics and benchmark data sets in Section 3.4. The chapter is concluded in Section 3.5.

3.2 Basics of MCCA and LDA

In this section, the basic concepts of both multiset canonical correlation analysis (MCCA) and linear discriminant analysis (LDA) are presented.

Notation

Uppercase letters and lowercase letters denote the matrices and column vectors, respectively. The operators $\mathrm{Tr}(\cdot)$, $(\cdot)^{-1}$ and $(\cdot)^T$ denote the trace, inverse and transpose of a matrix, respectively; $\|\cdot\|_F$ is the Frobenius-norm of a matrix; $\|\cdot\|_2$ is the l_2-norm of a vector; A/a represents variable in the optimization problem whereas A^*/a^* is the critical value of the variables; $\mathbf{1}$ is the vector consists of all ones; I_n is identity matrix of order n.

3.2.1 Multiset Canonical Correlation Analysis

The MCCA [106] is a generalization of the CCA. It can analyze the linear relationship among multiple sets of data (more than two). There exist several models of the MCCA, depending on different criteria and constraints. The sum of correlation (SUMCOR) criterion of MCCA, given here, is taken from [195], which maximizes the inter-set correlation (ISC) between multiple views. Suppose, there are m sets of data $\{X_i | X_i \in \mathbb{R}^{n\times p_i}\}$, where n is the number of common samples and p_i is the number of features in the data of the i-th view X_i, $i = 1, 2, \ldots, m$. The ISC is defined by the ratio of between set covariance r_B and the within set covariance r_W, as follows:

$$\rho = \frac{1}{m-1}\left(\frac{r_B}{r_W}\right). \tag{3.1}$$

The objective of MCCA is to find m sets of projection vectors, $\{v_i\}_{i=1}^m$ for m data blocks, such that the ISC between m data blocks in the projected space is maximized. The between

set covariance (r_B) and within set covariance (r_W) in the projected space are defined as

$$r_B = \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} v_i^T \Sigma_{ij} v_j, \tag{3.2}$$

$$r_W = \sum_{i=1}^{m} v_i^T \Sigma_{ii} v_i, \tag{3.3}$$

where $\Sigma_{ij} = X_i^T X_j$ $(i \neq j)$ is the between set covariance and $\Sigma_{ii} = X_i^T X_i$ is the within set covariance of the data blocks in the original space. By substituting the values of r_B and r_W in (3.1), the ISC in the projected space, ρ_{proj}, can be obtained as follows:

$$\rho_{proj} = \frac{1}{m-1} \left(\frac{\sum_{i=1}^{m} \sum_{j=1, j \neq i}^{m} v_i^T \Sigma_{ij} v_j}{\sum_{i=1}^{m} v_i^T \Sigma_{ii} v_i} \right). \tag{3.4}$$

The objective function for this formulation of MCCA can be written as

$$\begin{aligned} \max \quad & \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} v_i^T \Sigma_{ij} v_j \\ \text{subject to} \quad & \sum_{i=1}^{m} v_i^T \Sigma_{ii} v_i = 1. \end{aligned} \tag{3.5}$$

The solution of the optimization problem of (3.5) can be obtained using Lagrange's multiplier method and is given by the eigenvectors $\mathbf{v}$ of the generalized eigenvalue problem:

$$\mathbf{Rv} = \lambda \mathbf{Dv}; \tag{3.6}$$

where $\mathbf{R} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \dots & \Sigma_{1m} \\ \Sigma_{21} & \Sigma_{22} & \dots & \Sigma_{2m} \\ \vdots & \vdots & \dots & \vdots \\ \Sigma_{m1} & \Sigma_{m2} & \dots & \Sigma_{mm} \end{bmatrix}$, $\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{bmatrix}$, and $\mathbf{D} = diag[\Sigma_{11}, \Sigma_{22}, \dots, \Sigma_{mm}]$.

3.2.2 Linear Discriminant Analysis

The LDA [76] is a popular supervised technique for single view learning. It utilizes the class label information of a given data set to find a projective space, where different classes or groups are maximally separated. Let $X = \{x_1, x_2, x_3, \dots, x_n\}$ be a set of n data points, where each $x_i \in \mathbb{R}^p$ and p is the dimension or number of features for each sample x_i. The LDA finds an optimal direction vector w, for which the between-class scatter of the data, $w^T S_b w$, is maximized and the within-class scatter of the data, $w^T S_w w$, is minimized in the projected space, where S_b and S_w are the between-class scatter and within-class scatter in

the original space. The ratio

$$\frac{w^T S_b w}{w^T S_w w} \tag{3.7}$$

is known as Fisher's criterion, which is maximized to obtained the optimal direction vector w as follows:

$$\max_w \frac{w^T S_b w}{w^T S_w w}. \tag{3.8}$$

The scatter matrices, S_b and S_w, in the original space, can be constructed using the class label of the samples, and are defined as

$$S_b = \sum_c n_c(\mu_c - \bar{x})(\mu_c - \bar{x})^T; \tag{3.9}$$

$$S_w = \sum_c \sum_{i \in c} (x_i - \mu_c)(x_i - \mu_c)^T; \tag{3.10}$$

where $\bar{x}$ is the mean over all the samples, μ_c is the mean of the samples belonging to the c-th class and n_c is the number of samples present in class c.

The objective function of the LDA can also be expressed in terms of the total scatter $S_t (= S_b + S_w)$ of the data as follows:

$$\max_w \frac{w^T S_t w}{w^T S_w w}. \tag{3.11}$$

The formulation proposed in Section 3.3 uses the above form.

3.3 Class-Structure Preserving Multi-View Correlated Discriminant Analysis

The MCCA is unsupervised in nature and does not use the available class label information present in the data. In this section, a new multi-view supervised method, named class-structure preserving multi-view correlated discriminant analysis (CSP-MvCDA), is introduced. The proposed method not only considers pair-wise correlation to incorporate the consensus information across all the views, but also uses class-label information to preserve the global geometry of the classes present in each view. It also preserves the local structure of each class in the projected space by using a class-structure preserving norm.

3.3.1 Formulation of CSP-MvCDA Model

Let there be m data sources $\{X_i \mid X_i \in \mathbb{R}^{n \times p_i}\}_{i=1}^m$, corresponding to m views of an object, where n is the number of samples and p_i is the number of features in the i-th view X_i. Assume that there are K pattern classes $\{c_1, c_2, \ldots, c_K\}$. Consider m projection vectors $\omega_1, \omega_2, \ldots, \omega_m$ for the data views $X_1, X_2, \ldots, X_m$, such that the new features in the

projected space are given by $z_i = X_i\omega_i$, $i = 1, 2, \ldots, m$.

3.3.1.1 Consensus Information through Correlation Structure

The consensus information from all the views can be captured through the ISC of the m views. From (3.4), the ISC of the m views in the projected space ρ_{proj} is given by

$$\rho_{proj} = \frac{1}{m-1}\left(\frac{\sum_{i=1}^{m}\sum_{j=1,j\neq i}^{m}\omega_i^T\Sigma_{ij}\omega_j}{\sum_{i=1}^{m}\omega_i^T\Sigma_{ii}\omega_i}\right). \tag{3.12}$$

3.3.1.2 Global Class-Structure through LDA

The LDA [76] preserves the global class-structure of the data in the projected space by using the within-class scatter and between-class scatter. Correspondingly, the global class-structure of each view X_i in the transformed space can be obtained by applying the LDA. Let S_t^i and S_w^i be the total scatter and within-class scatter in the i-th view X_i of the original space. Define intra-set discrimination (ISD) ratio D_{proj} as the ratio between the sum of the total scatter and the sum of the within-class scatter across all the views in the projected space, as follows:

$$D_{proj} = \frac{\sum_{i=1}^{m}\omega_i^T S_t^i\omega_i}{\sum_{i=1}^{m}\omega_i^T S_w^i\omega_i}. \tag{3.13}$$

Maximizing the ISD ratio (D_{proj}) in the projected space would minimize the within-class scatter and maximize the between-class scatter of each view. It helps to find a direction vector that separates the classes maximally by considering the global class-structure of each view.

The main objective of the proposed formulation is to optimize jointly the ISC (ρ_{proj}) and ISD ratio (D_{proj}), so that the transformed space contains discriminative information from each view as well as the correlation information across all the views. Hence, the objective function of the proposed method maximizes the product of ρ_{proj} and D_{proj}, which is given by

$$\rho_{proj}D_{proj} = \frac{\sum_{i=1}^{m}\sum_{j=1,j\neq i}^{m}\omega_i^T\Sigma_{ij}\omega_j}{\sum_{i=1}^{m}\omega_i^T\Sigma_{ii}\omega_i}\frac{\sum_{i=1}^{m}\omega_i^T S_t^i\omega_i}{\sum_{i=1}^{m}\omega_i^T S_w^i\omega_i}. \tag{3.14}$$

Since the total scatter (S_t^i) of each view X_i is equal to the total variance Σ_{ii} of that view, therefore

$$\sum_{i=1}^{m}\omega_i^T\Sigma_{ii}\omega_i = \sum_{i=1}^{m}\omega_i^T S_t^i\omega_i. \tag{3.15}$$

So, the objective function of (3.14) can be simplified using (3.15) as follows:

$$\rho_{proj} D_{proj} = \frac{\sum_{i=1}^{m} \sum_{j=1, j\neq i}^{m} \omega_i^T \Sigma_{ij} \omega_j}{\sum_{i=1}^{m} \omega_i^T S_w^i \omega_i}. \tag{3.16}$$

3.3.1.3 Local Class-Structure through Locality Preserving Norm

The discriminative information through global class-structure may not always be useful for multi-class discrimination. There may be some local structures of the multiple classes in the data set that may be omitted when the global class-structure is only considered. Therefore, the projective space must be modified in such a way that the samples belonging to the same class come closer in the projective space and also maintain the global structure. This can be achieved by using a local class-structure preserving norm [284], inspired by the locality preserving projection, which is given by

$$\|\omega\|_{LCP} = \alpha \omega^T X^T L_w X \omega - (1-\alpha) \omega^T X^T L_b X \omega, \tag{3.17}$$

where α is a trade-off parameter. The parameter α is used to balance between the within-class and between-class discrimination information through two graph Laplacians L_w and L_b of the graphs G_w and G_b, respectively, constructed from the class labels of the samples.

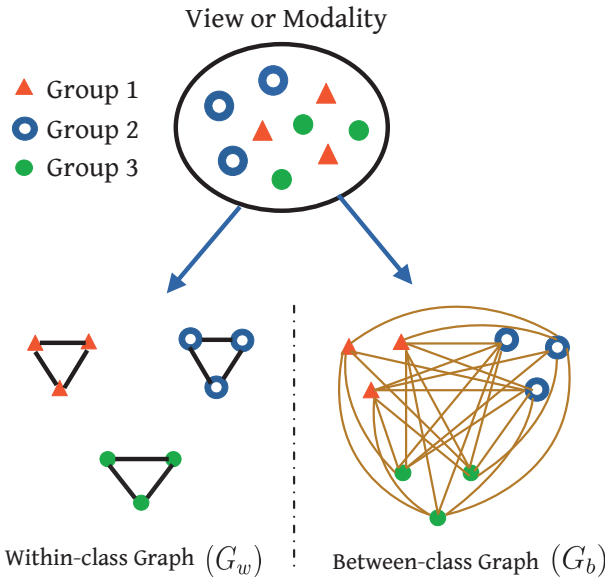


Figure 3.1: Construction of within-class and between-class graph.

Using the prior class label information, two graphs, namely, G_w and G_b, are first constructed, where a vertex corresponds to a sample. In the graph G_w, the vertices those belong to the same class are connected, whereas in the graph G_b, the vertices belonging to different classes are connected. So, the $n \times n$ adjacency matrices A_w and A_b, corresponding

to the graphs G_w and G_b, respectively, are defined as follows:

$$A_w(p,q) = \begin{cases} 1, & \text{if the class labels of samples } x_p \text{ and } x_q \text{ are same} \\ 0, & \text{otherwise} \end{cases} \tag{3.18}$$

$$A_b(p,q) = \begin{cases} 1, & \text{if the class labels of samples } x_p \text{ and } x_q \text{ are different} \\ 0, & \text{otherwise.} \end{cases} \tag{3.19}$$

The construction of these two graphs is shown in Fig. 3.1. Following the definition of graph Laplacian, $L_b = D_b - A_b$ and $L_w = D_w - A_w$ are the graph Laplacians of two graphs G_b and G_w, respectively. Here, D_b and D_w are the degree matrices of the graphs G_b and G_w, respectively.

The norm defined in (3.17) has the importance in increasing the closeness of the samples belonging to the same class and distance between the samples belonging to different classes in the projected space. Taking the norm of (3.17) into consideration, the objective function of (3.16) can be written as

$$\begin{aligned} &\max_{\omega_1,\omega_2,\ldots,\omega_m} \frac{\sum_{i=1}^{m}\sum_{j=1,j\neq i}^{m} \omega_i^T \Sigma_{ij} \omega_j}{\sum_{i=1}^{m} \omega_i^T S_w^i \omega_i}, \\ &\text{subject to} \quad \sum_{i=1}^{m} \|\omega_i\|_{LCP} \leqslant \kappa, \end{aligned} \tag{3.20}$$

where κ is some constant and

$$\|\omega_i\|_{LCP} = \alpha \omega_i^T X_i^T L_w X_i \omega_i - (1-\alpha) \omega_i^T X_i^T L_b X_i \omega_i. \tag{3.21}$$

The objective function (3.20) can further be expressed as

$$\begin{aligned} &\max_{\omega_1,\omega_2,\ldots,\omega_m} \sum_{i=1}^{m}\sum_{\substack{j=1 \\ j\neq i}}^{m} \omega_i^T \Sigma_{ij} \omega_j, \\ &\text{subject to} \quad \sum_{i=1}^{m} \|\omega_i\|_{LCP} \leqslant \kappa; \ \sum_{i=1}^{m} \omega_i^T S_w^i \omega_i = 1. \end{aligned} \tag{3.22}$$

For simplicity, adding the first constraint of (3.22) as a regularizer term into the objective function, the final objective function can be written as follows:

$$\begin{aligned} &\max_{\omega_1,\omega_2,\ldots,\omega_m} \sum_{i=1}^{m}\sum_{\substack{j=1 \\ j\neq i}}^{m} \omega_i^T \Sigma_{ij} \omega_j - \frac{\gamma}{2} \sum_{i=1}^{m} \|\omega_i\|_{LCP}, \\ &\text{subject to} \quad \sum_{i=1}^{m} \omega_i^T S_w^i \omega_i = 1, \end{aligned} \tag{3.23}$$

where γ is some constant, and is interdependent with κ. The proposed CSP-MvCDA method is pictorially represented in Fig. 3.2, along with MCCA and LDA. It also gives a brief overview of the construction of correlated discriminative subspace.

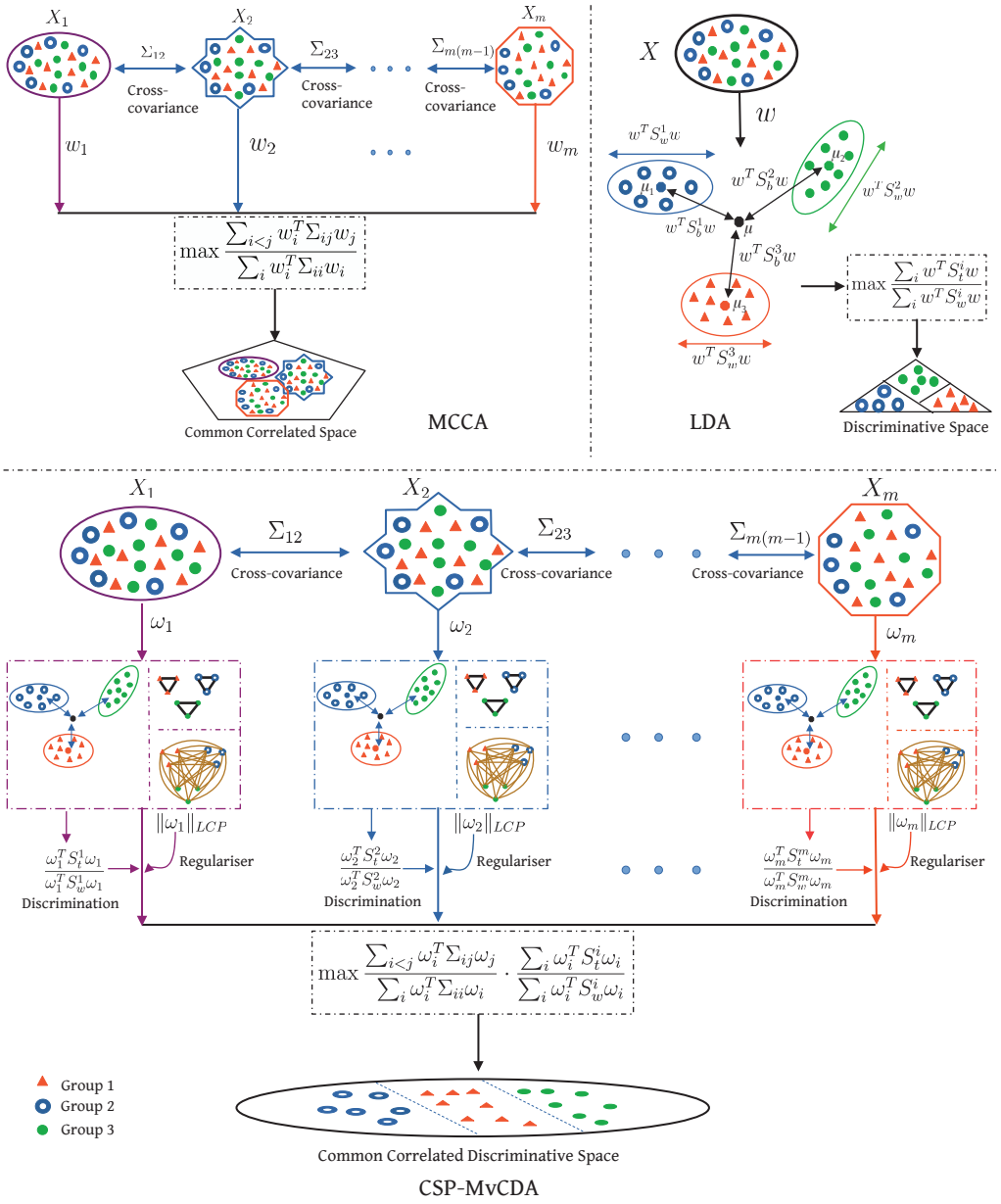


Figure 3.2: Pictorial representation of MCCA, LDA, and proposed CSP-MvCDA.

3.3.2 Analytical Solution

To solve the constraint optimization problem of (3.23), the following Lagrangian function is defined:

$$F(\omega_1, \omega_2, \ldots, \omega_m) = \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} \omega_i^T \Sigma_{ij} \omega_j - \frac{\gamma}{2} \sum_{i=1}^{m} \|\omega_i\|_{LCP} - \lambda \left(\sum_{i=1}^{m} \omega_i^T S_w^i \omega_i - 1 \right), \quad (3.24)$$

where λ is the Lagrange's multiplier. For the maximization of $F(\omega_1, \omega_2, \ldots, \omega_m)$,

$$\frac{\partial F(\omega_1, \omega_2, \ldots, \omega_m)}{\partial \omega_i} = 0$$

$$\Rightarrow \quad \frac{\partial}{\partial \omega_i} \left(\sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} \omega_i^T \Sigma_{ij} \omega_j \right) - \frac{\gamma}{2} \frac{\partial}{\partial \omega_i} \left(\sum_{i=1}^{m} \|\omega_i\|_{LCP} \right) - \lambda \frac{\partial}{\partial \omega_i} \left(\sum_{i=1}^{m} \omega_i^T S_w^i \omega_i - 1 \right) = 0$$

$$\Rightarrow \quad \frac{\partial}{\partial \omega_i} \left(\sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} \omega_i^T \Sigma_{ij} \omega_j \right) - \frac{\gamma}{2} \frac{\partial}{\partial \omega_i} \left(\sum_{i=1}^{m} \left(\alpha \omega_i^T X_i^T L_w X_i \omega_i - (1-\alpha) \omega_i^T X_i^T L_b X_i \omega_i \right) \right) - \lambda \frac{\partial}{\partial \omega_i} \left(\sum_{i=1}^{m} \omega_i^T S_w^i \omega_i - 1 \right) = 0$$

$$\Rightarrow \quad \sum_{\substack{j=1 \\ j \neq i}}^{m} \Sigma_{ij} \omega_j - \gamma \left(\alpha X_i^T L_w X_i \omega_i - (1-\alpha) X_i^T L_b X_i \omega_i \right) - \lambda S_w^i \omega_i = 0$$

$$\Rightarrow \quad \sum_{\substack{j=1 \\ j \neq i}}^{m} \Sigma_{ij} \omega_j = \lambda S_w^i \omega_i + \gamma L_i \omega_i, \qquad i = 1, 2, \ldots, m \quad (3.25)$$

$$\text{where} \quad L_i = \alpha X_i^T L_w X_i - (1-\alpha) X_i^T L_b X_i. \quad (3.26)$$

The set of equations in (3.25) can be written in the following form using the block matrix notation:

$$\mathbf{A}\Omega = \lambda \mathbf{S}\Omega + \gamma \mathbf{L}\Omega, \quad (3.27)$$

$$\text{where } \mathbf{A} = \begin{bmatrix} 0 & \Sigma_{12} & \dots & \Sigma_{1m} \\ \Sigma_{21} & 0 & \dots & \Sigma_{2m} \\ \vdots & \vdots & \dots & \vdots \\ \Sigma_{m1} & \Sigma_{m2} & \dots & 0 \end{bmatrix}, \ \Omega = \begin{bmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_m \end{bmatrix}, \tag{3.28}$$

$$\mathbf{S} = \begin{bmatrix} S_w^1 & 0 & \dots & 0 \\ 0 & S_w^2 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & S_w^m \end{bmatrix}, \text{and } \mathbf{L} = \begin{bmatrix} L_1 & 0 & \dots & 0 \\ 0 & L_2 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & L_m \end{bmatrix}. \tag{3.29}$$

The relation of (3.27) can also be expressed in a generalized eigenvalue problem as follows:

$$(\mathbf{A} - \gamma\mathbf{L})\,\Omega = \lambda\mathbf{S}\Omega. \tag{3.30}$$

The solution $\Omega = [\omega_1, \omega_2, \dots, \omega_m]$ of the generalized eigenvalue problem of (3.30) gives m projection vectors $\omega_1, \omega_2, \dots, \omega_m$ for m views X_1, X_2, ..., X_m. Suppose, the top d eigenvectors, Ω_1, Ω_2, ..., Ω_d are obtained from (3.30), where $\Omega_i = [\omega_1^i, \omega_2^i, \dots, \omega_m^i]$, $i = 1, 2, \dots, d$. Let W_j be the transformation matrix for the j-th view X_j, formed by the top d projection vector $\omega_j^1, \omega_j^2, \dots, \omega_j^d$, $j = 1, 2, \dots, m$. The data sets in the projected space become $Z_1 = X_1W_1$, $Z_2 = X_2W_2$, ..., $Z_m = X_mW_m$. These projected sample sets are sum up to construct the new feature set in the common discriminative latent space. This new feature set is then used for the downstream classification tasks. Each step of the proposed method is explained using the flowchart of Fig. 3.3, while the algorithmic steps are presented in Algorithm 3.1.

Algorithm 3.1 Algorithm for the Optimization Problem of (3.23)

1: **INPUT:** $X_1, X_2, X_3, \dots, X_m$, class labels of samples.
2: **OUTPUT:** $\Omega = [\omega_1, \omega_2, \dots, \omega_m]^T$
3: Compute the cross-covariance matrix $\Sigma_{ij} = X_i^T X_j (i \neq j)$ for each pair of views.
4: Compute the block matrix $\mathbf{A}$ using (3.28).
5: Compute the within-class scatter matrix $S_w^i = \sum_{c=1}^{K} X_i^{cT} X_i^c$ for each view.
6: Construct the within-class graph G_w and between-class graph G_b using the class labels of the samples.
7: Compute the graph Laplacians L_w and L_b, corresponding to G_w and G_b.
8: Compute the matrix L_i for each view using (3.26).
9: Compute the block diagonal matrices $\mathbf{S}$ and $\mathbf{L}$ using (3.29).
10: Solve the generalized eigenvalue problem of (3.30).

3.3.3 Comparison with Other MVL Techniques

In this section, the merits of the proposed model are discussed with respect to three multi-view data integration techniques, namely, MCCA, GMA and MvDA.

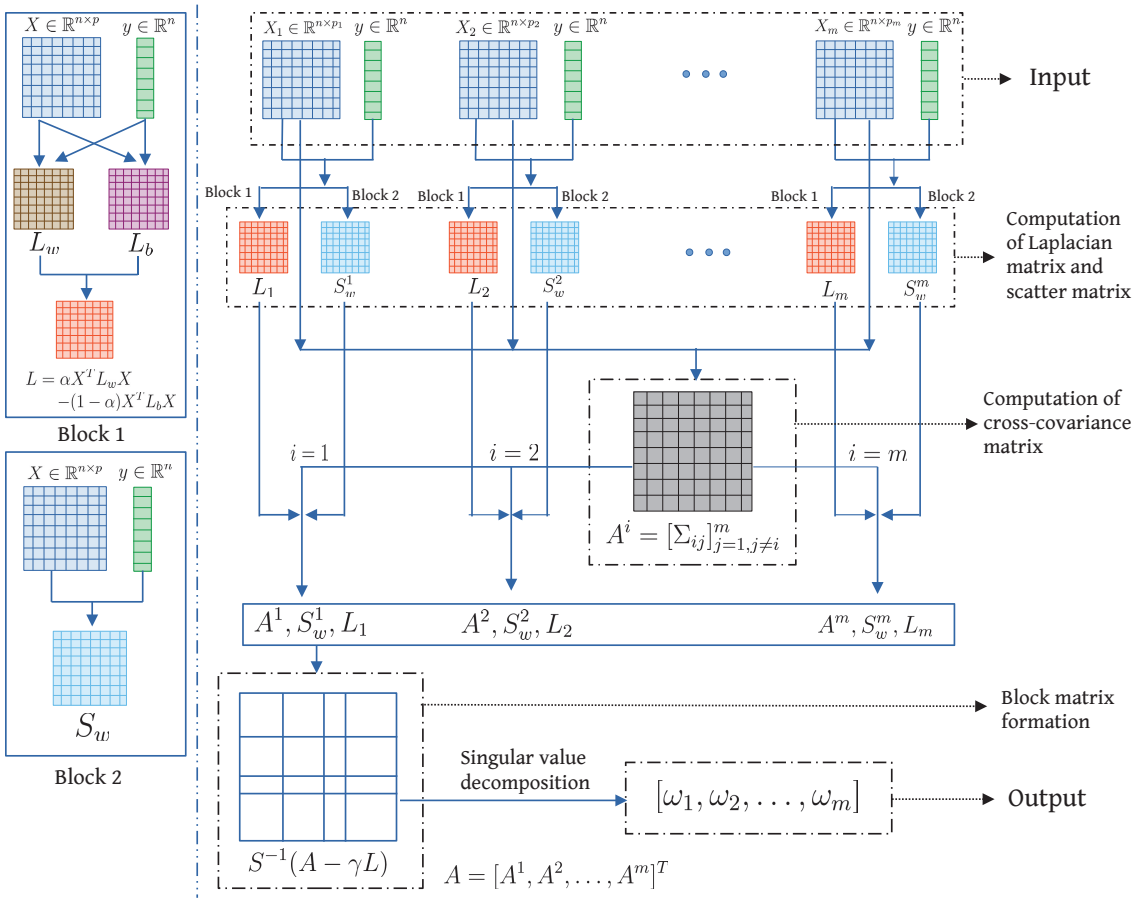


Figure 3.3: Flowchart of the proposed CSP-MvCDA method.

3.3.3.1 CSP-MvCDA versus Multiset CCA (MCCA)

Both MCCA [105] and CSP-MvCDA are the extended forms of the two-view method CCA and belong to the category of subspace learning technique. The basic difference between MCCA and CSP-MvCDA is that CSP-MvCDA utilizes the class label information of the data, whereas the MCCA cannot. The objective functions of MCCA and CSP-MvCDA differ in their constraints. The objective function of the MCCA is as follows:

$$\max_{\substack{\omega_1,\omega_2,\ldots,\omega_m \\ \sum_1^m \omega_i^T \Sigma_{ii} \omega_i = 1}} \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} \omega_i^T \Sigma_{ij} \omega_j, \tag{3.31}$$

whereas the objective function of the CSP-MvCDA is as follows:

$$\max_{\substack{\omega_1,\omega_2,\ldots,\omega_m \\ \sum_1^m \omega_i^T S_w^i \omega_i = 1 \\ \sum_1^m \|\omega_i\|_{LCP} \leqslant k}} \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} \omega_i^T \Sigma_{ij} \omega_j. \tag{3.32}$$

So, in the constraints of the proposed method, the sum of class-wise variance, $S_w^i = \sum_{c=1}^{K} X_i^{cT} X_i^c$, replaces the total variance $\Sigma_{ii} = X_i^T X_i$ in the constraint of MCCA. Moreover, there is a restriction on the projection vectors, $\{\omega_i\}_{i=1}^{M}$, in the proposed model, which makes the model cope with the non-linear class-structure. The constraints in CSP-MvCDA play an important role to include the discriminative information into the projection vector by moving the direction of the projection vector so that the class-wise variance of each data is minimized and preserve the local class-structure as well. The proposed method can handle the data non-linearity and is expected to be more significant towards the construction of discriminative latent space unlike MCCA.

3.3.3.2 CSP-MvCDA versus Generalized Multi-View Analysis (GMA)

The GMA [222], a general multi-view feature extraction method, is intuitively a supervised extension of CCA. It learns a single linear subspace by optimizing a relaxed quadratic constrained quadratic program (QCQP) over different feature spaces. The objective function of the GMA is given by

$$\begin{aligned} &\max_{w_1,w_2,\ldots,w_m} \sum_{i=1}^{m} \mu_i w_i^T A_i w_i + \sum_{i<j} 2\lambda_{ij} w_i X_i^T X_j w_j, \\ &\text{subject to} \quad \sum_{i=1}^{m} \gamma_i w_i^T B_i w_i = 1 \end{aligned} \tag{3.33}$$

where $\mu_i(\mu_1 = 1), \lambda_{ij}$ and $\gamma_i(\gamma_1 = 1)$ are the balance parameters, both A_i and B_i are square symmetric matrices. In supervised CCA version of GMA, $A_i = S_b^i$ and $B_i = S_w^i$ are the between-class and within-class scatter matrices for the i-th view.

The objective function of the GMA maximizes the sum of between-class scatter of each view and cross-covariance across multiple views to achieve the common latent space. Though the GMA is a more general method in MVL, the formulation of CSP-MvCDA is much simpler than the GMA with respect to formation of the common discriminative latent space. Also, the CSP-MvCDA has a closed form solution and is capable of preserving the non-linear class-structure of each view, whereas the GMA is unable to preserve the non-linearity of each view. Besides, the objective function of the GMA includes $\left[\frac{m(m+3)}{2} - 2\right]$ number of parameters for m views, and thus it becomes computationally very expensive. On the other hand, the proposed CSP-MvCDA method has only two parameters α and γ to be tuned. So, the proposed method has less computational burden for parameter tuning.

3.3.3.3 CSP-MvCDA versus Multi-View Discriminant Analysis (MvDA)

The MvDA [124] seeks a discriminant common subspace from multiple views and can be considered as a multi-view extension of LDA. Let $\{x_{ij}^v | i = 1, 2, \ldots, K; j = 1, 2, \ldots, n_i\}$ be the j-th sample of the i-th class in v-th view, where n_i is the number of samples in the i-class and $v = 1, 2, \ldots, m$. Let y_{ij}^v be the projected samples, which are obtained by projecting the original samples x_{ij}^v through the m view-specific linear transformations $w_1, w_2, \ldots, w_m$, that is, $y_{ij}^v = w_v^T x_{ij}^v$. The objective function of the MvDA can be written

as

$$\max_{w_1,w_2,\ldots,w_m} \frac{trace(\mathbf{S}_B^y)}{trace(\mathbf{S}_W^y)}; \tag{3.34}$$

where the between-class scatter matrix $\mathbf{S}_B^y$ and within-class scatter matrix $\mathbf{S}_W^y$ of the projected samples are defined by

$$\mathbf{S}_W^y = \sum_{v=1}^{m}\sum_{i=1}^{K}\sum_{k=1}^{n_i}(y_{ij}^v - \mu_i)(y_{ij}^v - \mu_i)^T \tag{3.35}$$

$$\mathbf{S}_B^y = \sum_{i=1}^{K} n_i(\mu_i - \mu)(\mu_i - \mu)^T. \tag{3.36}$$

The MvDA takes the inter-view and intra-view discriminative information into account and ignores the shared correlated information across multiple views, which might loose some important information and degrade the performance. However, the proposed CSP-MvCDA method considers both the shared and complementary information through ISC and ISD ratio, respectively, which makes the model more consistent towards multi-view data integration. While the subspace learned by the MvDA preserves the global class geometry of multiple views, the proposed model has the ability to preserve both the local and global class geometry from each view in the projected space.

3.3.4 Computational Complexity Analysis

In this section, the computational complexity of the proposed method is analyzed, assuming $p = \max_i\{p_i\}$, where $p_i\,(i = 1, 2, \ldots m)$ is the dimension of the i-th view X_i. The time complexity for computing the cross-covariance matrix $\Sigma_{ij}\,(i \neq j)$, $(i, j = 1, 2, \ldots, m)$ in Step 3 of Algorithm 6.1 is $\mathcal{O}(m^2p^2n)$. Similarly, the time complexity to compute the class covariance matrix $\{S_w^i\}_{i=1}^m$ is also $\mathcal{O}(m^2p^2n)$. The computational complexity to construct the within-class graph $\{G_w^i\}_{i=1}^m$ and between-class graph $\{G_b^i\}_{i=1}^m$ is $\mathcal{O}(mn^2)$. To compute their graph Laplacian L_w^i and L_b^i, the time complexity is also $\mathcal{O}(mn^2)$. The time complexity to solve the generalized eigenvalue problem of (3.30) is $\mathcal{O}(m^3p^3)$. So, the total computational complexity of the proposed CSP-MvCDA method is $\mathcal{O}(m^2p^2n + m^2p^2n + mn^2 + mn^2 + m^3p^3) \approx \mathcal{O}(m^2p^2n + mn^2 + m^3p^3)$. In case of "large p small n" problem, $p >> n$ implies $m^3p^3 >> mn^2$, and the computational complexity of the proposed method becomes $\mathcal{O}(m^2p^2n + m^3p^3)$.

In Table 3.1, the computational complexity of different state-of-the-art methods and the proposed CSP-MvCDA is compared. Here, d represents the number of extracted features in the projected space, c is number of classes, and T is the number of iterations required to converge MvCCDA [292], LACK [296], GPSNMF [153] and CONMF [152]. Parameters K_c and K_s denote the number of common latent factors and view specific latent factors, respectively, for GPSNMF; while T_1 and T_2 are the number of iterations required for the subproblems I and II, respectively, in the t-th iteration of MvCCDA. The parameter τ is the total number of possible regularization parameters in ReDMiCA [164].

Table 3.1: Complexity of Proposed and Existing Algorithms

Different Methods	Computational Complexity
MCCA [106]	$\mathcal{O}(m^2p^2n + m^3p^3)$
GMA [222]	$\mathcal{O}(m^2p^2n + m^3p^3)$
MvDA [124]	$\mathcal{O}(m^2p^2(n + c) + m^3p^3)$
MULDA [239]	$\mathcal{O}(dp^3)$
MvCCDA [292]	$\mathcal{O}(T((mndp + m^2p)T_1 + mnp^2T_2))$
ReDMiCA [164]	$\mathcal{O}(d\tau p^3)$
LACK [296]	$\mathcal{O}(T(m(c^3 + c^2p + pnc)))$
GPSNMF [153]	$\mathcal{O}(T((K_c + mK_s)^3 + (K_c + mK_s)^2(n^2 + pn)))$
CONMF [152]	$\mathcal{O}(T(m(n^2c + c^2p)))$
CSP-MvCDA	$\mathcal{O}(m^2p^2n + mn^2 + m^3p^3)$

3.4 Experimental Results and Discussion

In this section, the performance of the proposed CSP-MvCDA method is compared comprehensively with that of several existing multi-view learning methods, namely, MCCA [106], GMA [222], MvDA [124], MvDA-VC [124], MLDA [286], MULDA [239], MvCCDA [292], ReDMiCA [164], LACK [296], GPSNMF [153] and CONMF [152], and two deep learning based multi-view classification methods, namely, DMvTSVM-AE [278] and SMDDRL [116]. Three real life cancer data sets and five benchmark data sets are used for the comparative performance analysis in the current study. The cancer data sets are breast invasive carcinoma (BRCA), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) and lower grade glioma (LGG), which are collected from The Cancer Genome Atlas (TCGA)[1]. The benchmark data sets are Cora[2], CiteSeer[3], Nus-Wide-Object[4](NW-Object), Caltech-101[5] and ALOI[6]. The statistical information of these data sets is given in Table 3.2 and their detailed description is given in Appendix A. The source code of the proposed CSP-MvCDA algorithm, written in Python, is available at https://github.com/sankarML/CSP-MvCDA.

Both training-testing and 10-fold cross-validation (CV) are used to study the effectiveness of different algorithms. In training-testing, 50% samples are used for training and the rest is used for testing. For all the methods, top 25 features are extracted, except for MvCCDA, LACK, CONMF and two deep learning methods DMvTSVM-AE and SMDDRL. The methods MvCCDA, LACK, and CONMF have the capability of generating at most c numbers of features, where c is the number of classes present in the data. For these five methods, the optimal number of features is used to report the classification accuracy. To compute the classification accuracy, support vector machine (SVM) with linear kernels is used. To establish the performance of the proposed method statistically, two significance tests, namely, paired sample t-test and Wilcoxon's signed rank test, are performed, and the corresponding p-value is reported.

[1] https://cancergenome.nih.gov/
[2] http://lig-membres.imag.fr/grimal/data.html
[3] http://lig-membres.imag.fr/grimal/data.html
[4] https://lms.comp.nus.edu.sg/wp-content/uploads/2019/research/nuswide/NUS-WIDE.html
[5] http://www.vision.caltech.edu/datasets/
[6] https://elki-project.github.io/datasets/multi_view

Table 3.2: Descriptions of Data Sets Used

Data Sets	Sample	Views/Modalities	Class
BRCA	134	RNA Sequence (12268) Protein expression (216) miRNA Sequence (178) Copy number segmentation (2258)	4
CESC	104	RNA Sequence (12028) Protein expression (192) miRNA Sequence (174) DNA methylation (291368)	3
LGG	374	RNA Sequence (11973) Protein expression (181) miRNA Sequence (139) DNA methylation (293965) Copy number segmentation (6261)	3
CiteSeer	3312	Content (3703) Inbound (3312) Outbound (3312) Citation (3312)	6
Cora	2708	Content (1433) Inbound (2708) Outbound (2708) Citation (2708)	7
NW-Object	30000	Color histogram (64) Block-wise color moments (225) Color correlogram (144) Edge direction histogram (73) wavelet texture (128)	31
Caltech-101	9144	Gabor (48) Wavelet moments (40) Cenhist (254) Hog (1984) Gist (512) Local binary patterns (LBP) 928	102
ALOI	10800	RGB color histograms (64) HSB color histograms (125) Color similarity (77) Haralick features (13)	100

3.4.1 Selection of Optimal Parameters

The proposed CSP-MvCDA model has two parameters, namely, α and γ, to be tuned. The parameter α controls the balance between within-class and between-class discriminative information in class-structure preserving norm, while the parameter γ represents the contribution of the class-structure preserving norm into the performance of the model.

To find out the optimal values of α and γ parameters, the geometrical separability index (GSI) [249] is used. The GSI, also known as the Thornton index, calculates the average number of samples that share the same class label as those of their first nearest neighbor. The GSI is mathematically defined as follows:

$$T = \frac{1}{n}\sum_{i=1}^{n} f(x_i, x_i') \tag{3.37}$$

where x_i' denotes the first nearest neighbor of the sample x_i, n is the total number of samples and f is a binary function, defined as follows:

$$f(x_i, x_i') = \begin{cases} 1 & \text{if label of } x_i = \text{label of } x_i' \\ 0 & \text{if label of } x_i \neq \text{label of } x_i'. \end{cases} \tag{3.38}$$

The value of T belongs to the closed interval $[0, 1]$. For the set of samples, where the individuals with opposite labels exist in well-separated groups, the value of T will be closer to 1. The index value drops as the groupings get closer together and samples from different classes start to geometrically overlap. Lastly, this separability index will be near 0.5 if the centroids coincide or the samples are evenly distributed throughout the space. As a result, the best approximated sample grouping will be represented by the index with the highest value.

The optimal parameters α^* and γ^* are obtained using a grid search approach using GSI. The value of α is varied from 0.0 to 1.0 having a step size of 0.1 and the value of γ is taken from the set {0.0001, 0.001, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}. The values of α and γ for which the GSI is maximum are considered as the optimal parameters α^* and γ^* of the proposed CSP-MvCDA model, that is,

$$\{\alpha^*, \gamma^*\} = \arg \max_{\{\alpha,\gamma\}} \{GSI(\alpha, \gamma)\}, \tag{3.39}$$

where $GSI(\alpha, \gamma)$ denotes the GSI value for the parameter set $\{\alpha, \gamma\}$. The optimal parameters of the proposed method are listed in Table 3.3 for different data sets. To establish the effectiveness of the proposed parameter optimization technique, the classification accuracy of CSP-MvCDA on test samples is reported in Table 3.3 considering optimal parameters and parameters for which training accuracy is maximum (Max. Training). The best possible test accuracy (Best) is also presented for comparison, which is obtained by searching the parameter space exhaustively. From the results reported in Table 3.3, it can be seen that the CSP-MvCDA attains the highest accuracy in 13 cases out of total 16 cases, considering optimal parameters. Also, the performance for optimal parameters is better than that of Max. Training in 2 out of remaining 3 cases. All the results reported here establish the effectiveness of the proposed parameter optimization technique. The variation of classification accuracy and GSI, with respect to two parameters γ and α, is presented in the supplementary material.

Table 3.3: Parameter Analysis of CSP-MvCDA on Different Data Sets

	Different Data Sets	Parameters	Classification Accuracy on Test Data		
		(γ^*, α^*)	Optimal Parameters	Max. Training	Best
10-fold CV	BRCA	(0.4, 0.8)	0.8325	0.8313	**0.8375**
	CESC	(0.2, 0.5)	**0.9416**	**0.9416**	**0.9416**
	LGG	(0.0001, 0.1)	**0.9842**	**0.9842**	**0.9842**
	CiteSeer	(0.01, 0.8)	**0.6891**	0.6874	**0.6891**
	Cora	(0.01, 0.7)	0.7847	**0.7949**	**0.7949**
	NW-Object	(0.001, 0.1)	**0.4450**	**0.4450**	**0.4450**
	Caltech-101	(0.0001, 0.1)	**0.8251**	0.8234	**0.8251**
	ALOI	(0.0001, 0.1)	**0.9926**	**0.9926**	**0.9926**
Train-Test	BRCA	(0.1, 0.2)	**0.8030**	**0.8030**	**0.8030**
	CESC	(0.8, 0.1)	**0.8654**	**0.8654**	**0.8654**
	LGG	(0.01, 0.3)	**0.9839**	**0.9839**	**0.9839**
	CiteSeer	(0.01, 0.0)	**0.6975**	**0.6975**	**0.6975**
	Cora	(0.2, 0.5)	0.7780	0.7747	**0.7795**
	NW-Object	(0.001, 0.0)	**0.4395**	0.4375	**0.4395**
	Caltech-101	(0.001, 0.8)	**0.6100**	**0.6100**	**0.6100**
	ALOI	(0.001, 0.4)	**0.9815**	**0.9815**	**0.9815**

3.4.2 Importance of Class-Structure Preserving Norm

In order to establish the importance of class-structure preserving norm used in the proposed CSP-MvCDA method, the performance of the proposed method with $\gamma \neq 0$ and $\gamma = 0$ is studied on all the eight data sets. The corresponding results are reported in Table 3.4 with respect to classification accuracy on test data. The results corresponding to $\gamma \neq 0$ imply the performance of the proposed method for optimal γ. All the results reported here confirm that the performance of the CSP-MvCDA with $\gamma \neq 0$ is significantly better than that of $\gamma = 0$, irrespective of the data sets and experimental set-up used. This signifies the necessity of including the class-structure preserving norm into the proposed model.

Table 3.4: Classification Accuracy of CSP-MvCDA on Different Data

Different Data Sets	$\gamma = 0$		$\gamma \neq 0$	
	Train-Test	10-fold CV	Train-Test	10-fold CV
BRCA	0.606	0.744	**0.803**	**0.833**
CESC	0.712	0.683	**0.865**	**0.942**
LGG	0.973	0.984	**0.984**	**0.984**
CiteSeer	0.662	0.661	**0.698**	**0.689**
Cora	0.740	0.760	**0.778**	**0.785**
NW-Object	0.406	0.407	**0.440**	**0.445**
Caltech-101	0.568	0.713	**0.610**	**0.825**
ALOI	0.976	0.983	**0.981**	**0.993**

3.4.3 Superiority of CSP-MvCDA over MCCA

In order to establish the superiority of CSP-MvCDA over MCCA, these two methods are extensively studied and evaluated on the same eight data sets, namely, BRCA, CESC, LGG, Cora, CiteSeer, NW-Object, Caltech-101 and ALOI. The first and last columns of the scatter plots presented in Fig. 3.4 give insight into the quality of top two features extracted by CSP-MvCDA and MCCA. In case of LGG, the CSP-MvCDA is able to isolate the three classes almost accurately, and for CiteSeer and Caltech-101, it shows a discriminating nature, whereas MCCA has no pattern of differentiating classes for all these data sets. This is because the MCCA finds correlated features that does not help in class discrimination most of the time.

In Fig. 3.5, the variation of classification accuracy with respect to the number of extracted features is reported, where deep green and blue curves correspond to MCCA and CSP-MvCDA, respectively. The graph signifies that the extracted features of CSP-MvCDA contain much more discriminative information than the features of the MCCA. The scatter plots and feature versus accuracy graphs for the other data sets are given in the supplementary material.

Finally, Table 3.5 and Table 3.6 presents the classification accuracy of both training-testing and 10-fold CV. Statistical significance test is also performed for 10-fold CV and corresponding p-values are reported in Table 3.5 and Table 3.6. The results show that the CSP-MvCDA achieves much higher classification accuracy than MCCA for each experimental set-up over all eight data sets and all the p-values are significant for 10-fold CV with respect to the MCCA. All these results establish that the CSP-MvCDA can be considered as the supervised alternative of MCCA and more suitable for multi-class classification problem unlike MCCA.

3.4.4 Comparative Performance Analysis

This section compares the performance of the proposed method with that of several state-of-the-art approaches.

3.4.4.1 Classification Results on Omics Data Sets

The performance of the proposed CSP-MvCDA model is compared with that of several existing algorithms on omics data sets. From the scatter plots presented in Fig. 3.4, it is seen that only the proposed method can distinguish between the three classes of LGG data set. For this data set, only two methods, namely, ReDMiCA and LACK are able to isolate one class from the other two. All the other methods are unable to identify the three classes of LGG. In brief, the top two extracted features of the proposed method contains significantly more discriminative information than the other algorithms. To study the quality of the features extracted by the proposed method as well as existing algorithms, the variation of classification accuracy with respect to the number of extracted features is shown in Fig. 3.5. The graphs in Fig. 3.5 show that the proposed method provides best performance for LGG. It establishes that the quality of the extracted features by the proposed method is better than that of the existing algorithms.

The classification accuracy for both the training-testing and 10-fold CV is provided in Table 3.5. The mean, median and standard deviation of classification accuracy are also re-

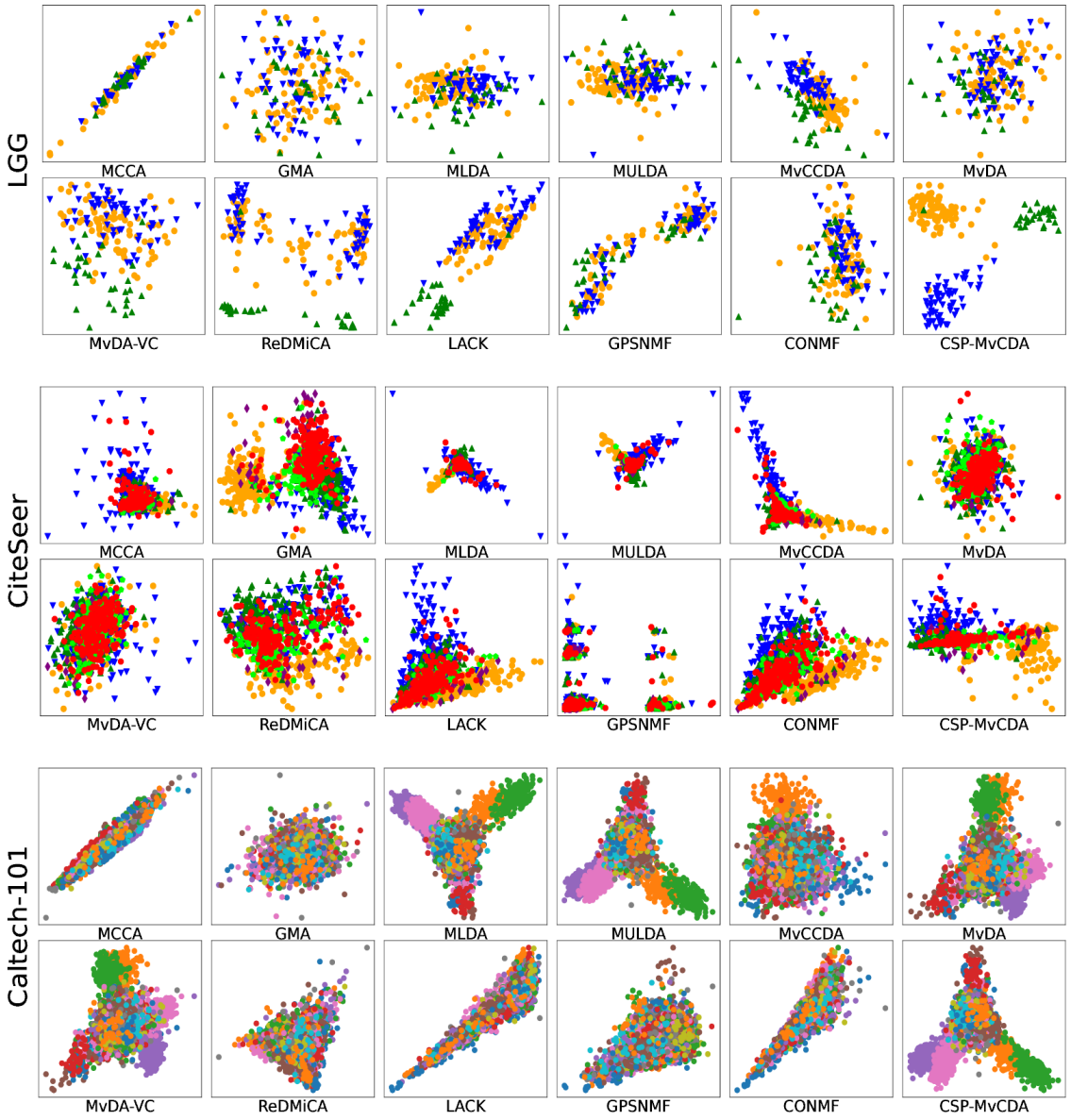


Figure 3.4: Scatter plots for the proposed and existing methods based on the first two extracted features.

ported in case of 10-fold CV, and the p-values corresponding to paired-t test and Wilcoxon signed rank test are presented to prove the statistical significance of these results. All the results reported in Table 3.5 show that, in case of training-testing, the classification accuracy of the proposed model corresponding to BRCA and CESC data sets is significantly higher than that of the existing algorithms. This is because the proposed method integrates the data by taking the combination of inter-set correlation and intra-set discrimination, rather than considering only the inter-set or intra-set discrimination which is considered in the MvCCDA, MvDA, and MvDA-VC. Taking only the intra-set or inter-set discriminative information may include unnecessary information that may lead to degradation in

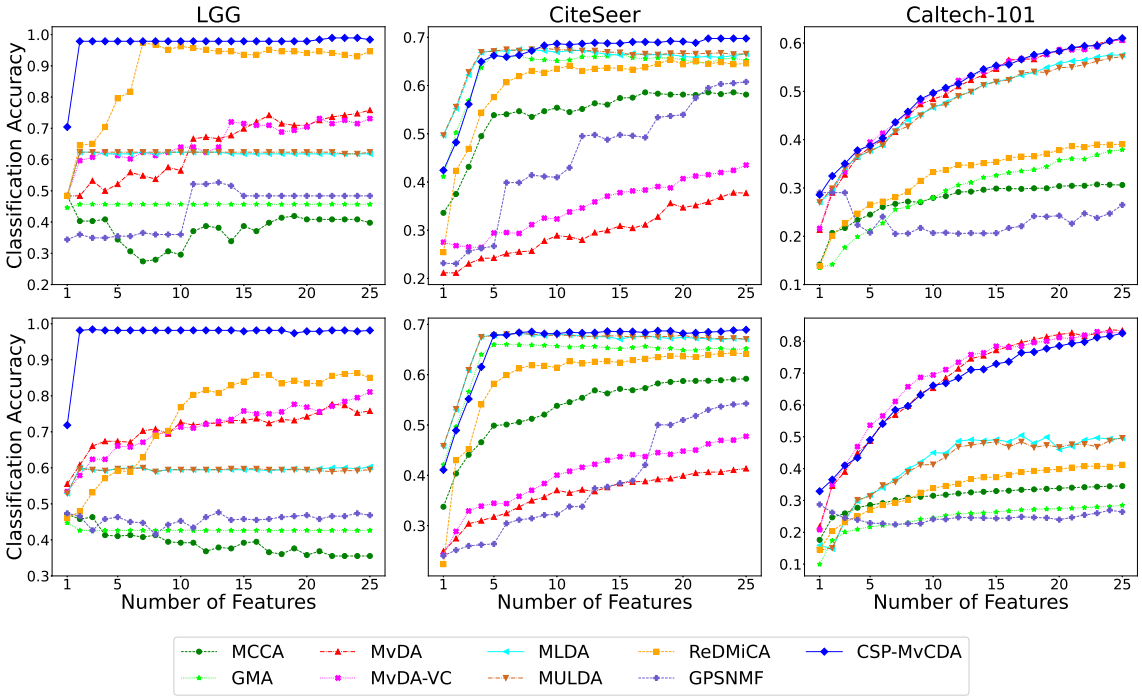


Figure 3.5: Variation of classification accuracy with respect to number of extracted features for the proposed CSP-MvCDA and existing methods on omics and benchmark data sets (top-row: training-testing, bottom-row: 10-fold CV).

the performance. Other than CSP-MvCDA, some of the existing models, namely, MLDA, MULDA, and MvCCDA, provide better classification accuracy on CESC, which is comparatively better among all the other methods. In case of LGG data set, the proposed model also attains the highest accuracy of 98.39% in case of training-testing. Similarly, for 10-fold CV, the proposed model has obtained highest classification accuracy with highest mean and median statistics over all the omics data sets compared to the existing algorithms. Other than the proposed method, the classification accuracy of MvCCDA is also noticeable in case of CESC data set. The deep learning based method, namely, SMDDRL achieves a descent classification score on LGG data set and another deep learning model, namely, DMvTSVM-AE performs well in case of BRCA and CESC. The ReDMiCA also attains better classification accuracy in case of LGG data set. From Table 3.5, it is also clear that all the p-values are significant, justifying that the proposed method has significantly better performance than all other existing methods on three omics data sets. This concludes that the proposed method has the ability to classify omics data sets better than any other state-of-the-art algorithms.

3.4.4.2 Classification Results on Benchmark Data Sets

The scatter plots in Fig. 3.4 show that the proposed method has significant class discrimination ability in case of CiteSeer and Caltech-101 data sets using its top two extracted features. The variation of the classification accuracy with respect to the number of extracted features of the proposed model, shown in Fig. 3.5, suggests that the performance

Table 3.5: Classification Accuracy of Proposed Method (CSP-MvCDA) and Different Algorithms on Omics and Benchmark Data Sets

Different Algorithms	Data Sets	Accuracy (Train-Test)	Accuracy and Significance Analysis for 10-fold CV					Time (Sec.)
			Mean	Median	StdDev	Paired-*t*:p	Wilcoxon:p	
MCCA [106]	BRCA	0.288	0.306	0.319	0.055	**2.61E-10**	**2.21E-03**	31.2
GMA [222]		0.424	0.538	0.531	0.109	**2.44E-05**	**3.91E-03**	33.0
MvDA [124]		0.318	0.431	0.375	0.100	**2.31E-10**	**1.95E-03**	14.1
MvDA-VC [124]		0.303	0.400	0.400	0.079	**2.53E-10**	**1.95E-03**	15.2
MLDA [286]		0.561	0.394	0.406	0.134	**3.49E-05**	**1.95E-03**	73.3
MULDA [239]		0.561	0.381	0.406	0.123	**2.11E-05**	**1.95E-03**	76.5
MvCCDA [292]		0.455	0.569	0.625	0.095	**4.91E-05**	**1.95E-03**	0.9
ReDMiCA [164]		0.424	0.413	0.400	0.119	**2.91E-10**	**1.95E-03**	5774.5
DMvTSVM-AE [278]		0.692	0.615	0.615	0.843	**2.21E-10**	**1.95E-03**	147.7
SMDDRL [116]		0.373	0.474	0.481	0.086	**2.29E-10**	**1.95E-03**	1.8
LACK [296]		0.439	0.444	0.438	0.110	**1.22E-05**	**1.95E-03**	0.1
GPSNMF [153]		0.409	0.388	0.375	0.047	**1.98E-07**	**1.95E-03**	0.8
CONMF [152]		0.591	0.725	0.719	0.098	**1.63E-02**	**2.69E-02**	0.2
CSP-MvCDA		**0.803**	**0.833**	**0.822**	0.081	-	-	5.2
MCCA [106]	CESC	0.385	0.458	0.458	0.137	**1.97E-10**	**1.95E-03**	6.0
GMA [222]		0.462	0.558	0.625	0.124	**9.93E-06**	**1.93E-03**	10.4
MvDA [124]		0.423	0.467	0.492	0.153	**2.05E-10**	**1.95E-03**	15.5
MvDA-VC [124]		0.404	0.500	0.467	0.142	**2.08E-10**	**1.95E-03**	21.9
MLDA [286]		0.769	0.483	0.417	0.170	**1.91E-04**	**7.26E-03**	51.4
MULDA [239]		0.769	0.483	0.417	0.170	**1.91E-04**	**7.26E-03**	50.7
MvCCDA [292]		0.712	0.833	0.833	0.091	**1.32E-02**	**2.40E-02**	0.4
ReDMiCA [164]		0.500	0.567	0.608	0.203	**1.92E-10**	**1.95E-03**	5162.4
DMvTSVM-AE [278]		0.650	0.730	0.700	0.078	**1.32E-02**	**2.40E-02**	147.4
SMDDRL [116]		0.423	0.390	0.381	0.079	**1.97E-10**	**1.95E-03**	1.8
LACK [296]		0.577	0.625	0.667	0.113	**1.55E-04**	**1.95E-03**	0.1
GPSNMF [153]		0.250	0.258	0.250	0.025	**2.81E-09**	**1.95E-03**	0.6
CONMF [152]		0.596	0.533	0.500	0.130	**1.55E-05**	**1.95E-03**	0.2
CSP-MvCDA		**0.865**	**0.942**	**1.000**	0.092	-	-	4.9
MCCA [106]	LGG	0.398	0.355	0.342	0.077	**2.17E-18**	**1.95E-03**	14.2
GMA [222]		0.457	0.426	0.447	0.082	**5.35E-09**	**1.95E-03**	135.3
MvDA [124]		0.758	0.758	0.763	0.080	**1.99E-18**	**1.95E-03**	16.4
MvDA-VC [124]		0.731	0.811	0.789	0.078	**1.99E-18**	**1.95E-03**	16.9
MLDA [286]		0.618	0.603	0.605	0.093	**5.62E-07**	**1.95E-03**	1171.3
MULDA [239]		0.624	0.595	0.592	0.090	**3.21E-07**	**1.95E-03**	1183.6
MvCCDA [292]		0.694	0.766	0.776	0.057	**1.83E-06**	**1.95E-03**	0.7
ReDMiCA [164]		0.946	0.850	0.842	0.035	**1.63E-18**	**1.95E-03**	5958.2
DMvTSVM-AE [278]		0.743	0.746	0.757	0.052	**1.99E-18**	**1.95E-03**	143.4
SMDDRL [116]		0.861	0.877	0.900	0.055	**1.63E-18**	**1.95E-03**	4.3
LACK [296]		0.731	0.945	0.947	0.018	**5.28E-04**	**7.11E-03**	0.2
GPSNMF [153]		0.484	0.476	0.474	0.078	**6.03E-09**	**1.95E-03**	1.4
CONMF [152]		0.532	0.568	0.539	0.100	**5.45E-07**	**1.95E-03**	0.8
CSP-MvCDA		**0.984**	**0.984**	**0.974**	0.012	-	-	8.8
MCCA [106]	ALOI	0.608	0.736	0.742	0.021	**2.41E-15**	**1.95E-03**	0.1
GMA [222]		0.931	0.941	0.940	0.005	**1.82E-07**	**1.95E-03**	0.2
MvDA [124]		0.971	0.986	0.986	0.004	**1.01E-02**	**1.95E-03**	0.5
MvDA-VC [124]		0.981	0.992	0.992	0.003	**2.89E-02**	*6.44E-02*	4.9
MLDA [286]		0.935	0.957	0.958	0.009	**2.76E-06**	**1.95E-03**	0.3
MULDA [239]		0.906	0.926	0.914	0.007	**1.66E-14**	**1.95E-03**	0.3
MvCCDA [292]		0.977	0.986	0.985	0.002	**1.73E-02**	**1.95E-03**	141.6
ReDMiCA [164]		0.941	0.949	0.950	0.011	**4.54E-05**	**1.95E-03**	436.6
DMvTSVM-AE [278]		0.808	0.871	0.881	0.008	**8.61E-05**	**1.95E-03**	133.6
SMDDRL [116]		0.820	0.838	0.839	0.006	**2.16E-08**	**1.95E-03**	60.6
LACK [296]		0.972	0.982	0.983	0.004	**2.52E-05**	**1.95E-03**	2.2
GPSNMF [153]		0.235	0.252	0.250	0.020	**2.31E-15**	**1.95E-03**	9.1
CONMF [152]		**0.981**	0.985	0.985	0.003	**7.39E-05**	**1.95E-03**	6.5
CSP-MvCDA		**0.981**	**0.993**	**0.992**	0.003	-	-	11.9

of the proposed method on benchmark data sets is better than that of the other existing algorithms. The results of Fig. 3.5 show that all the methods, except MvDA and MvDA-VC, provide good performance in case of CiteSeer data set, which indicates the importance of including the correlation information across all the views. The scatter plots and feature versus accuracy graphs for the other benchmark data sets, namely, Cora, NW-Object and ALOI, are given in the supplementary material. In brief, the graphs and scatter plots on benchmark data sets establish that the quality of the features extracted by the proposed method is much better than that of other existing algorithms.

Finally, the results of Table 3.5 and Table 3.6 on benchmark data sets show that the proposed model obtains the highest classification accuracy in case of training-testing as well as 10-fold CV except for Caltech-101. Although MvDA method achieves highest classification accuracy of 85.2% for the data set Caltech-101 in 10-fold CV, the proposed model also performs well to achieve a classification accuracy of 82.3%. In case of CiteSeer data set, both MvDA and MvDA-VC perform poorly, whereas CSP-MvCDA has better classification accuracy than these two methods, suggesting the importance of including the shared information through correlation across all the views. While all the existing methods do not perform well on NW-Object data set, the proposed method provides comparatively better accuracy in both the cases. Although the proposed model obtains the highest classification accuracy on Cora data set, MvCCDA has also a promising result on this data set.

For ALOI data set, the proposed method is not significant with respect to MvDA-VC in case of Wilcoxon test, which is shown in Table 3.5 in italics. In case of CiteSeer and NW-Object, one p-value is not significant in each case, obtained by Wilcoxon test for the model MvCCDA, as shown in Table 3.6 in italics. Similarly, for NW-Object, the performance of the proposed method is not significant with respect to LACK for both statistical tests. Out of total 130 cases of benchmark data sets, the proposed method achieves significantly better p-values in 125 cases. The non-significant p-value represents the similar performance of two models. Overall, the proposed method has better classification ability on benchmark data sets than the existing approaches.

From all the results reported in this chapter on both real-life cancer data sets and benchmark data sets, it can be concluded that the proposed CSP-MvCDA method outperforms all the existing methods. The results also signify that the CSP-MvCDA has the advantage over some of the non-linear deep learning based methods such as DMvTSVM-AE and SMDDRL, in terms of execution time and performance on large data sets. The comprehensive performance analysis also signifies that the proposed method can take the shared and complementary information across all the multiple views in a balanced and much simpler way with respect to other algorithms.

3.4.4.3 Performance of CSP-MvCDA on Deep Features

In order to evaluate the performance of the proposed CSP-MvCDA method on deep features, two deep learning models, namely, deep Boltzmann machines (DBM) [217] and deep convolutional neural networks (DCNN) [40], are considered to extract deep features. Two data sets, namely, CiteSeer and Cora, are taken as examples. For each view, ten deep features are extracted for the analysis. After extracting the deep features, the proposed CSP-MvCDA method is applied to generate the common correlated discriminative space

Table 3.6: Classification Accuracy of Proposed Method (CSP-MvCDA) and Different Algorithms on Benchmark Data Sets

Different Algorithms	Data Sets	Accuracy (Train-Test)	Accuracy and Significance Analysis for 10-fold CV					Time
			Mean	Median	StdDev	Paired-*t*:p	Wilcoxon:p	(Sec.)
MCCA [106]	CiteSeer	0.581	0.592	0.599	0.028	**1.75E-15**	**1.95E-03**	0.9
GMA [222]		0.652	0.652	0.646	0.025	**7.41E-05**	**1.95E-03**	11.3
MvDA [124]		0.377	0.414	0.416	0.020	**1.81E-14**	**1.95E-03**	30.5
MvDA-VC [124]		0.435	0.477	0.473	0.032	**1.34E-14**	**1.95E-03**	27.2
MLDA [286]		0.666	0.671	0.676	0.026	**7.21E-04**	**1.07E-02**	7.5
MULDA [239]		0.666	0.671	0.676	0.026	**1.01E-02**	**5.85E-03**	7.5
MvCCDA [292]		0.685	0.682	0.694	0.026	**1.66E-19**	*6.44E-02*	17.4
ReDMiCA [164]		0.646	0.641	0.643	0.027	**1.75E-15**	**1.95E-03**	447.7
DMvTSVM-AE [278]		0.489	0.480	0.480	0.016	**1.73E-14**	**1.95E-03**	132.6
SMDDRL [116]		0.515	0.618	0.625	0.028	**1.15E-14**	**1.95E-03**	64.3
LACK [296]		0.694	0.646	0.648	0.014	**2.54E-09**	**1.95E-03**	0.2
GPSNMF [153]		0.650	0.642	0.653	0.030	**1.33E-05**	**1.95E-03**	4.7
CONMF [152]		0.668	0.662	0.668	0.029	**6.28E-04**	**3.91E-03**	3.4
CSP-MvCDA		**0.698**	**0.689**	**0.710**	0.027	-	-	3.0
MCCA [106]	Cora	0.329	0.247	0.256	0.073	**2.15E-15**	**1.95E-03**	6.1
GMA [222]		0.609	0.604	0.608	0.024	**8.01E-08**	**1.95E-03**	57.4
MvDA [124]		0.539	0.495	0.504	0.029	**1.29E-15**	**1.95E-03**	71.0
MvDA-VC [124]		0.557	0.438	0.441	0.020	**1.13E-16**	**1.95E-03**	70.1
MLDA [286]		0.567	0.690	0.692	0.014	**2.16E-07**	**1.95E-03**	30.9
MULDA [239]		0.574	0.683	0.687	0.013	**2.92E-08**	**1.95E-03**	31.1
MvCCDA [292]		0.716	0.748	0.753	0.038	**2.45E-02**	**2.73E-02**	15.0
ReDMiCA [164]		0.573	0.615	0.610	0.035	**1.76E-15**	**1.95E-03**	1286.8
DMvTSVM-AE [278]		0.704	0.755	0.754	0.023	**2.61E-02**	**2.73E-02**	137.3
SMDDRL [116]		0.613	0.686	0.683	0.021	**2.92E-08**	**1.95E-03**	38.8
LACK [296]		0.769	0.783	0.789	0.022	**1.10E-09**	**1.95E-03**	0.2
GPSNMF [153]		0.587	0.607	0.593	0.030	**2.55E-08**	**1.95E-03**	7.6
CONMF [152]		0.669	0.687	0.685	0.032	**1.65E-05**	**1.95E-03**	6.7
CSP-MvCDA		**0.778**	**0.785**	**0.801**	0.021	-	-	6.3
MCCA [106]	NW-Object	0.303	0.322	0.321	0.005	**5.48E-17**	**1.95E-03**	0.5
GMA [222]		0.283	0.296	0.298	0.008	**1.11E-11**	**1.95E-03**	0.2
MvDA [124]		0.290	0.280	0.279	0.008	**5.25E-17**	**1.95E-03**	53.2
MvDA-VC [124]		0.286	0.279	0.281	0.008	**5.55E-17**	**1.95E-03**	73.8
MLDA [286]		0.392	0.323	0.322	0.012	**1.18E-10**	**1.95E-03**	0.9
MULDA [239]		0.370	0.314	0.314	0.007	**9.95E-12**	**1.95E-03**	0.9
MvCCDA [292]		0.424	0.431	0.432	0.007	**4.58E-02**	*6.44E-02*	1745.3
ReDMiCA [164]		0.377	0.382	0.382	0.009	**5.30E-17**	**1.95E-03**	1344.7
DMvTSVM-AE [278]		0.387	0.374	0.373	0.027	**1.29E-17**	**1.95E-03**	145.3
SMDDRL [116]		0.346	0.368	0.366	0.008	**5.48E-17**	**1.95E-03**	240.7
LACK [296]		0.424	0.434	0.434	0.006	*8.02E-01*	*8.46E-01*	2.8
GPSNMF [153]		0.166	0.181	0.182	0.009	**9.35E-15**	**1.95E-03**	84.1
CONMF [152]		0.359	0.369	0.369	0.007	**1.78E-11**	**1.95E-03**	49.5
CSP-MvCDA		**0.440**	**0.445**	**0.446**	0.008	-	-	72.6
MCCA [106]	Caltech-101	0.306	0.345	0.332	0.013	**1.72E-14**	**1.95E-03**	5.2
GMA [222]		0.379	0.285	0.286	0.013	**2.16E-15**	**1.95E-03**	4.4
MvDA [124]		0.607	**0.852**	**0.854**	0.091	**2.31E-08**	**1.95E-03**	1.1
MvDA-VC [124]		0.605	0.839	0.839	0.096	**2.01E-04**	**1.95E-03**	4.9
MLDA [286]		0.575	0.495	0.489	0.014	**1.07E-15**	**1.95E-03**	51.3
MULDA [239]		0.572	0.493	0.489	0.024	**1.65E-15**	**1.95E-03**	24.2
MvCCDA [292]		0.605	0.815	0.815	0.009	**1.66E-14**	**1.95E-03**	25.3
ReDMiCA [164]		0.391	0.412	0.400	0.013	**5.85E-09**	**1.95E-03**	4531.7
DMvTSVM-AE [278]		0.553	0.737	0.736	0.007	**1.66E-14**	**1.95E-03**	164.0
SMDDRL [116]		0.378	0.704	0.715	0.017	**3.51E-09**	**1.95E-03**	128.3
LACK [296]		0.372	0.417	0.420	0.013	**2.04E-14**	**1.95E-03**	5.9
GPSNMF [153]		0.258	0.265	0.271	0.039	**8.84E-13**	**1.95E-03**	22.4
CONMF [152]		0.362	0.401	0.401	0.015	**4.91E-16**	**1.95E-03**	30.6
CSP-MvCDA		**0.610**	0.825	0.825	0.009	-	-	14.8

for classification. The results corresponding to training-testing are reported in Table 3.7. From the results reported here, it can be seen that the performance of the proposed method on original features is comparable with or better than that of deep features. This is mainly due to the fact that the extracted deep features are small in number and do not consider the supervised information of class labels.

Table 3.7: Classification Accuracy of Proposed Method on Deep Features

Different	Deep Features		Original
Data Sets	DCNN [40]	DBM [217]	Features
CiteSeer	**69.85**	66.58	69.80
Cora	75.03	74.03	**77.80**

3.5 Conclusion

This chapter introduces a new supervised subspace learning technique, termed as CSP-MvCDA, by judiciously combining the merits of both MCCA and LDA. It also includes a class-structure preserving norm to preserve the class-structure of each view. The proposed method forms a common discriminative latent space, where both the global and local geometry of the classes in each view are preserved. It has been shown that the formulation of CSP-MvCDA can be solved easily by solving a generalized eigenvalue problem. The proposed CSP-MvCDA method can be used as an alternative to MCCA, in case of supervised multi-view subspace learning. The CSP-MvCDA can deal with non-linear data as well because of its class-structure preserving norm. To estimate the potency of the proposed method and to compare it with that of the state-of-the-art multi-view subspace learning techniques, several real-life cancer and benchmark data sets have been used under the 10-fold cross-validation and training-testing experimental set-up. The results on both types of data sets show that the proposed approach outperforms the state-of-the-art subspace learning methods.

The algorithm, proposed in this chapter, is based on the concept of the SUMCOR criterion of MCCA. To reduce the complexity, another supervised MVL algorithm, based on maximum variance criterion of MCCA, has been proposed in the next chapter.

Chapter 4

Supervised Graph Regularized Multi-View Canonical Correlation and Discrimination Analysis

4.1 Introduction

Due to the information explosion, the size of data representing an object from different perspectives is increasing rapidly in the fields of pattern recognition, data mining, and computer vision. These multiple perspectives create heterogeneous and distinct views of the common sources. As mentioned in Chapter 3, learning from these heterogeneous and distinct views to express the object more comprehensively is known as multi-view learning [238]. Numerous studies have demonstrated that the proper integration of data from various viewpoints will improve the performance in several learning tasks, including dimensionality reduction, classification, and clustering than the performance obtained from each view separately [37, 312].

In multi-view learning, the feature sets collected from different perspective can be diverse in nature, high in dimension, and may not be equally relevant to a desired task. Moreover, these multiple views should be integrated to make use of the consistent information across all the views and the complementary information of each view properly. As mentioned in Chapter 3, the canonical correlation analysis (CCA) [107] is a classical multi-view subspace learning technique, which forms a reduced low dimensional subspace by maximizing the correlation between two views. The multi-set CCA (MCCA) [129] is the generalization of the two view CCA, which seeks for the common low dimensional subspace by maximizing the pairwise correlation among multiple views. There are several MCCA formulations according to the optimizing criteria, such as SUMCOR, maximum variance (MAXVAR), sum of squared correlations, minimum variance, and generalized variance methods [105, 129]. However, MCCA is unable to use label of the data, can only explore the linear relationships among the multiple views, and suffers from singularity issue of the covariance matrices for "large p small n" problem.

The regularized CCA (RCCA) [259] is the extension of CCA, which makes the singular

covariance matrix invertible by adding some small positive quantity to the diagonal of the matrix and prevents the overfitting problem. The supervised RCCA [81] is the supervised extension of RCCA, which learns the optimal regularizer by some statistical test for better classification performance. The supervised and regularized version of MCCA has been explored in [164]. A supervised multi-view CCA (sMVCCA) has also been proposed in [145]. To explore the non-linear relationship between the data, several extension of CCA and MCCA, based on kernel learning [224], locality preserving projection [242] and deep neural networks [12] [79], have been introduced. Moreover, the CCA or MCCA based models do not explore the geometry of the data. In [44], a graph regularized CCA (gCCA) has been proposed, which incorporates the graph information of the data as a regularizer into the objective function of CCA. The graph multi-view CCA (GMCCA) [43] is the extension of the gCCA in multi-view settings. In [298], Laplacian MCCA (LapMCCA) has been introduced to uncover the non-linear correlation hidden in multiple views. Although these methods explore the prior geometry of the data, they are unsupervised in nature. Moreover, all the methods described above give equal importance to each of the views for constructing the common latent space. But, the integration of multiple views according to the relevance of each view is reasonable and may improve the performance of the models.

In this regard, a supervised subspace learning approach, based on MAXVAR formulation of MCCA is presented in this chapter. The proposed formulation is termed as multi-view canonical correlation and discrimination analysis (MCCDA). The MCCDA finds a latent low dimensional subspace, where the covariance between any two views is maximized and the within-class scatter of each view is minimized. The features obtained by this approach contain both the correlated information among multiple views and discriminative information of each view, which make the formulation more suitable for multi-class classification problem. The proposed model is further extended by considering the prior geometrical structure of the data encoded by the within-class and between-class graphs, which are used to construct a regularizer term and incorporated into the model. The model also learns a proper weight for each view while integrating the multiple views. The extended version of MCCDA is termed as supervised graph regularized MCCDA (SGR-MCCDA). The SGR-MCCDA is a unified framework, where label information as well as known geometry of data are explored and a suitable weight for each view is learned. Furthermore, an iterative rule is presented to solve the optimization problem of SGR-MCCDA. The convergence analysis and computational complexity analysis are also given to justify the efficacy of the proposed method. Several cancer and benchmark data sets with different characteristics are used to show the effectiveness of the proposed model. The results confirm that the SGR-MCCDA outperforms the existing algorithms in terms of classification accuracy, and has comparatively lower execution time than the existing ones. Some of the results of this chapter can be found in [179].

The remaining sections of this chapter are as follows: Section 4.2 presents brief ideas of the MAXVAR criterion of MCCA, and GMCCA. Section 4.3 presents the proposed algorithm. An iterative rule has been provided in this section for solving the proposed optimization problem. Computational complexity analysis for different algorithms is also given in this section. The effectiveness of the proposed algorithm is shown by comparing the performance of it with the other state-of-the-art algorithms over several omics and benchmark data sets in Section 4.5. The chapter is concluded in Section 4.6.

4.2 Basics of MCCA and GMCCA

This section gives a brief overview of MAXVAR formulation of MCCA [129] and GMCCA [43].

4.2.1 MAXVAR Criterion of MCCA

Suppose M feature sets, $\{X_v \in \mathbb{R}^{N\times d_v}\}_{v=1}^{M}$, are given for M number of views which are obtained from a common data source matrix $\hat{S} \in \mathbb{R}^{N\times\rho}$, where N is the number of common source vector $\hat{s}_n(n = 1, 2, \ldots, N)$, $d_v(v = 1, 2, \ldots, M)$ is the number of feature in the v-th feature set and ρ is the original dimension of the source vector with $\rho << \min_v \{d_v\}_{v=1}^{M}$. The classical CCA for two views X_1 and X_2 looks for two linear transformations, $W_1 \in \mathbb{R}^{d_1\times d}$ and $W_2 \in \mathbb{R}^{d_2\times d}$ with $d \leqslant \rho$, so that the transformed views X_1W_1 and X_2W_2 have maximum correlation or it can be said that their euclidean distance is minimum. This is known as the sum of correlation (SUMCOR) formulation of the CCA. The MCCA is multi-view extension of the two view CCA. The MAXVAR formulation of the MCCA finds a shared low dimensional space, S, such that the euclidean distance between each view's linear projection X_vW_v and S is minimized. The formulation is given by

$$\begin{aligned} &\min_{\{W_v\}_{v=1}^{M}} \sum_{v=1}^{M} \|X_vW_v - S\|_F^2, \\ &\text{subject to } S^TS = I_d. \end{aligned} \tag{4.1}$$

If the sample covariance matrices, $\{X_v^TX_v\}_{v=1}^{M}$, are all non-singular, then by fixing S, the W_v-minimizer of (4.1) can be obtained from the following equation

$$W_v = (X_v^TX_v)^{-1}X_v^TS.$$

Replacing the values of W_v into (4.1), the S-minimizer would be the solution of the following eigendecomposition problem:

$$\begin{aligned} &S^* := \arg\max_S \operatorname{Tr}\left[S^T\left(\sum_{v=1}^{M} X_v(X_v^TX_v)^{-1}X_v^T\right)S\right], \\ &\text{subject to } S^TS = I_d. \end{aligned} \tag{4.2}$$

The top d principal component of the matrix $\sum_{v=1}^{M} X_v(X_v^TX_v)^{-1}X_v^T$ would give the d columns of S^*. In succession, the optimal linear transformations are given by

$$W_v{}^* = (X_v^TX_v)^{-1}X_v^TS. \tag{4.3}$$

It is worth noting that the solution of SUMCOR and MAXVAR criteria of MCCA are different. In particular, for $M = 2$, both the variant can be solved analytically by eigen-decomposition; but for $M > 2$, the SUMCOR cannot be solved analytically, while the MAXVAR still admits analytical solution though there is an extra variable S to compute.

4.2.2 GMCCA

The GMCCA is the extension of the MAXVAR criterion of the MCCA, where the known geometry of the data is incorporated into the model as a graph regularizer. Suppose $\mathcal{G} : (\mathcal{V}, \mathcal{W})$ is the graph representing structural geometry of the N common sources, where $\mathcal{V}$ denotes the vertex set $\{\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_N\}$ and $\mathcal{W}$ is the set consisting of all the edge weights $w_{ij}(> 0)$ between the vertices $\hat{s}_i$ and $\hat{s}_j$ $(i, j \in \mathcal{V})$. Let $W \in \mathbb{R}^{N\times N}$ be the weighted adjacency matrix of $\mathcal{G}$, whose (i, j)-entry is the edge weight w_{ij}. Let the degree matrix be $D := \text{diag}(\{d_i\}_{i=1}^N)$, where $d_i = \sum_j w_{ij}$. The graph Laplacian $L_\mathcal{G}$ of $\mathcal{G}$, which can be computed as: $L_\mathcal{G} = D - W$, is used to form the following regularizer term:

$$\text{Tr}\,(S^T L_\mathcal{G} S) = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\|s_i - s_j\|_2^2. \tag{4.4}$$

The term is then invoked into the model of MAXVAR criterion of MCCA to form the objective function of GMCCA as given by:

$$\begin{aligned} &\min_{\{W_d\}_{d=1}^M, S} \sum_{v=1}^{M} \|X_v W_v - S\|_F^2 + \gamma\,\text{Tr}\,(S^T L_\mathcal{G} S), \\ &\text{subject to } S^T S = I_d, \end{aligned} \tag{4.5}$$

where γ is the trade-off parameter, which controls the minimization of distance between canonical variables and common source vectors, and smoothness of the common source estimates over the graph $\mathcal{G}$. Similar to the MAXVAR criterion of MCCA, the optimal solution S^* of the problem (4.5) can be obtained from the top d eigenvalues of the matrix

$$C = \sum_{v=1}^{M} X_v (X_v^T X_v)^{-1} X_v^T - \gamma L_\mathcal{G}$$

and ${W_v}^*$ as given in (4.3).

4.3 Formulation of SGR-MCCDA

The MCCA finds the optimal linear transformation without using the available prior information of class label or the known geometry of source vectors. Incorporating the prior knowledge of class label or structural geometry of the data may improve the performance of MCCA. Moreover, most of the MCCA based models consider equal weight of the views to construct the low dimensional subspace. In this section, a supervised graph regularized multi-view learning algorithm, based on the MAXVAR criterion of MCCA, has been proposed, which uses the prior information of the data in a weighted average manner to improve the performance in multi-class classification.

4.3.1 MCCDA: Multi-View Canonical Correlation and Discrimination Analysis

Suppose, the N source vectors $\{\hat{s}_n\}_{n=1}^{N}$ are categorized into K number of classes $\{c_1, c_2, \ldots, c_K\}$. Then, the within-class scatter matrix (S_w^v) and between-class scatter matrix (S_b^v) for v-th view can be computed from the following formula:

$$S_w^v = \sum_{k=1}^{K} \sum_{i \in c_k} (x_i^v - \mu_{c_k}^v)(x_i^v - \mu_{c_k}^v)^T; \tag{4.6}$$

$$S_b^v = \sum_{k=1}^{K} n_k (\bar{x}^v - \mu_{c_k}^v)(\bar{x}^v - \mu_{c_k}^v)^T, \tag{4.7}$$

where $\bar{x}^v$ is the mean over all the samples and $\mu_{c_k}^v$ is the mean over the samples of the class c_k for the v-th view X_v. The MAXVAR formulation of MCCA in (4.1) is modified to incorporate the supervised information by adding an extra term $-\lambda \operatorname{Tr}(W_v^T S_b^v W_v)$ into the objective function to obtain a new formulation given by:

$$\begin{aligned} &\min_{\{W_v\}_{v=1}^{M}, S} \sum_{v=1}^{M} \left[\|X_v W_v - S\|_F^2 - \lambda \operatorname{Tr}(W_v^T S_b^v W_v) \right], \\ &\text{subject to } S^T S = I_d. \end{aligned} \tag{4.8}$$

The above formulation, different from the MAXVAR criterion of MCCA, maximizes the between-class variance of each view in the projected space and hence minimizes the within-class variance of each view while keeping the projected space as near as possible to the common space S. The formulation is termed as multi-view canonical correlation and discrimination analysis (MCCDA). The optimization problem of MCCDA can be solved using Lagrangian multiplier method. Consider the Lagrangian function of the above problem:

$$f(\{W_v\}_{v=1}^{M}, S) = \sum_{v=1}^{M} \left[\|X_v W_v - S\|_F^2 - \lambda \operatorname{Tr}(W_v^T S_b^v W_v) \right] + Tr\{(S^T S - I_d)\Lambda\}, \tag{4.9}$$

where Λ is the diagonal matrix consisting of the Lagrange's multiplier of each sub problem of (4.8).

After simplifying, the expression (4.9) reduces to

$$f(\{W_v\}_{v=1}^{M}, S) = \sum_{v=1}^{M} \left[d - 2 \operatorname{Tr}\{S^T X_v W_v + W_v^T (X_v^T X_v - \lambda S_b^v)^b W_v\} \right] + \operatorname{Tr}\{(S^T S - I_d)\Lambda\}. \tag{4.10}$$

Assuming $\lambda = 1$ and substituting $X_v^T X_v - \lambda S_b^v = X_v^T X_v - S_b^v = S_w^v$ into (4.10), the

equation becomes

$$f(\{W_v\}_{v=1}^M, S) = \sum_{v=1}^{M} \left[d - \mathrm{Tr}(2S^T X_v W_v - W_v^T S_w^v W_v)\right] + \mathrm{Tr}\{(S^T S - I_d)\Lambda\}. \tag{4.11}$$

The function f is convex in S, when $\{W_v\}_{v=1}^M$ is fixed and also convex in each of the W_v, when S is fixed. The problem can be solved using alternating optimization method. Keeping S fixed, the W_v-minimizer is given by the equation:

$$\begin{aligned} \frac{\partial f}{\partial W_v} = 0 \Rightarrow & -2X_v^T S + 2S_w^v W_v = 0 \\ \Rightarrow & W_v = (S_w^v)^{-1} X_v^T S. \end{aligned} \tag{4.12}$$

Keeping each of the W_v-fixed, the S-minimizer is given by the equation

$$\begin{aligned} \frac{\partial f}{\partial S} = 0 \Rightarrow & -2 \sum_{v=1}^{M} X_v W_v + 2S\Lambda = 0 \\ \Rightarrow & -2 \sum_{v=1}^{M} X_v (S_w^v)^{-1} X_v^T S + 2S\Lambda = 0 \\ \Rightarrow & S^T \left[\sum_{v=1}^{M} X_v (S_w^v)^{-1} X_v^T\right] S = \Lambda. \end{aligned} \tag{4.13}$$

The above expression suggests that the columns of S^* consist of the top d principal eigenvectors of the matrix $\sum_{v=1}^M X_v (S_w^v)^{-1} X_v^T$ and in succession the W_v-minimizer can be obtained from the equation $W_v^* = (S_w^v)^{-1} X_v^T S^*$. The optimization procedure of MCCDA is given in Algorithm 4.1.

Algorithm 4.1 Algorithm for MCCDA

1: **INPUT:** $\{X_v\}_{v=1}^M$, class label of the samples.
2: Compute the within-class scatter matrices $\{S_w^v\}_{v=1}^M$.
3: Construct $C = \sum_{v=1}^M X_v (S_w^v)^{-1} X_v^T$.
4: Apply eigendecomposition on C and take the d eigenvectors corresponding to the d largest eigenvalues which will constitute the columns of S.
5: Compute $\{W_v = (S_w^v)^{-1} X_v^T S\}_{v=1}^M$.
6: **OUTPUT:** S^* and $\{W_v^*\}_{v=1}^M$

Remark 4.1. *At the optimum point, the minimum value attained by the objective function of MCCDA will be* $\left(Md - \sum_{i=1}^d \lambda_i\right)$*, where* λ_i *is the i'th largest eigenvalue of the matrix* $\sum_{v=1}^M X_v (S_w^v)^{-1} X_v^T$.

Proof. Consider the objective function in (4.8) as

$$\begin{aligned}\mathcal{J}(\{W_v\}_{v=1}^M, S) &= \sum_{v=1}^{M} \left[\|X_v W_v - S\|_F^2 - \lambda \operatorname{Tr}(W_v^T S_b^v W_v)\right] \\ &= \sum_{v=1}^{M} \left[d - \operatorname{Tr}\left(2S^T X_v W_v - W_v^T S_w^v W_v\right)\right] \quad \text{(See 4.8 to 4.11)} \end{aligned} \tag{4.14}$$

Substituting the values of $\{W_v = (S_w^b)^{-1} X_v^T S\}_{v=1}^M$ into (4.14), it reduces to

$$\begin{aligned}\mathcal{J}(\{W_v\}_{v=1}^M, S) &= \sum_{v=1}^{M} \left[d - \operatorname{Tr}\left\{\left(2S^T X_v (S_w^b)^{-1} X_v^T S\right) - \left((S_w^b)^{-1} X_v^T S\right)^T S_w^v ((S_w^b)^{-1} X_v^T S)\right\}\right] \\ &= \sum_{v=1}^{M} d - \operatorname{Tr}\left[S^T \left(\sum_{v=1}^{M} X_v (S_w^b)^{-1} X_v^T\right) S\right] \\ &= Md - \operatorname{Tr}(\Lambda) \quad \text{(Using 4.13)} \\ &= Md - \sum_{i=1}^{d} \lambda_i,\end{aligned}$$

where λ_i is the i'th largest eigenvalue of the matrix $\sum_{v=1}^M X_v \left(S_w^v\right)^{-1} X_v^T$. This concludes the remark. ■

4.3.2 Supervised Graph Regularized MCCDA (SGR-MCCDA)

The objective function of (4.8) is modified further to consider a suitable weight, ω_v, for each view, X_v, so that multiples views are integrated according to their relevance. An entropy term involving ω_v is also added to the objective function to avoid the trivial solution on the weights. The modified objective function is given by:

$$\begin{aligned}&\min_{\{W_v, \omega_v\}_{v=1}^M, S} \sum_{v=1}^{M} \omega_v \left[\|X_v W_v - S\|_F^2 - \lambda \operatorname{Tr}(W_v^T S_b^v W_v)\right] + \gamma \sum_{v=1}^{M} \omega_v \log \omega_v \\ &\text{subject to } S^T S = I_d,\ \boldsymbol{\omega}^T \mathbf{1} = 1,\ 0 \leqslant w_v \leqslant 1,\ \boldsymbol{\omega} = [\omega_1, \omega_2, \ldots, \omega_M]^T. \end{aligned} \tag{4.15}$$

Consider that the structural geometry of N common source vectors can be represented as a graph $\mathcal{G} : (\mathcal{V}, \mathcal{W})$, where $\mathcal{V}$ denotes the vertex set $\{\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_N\}$ and $\mathcal{W}$ is the set consisting of all the edge weights $w_{ij} (> 0)$ between the vertices i and j $(i, j \in \mathcal{V})$. Exploring the idea of [284], two new graphs $\mathcal{G}_W : (\mathcal{V}_1, \mathcal{W}_1)$ and $\mathcal{G}_B : (\mathcal{V}_2, \mathcal{W}_2)$ are constructed using the class labels, where $\mathcal{V}_1 = \mathcal{V}_2 = \mathcal{V}$. The $\mathcal{W}_1$ is the set of edge weights w_{ij}^1 defined by $w_{ij}^1 = w_{ij}$, if the vertices i and j belong to the same class and $w_{ij}^1 = 0$ elsewhere. The $\mathcal{W}_2$ denotes the edge weights w_{ij}^2 defined by $w_{ij}^2 = w_{ij}$, if the vertices i and j belong to different classes and $w_{ij}^2 = 0$ elsewhere. The $\mathcal{G}_W$ and $\mathcal{G}_B$ are known as the within-class and between-class graphs. Let A_1 and A_2 be the weighted adjacency matrices formed by the edge weights w_{ij}^1 and w_{ij}^2, respectively. Define $d_i^1 = \sum_{j=1}^N w_{ij}^1$ and $d_i^2 = \sum_{j=1}^N w_{ij}^2$,

and construct the degree matrices as $D_1 := \text{diag}(\{d^1 i\}_{i=1}^N)$ and $D_2 := \text{diag}(\{d^2 i\}_{i=1}^N)$. The Laplacian matrices for the graphs $\mathcal{G}_W$ and $\mathcal{G}_B$ can be computed as $L_{\mathcal{G}_W} = D_1 - A_1$, and, $L_{\mathcal{G}_B} = D_2 - A_2$. Now, define a new Laplacian graph by using $L_{\mathcal{G}_W}$ and $L_{\mathcal{G}_B}$ as

$$L_{\mathcal{G}'} = \alpha L_{\mathcal{G}_W} - (1-\alpha) L_{\mathcal{G}_B} \tag{4.16}$$

The prior knowledge of graph structure of the source vectors is incorporated into the proposed model. Assume that the source vectors $\{\hat{s}_i\}_{i=1}^N$ are smooth over $\mathcal{G}$, which suggests that two source vectors, s_i and s_j corresponding to the connected nodes i and j, respectively, stay close to each other in euclidean distance. The graph $\mathcal{G}_W$ suggests that the nodes belonging to the same class are connected, whereas the graph $\mathcal{G}_B$ establishes the connection between the nodes of different classes. A graph regularizer term, which is constructed using the knowledge of both the within-class and between-class graph, is given by

$$\text{Tr}\left(S^T L_{\mathcal{G}'} S\right) = \sum_{i=1}^{N} \sum_{j=1}^{N} w'_{ij} \|s_i - s_j\|_2^2, \tag{4.17}$$

where s_i and s_j are d-dimensional approximations of the source vectors $\hat{s}_i$ and $\hat{s_j}$, respectively, and, $w'_{ij} = \alpha w^1_{ij} - (1-\alpha) w^2_{ij}$.

It is clear from the definition of w'_{ij} that the large positive value of w'_{ij} will bring the adjacent nodes i and j, belonging to the same class, closer and larger negative value of w'_{ij} will make the nodes i and j, belonging to different classes, far apart in the projected space. To use the prior geometry of the data and labels, the term (4.17) is included as a regularizer into (4.15), resulting in a new formulation

$$\begin{aligned} &\min_{S, \{W_v, \omega_v\}_{v=1}^M} \sum_{v=1}^{M} \omega_v \left[\|X_v W_v - S\|_F^2 - \lambda \,\text{Tr}(W_v^T S_b^v W_v) + \gamma \log \omega_v\right] + \beta \,\text{Tr}\left(S^T L_{\mathcal{G}'} S\right) \\ &\text{subject to } S^T S = I_d,\ \boldsymbol{\omega}^T \mathbf{1} = 1,\ 0 \leqslant \omega_v \leqslant 1,\ \boldsymbol{\omega} = [\omega_1, \omega_2, \ldots, \omega_M]^T. \end{aligned} \tag{4.18}$$

This formulation is termed as supervised graph-regularized weighted MCCDA (SGR-MCCDA). The proposed SGR-MCCDA method is pictorially represented in Fig. 4.1, along with MCCA-MAXVAR and MCCDA. It also gives a brief overview of the construction of the discriminative subspace.

4.3.3 Optimization of SGR-MCCDA

Consider the objective function of SGR-MCCDA as

$$\mathcal{J}(\{W_v, \omega_v\}_{v=1}^M, S) = \sum_{v=1}^{M} \omega_v \left[\|X_v W_v - S\|_F^2 - \lambda \,\text{Tr}(W_v^T S_b^v W_v) + \gamma \log \omega_v\right] + \beta \,\text{Tr}\left(S^T L_{\mathcal{G}'} S\right) \tag{4.19}$$

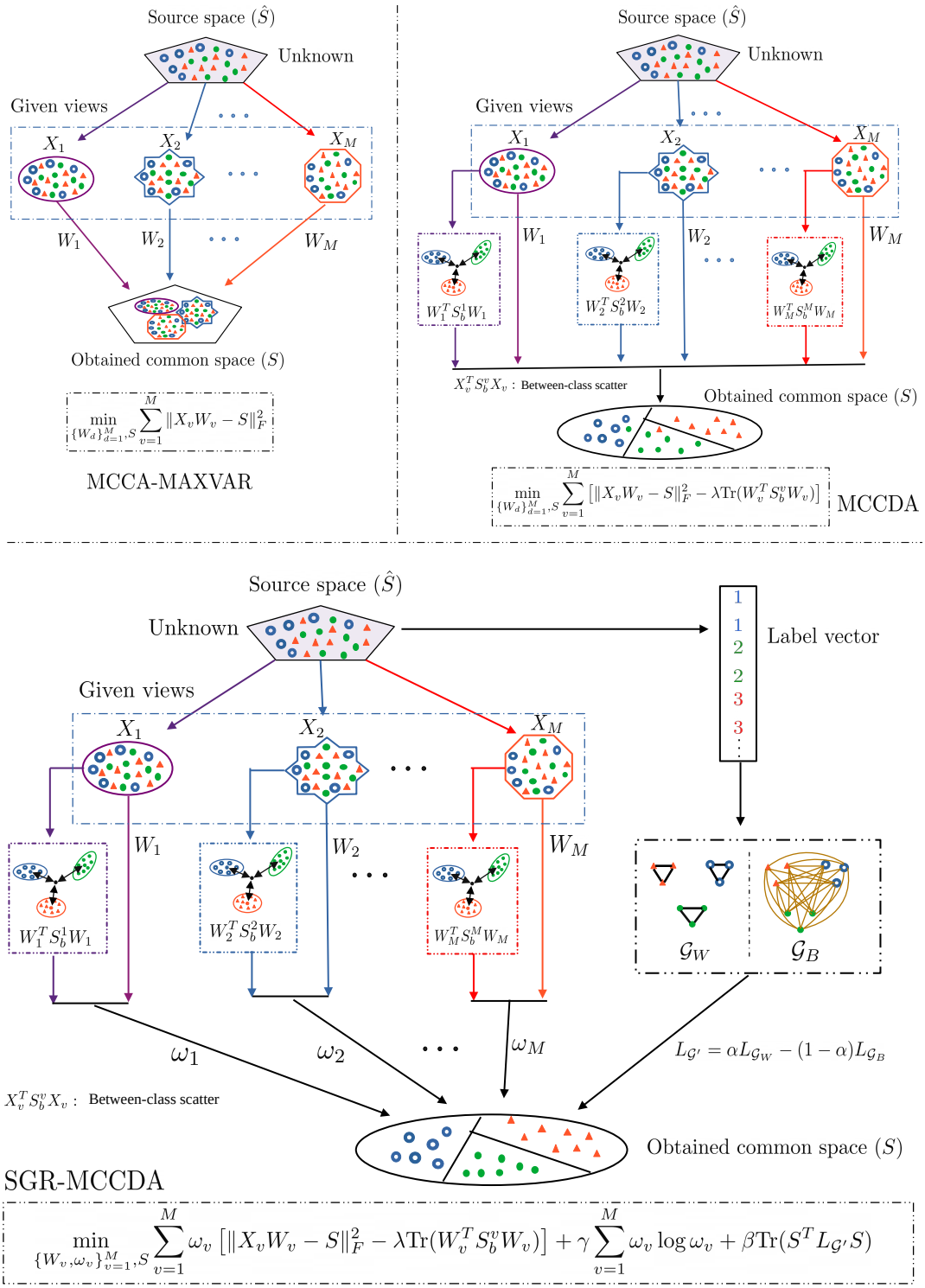


Figure 4.1: Pictorial representation of MCCA-MAXVAR, proposed MCCDA, and proposed SGR-MCCDA.

The objective function, $\mathcal{J}(\{W_v, \omega_v\}_{v=1}^M, S)$, involves the variables $\{W_v\}_{v=1}^M, S, \{\omega_v\}_{d=1}^M$ to be optimized. To solve this, an alternating optimization technique is adopted, where each variable is updated iteratively while other variables remain fixed. Consider the Lagrangian function of (4.18) as:

$$L = \sum_{v=1}^{M} \omega_v \left[\|X_v W_v - S\|_F^2 - \lambda \operatorname{Tr}(W_v^T S_b^v W_v) + \gamma \log \omega_v \right] + \beta \operatorname{Tr}\left(S^T L_{\mathcal{G}'} S\right) + \operatorname{Tr}\{(S^T S - I_d)\Lambda\} - \lambda' \left(\sum_{v=1}^{M} \omega_v - 1 \right) \tag{4.20}$$

Simplifying the above expression and taking $\lambda = 1$, the Lagrangian in (4.20) reduces to

$$L = \sum_{v=1}^{M} \omega_v \left[d - \operatorname{Tr}(2S^T X_v W_v - W_v^T S_w^v W_v) + \gamma \log \omega_v \right] + \beta \operatorname{Tr}\left(S^T L_{\mathcal{G}'} S\right) + \operatorname{Tr}\{(S^T S - I_d)\Lambda\} - \lambda' \left(\sum_{v=1}^{M} \omega_v - 1 \right) \tag{4.21}$$

4.3.3.1 Update Projection Matrix

If S and ω_v are fixed, the last three terms of the Lagrangian in (4.21) can be treated as constant. Taking partial differentiation of (4.21) with respect to W_v and equating it to 0 gives

$$\begin{aligned} \frac{\partial L}{\partial W_v} = 0 \Rightarrow & -2X_v^T S + 2S_w^v W_v = 0 \\ \Rightarrow & W_v = (S_w^v)^{-1} X_v^T S. \end{aligned} \tag{4.22}$$

4.3.3.2 Update Common Space

Fixing W_v and ω_v, the terms involving them in (4.21) can be treated as constant and thus taking the partial differentiation of L with respect to S and equating it to 0 gives

$$\frac{\partial L}{\partial S} = 0 \Rightarrow -2 \sum_{v=1}^{M} \omega_v X_v W_v + 2\beta L_{\mathcal{G}'} S + 2S\Lambda = 0$$

$$\Rightarrow - \sum_{v=1}^{M} \omega_v X_v \left(S_w^v\right)^{-1} X_v^T S + \beta L_{\mathcal{G}'} S + S\Lambda = 0$$

$$\Rightarrow S^T \left[\sum_{v=1}^{M} \omega_v X_v \left(S_w^v\right)^{-1} X_v^T - \beta L_{\mathcal{G}'} \right] S = \Lambda. \tag{4.23}$$

The relation (4.23) suggests that S consists of top d eigenvecctors of the matrix

$$C = \sum_{v=1}^{M} \omega_v X_v \left(S_w^v\right)^{-1} X_v^T - \beta L_{\mathcal{G}'}. \tag{4.24}$$

4.3.3.3 Update View Coefficient

Fixing W_v and S, the terms involving them in (4.21) can be taken as constant. Therefore, taking the partial differentiation of L with respect to ω_v and equating it to 0 gives

$$\frac{\partial L}{\partial \omega_v} = 0 \Rightarrow r_v + \gamma(1 + \log \omega_v) - \lambda' = 0 \tag{4.25}$$

$$\Rightarrow \omega_v = e^{\frac{\lambda' - r_v}{\gamma} - 1}, \tag{4.26}$$

where $r_v = d - \mathrm{Tr}(2S^T X_v W_v - W_v^T S_w^v W_v)$. Again, from (4.25), the value of λ' is given by

$$\lambda' = r_v + \gamma(1 + \log \omega_v). \tag{4.27}$$

Now, substituting the values of ω_v from (4.26) into the equation $\sum_{v=1}^{M} \omega_v = 1$ gives

$$\sum_{v=1}^{M} e^{\frac{\lambda' - r_v}{\gamma} - 1} = 1. \tag{4.28}$$

Substituting the value of λ' and simplifying the above equation, the updation rule of ω_v can be obtained as

$$\omega_v = \frac{e^{-\frac{r_v}{\gamma}}}{\sum_{v=1}^{M} e^{-\frac{r_v}{\gamma}}} \tag{4.29}$$

The combined iterative rule (4.22), (4.23) and (4.29) gives the required shared space S and projection matrices $\{W_v\}_{v=1}^{M}$ upon convergence. Later the projection matrices are used to project the data into common latent space. The optimization procedure of the proposed SGR-MCCDA is described in Algorithm 4.2.

Remark 4.2. *When the dimension of the data matrix is very high compared to the sample size, that is, $d_v >> N$, the within-class scatter matrix, S_w^v, will become singular, which makes the matrix non-invertible and yields undesirable solution to the problems of (4.8) and (4.18). Similar to an analysis reported in [43], a dual formulation of SGR-MCCDA can be adopted for such cases. However, to tackle this problem, one may also use a low rank approximation of the scatter matrices to find the inverse of the matrices. In the current study, eigendecomposition is used to approximate the scatter matrix and find the inverse of it. The following formula is used to find the inverse of the within-class scatter matrix of*

Algorithm 4.2 Algorithm for SGR-MCCDA

INPUT: $\{X_v\}_{v=1}^{M}$, d, γ, β, α, class label of samples.
INITIALIZE $\omega_v = 1/M$ for $v = 1, 2, \ldots, M$.
Construct $L_{\mathcal{G}'}$ using (4.16).
while not converge **do**
    Compute $C = \sum_{v=1}^{M} \omega_v X_v \left(S_w^v\right)^{-1} X_v^T - \beta L_{\mathcal{G}'}$
    Apply eigendecomposition on C and take the d eigenvectors corresponding to the d largest eigenvalues which constitute the columns of S according to (4.23).
    Update W_v according to (4.22).
    Update ω_v according to (4.29).
end while
OUTPUT: S^* and $\{W_v^*\}_{v=1}^{M}$.

the v-th view.

$$S_w^v = \sum_{r=1}^{R} \delta_r^v u_r^v u_r^{vT}, \tag{4.30}$$

where δ_r^v is the r'th largest eigenvalue and u_r^v is the corresponding eigenvector of the v-th within-class scatter matrix; R is the optimal rank, chosen in such a way that the 99% variance of the matrix is retained.

Remark 4.3. *In the experiment, the prior graph information ($L_{\mathcal{G}'}$) of the estimated source vectors can effectively be obtained from the existing views [43]. Consider, $A^{(v)}$ is the adjacency matrix of the v'th view, X_v, consisting of the edge weights, w_{ij}^v, between the i'th and j'th samples. The w_{ij}^v is defined below as:*

$$w_{ij}^v = \begin{cases} e^{-\frac{\|x_i^v - x_j^v\|^2}{\sigma^2}} & x_i^v \in \mathcal{N}_k(x_j^v) \text{ or } x_j^v \in \mathcal{N}_k(x_i^v) \\ 0 & \text{otherwise} \end{cases} \tag{4.31}$$

where $\mathcal{N}_k(x_j^v)$ is the set of all points belonging to k-nearest neighbor of x_j^v. The edge weights, w_{ij}, used in the current study can be taken as $w_{ij} = \sum_{v=1}^{M} w_{ij}^v$.

4.4 Theoretical Analysis

This section presents some of the important properties, convergence analysis and computational complexity analysis of the proposed SGR-MCCDA algorithm.

4.4.1 Important Properties

Theorem 4.1. *Given α, β and γ, the minimum value of the objective function of SGR-MCCDA for a fixed set $\{\omega_v\}$, is $\left[d - \sum_{i=1}^{d} \lambda_i + \gamma \sum_{i=1}^{M} \omega_v \log \omega_v\right]$, where λ_i is the i-th largest eigenvalue of the matrix C of (4.24).*

Proof. Consider the reduced form of the objective function in (4.18) as

$$\mathcal{J}(\{W_v, \omega_v\}_{v=1}^{M}, S) = \sum_{v=1}^{M} \omega_v \left[d - \operatorname{Tr}(2S^T X_v W_v)\right] + \sum_{v=1}^{M} \omega_v \operatorname{Tr}(W_v^T S_w^v W_v) + \gamma \sum_{v=1}^{M} \omega_v \log \omega_v + \beta \operatorname{Tr}(S^T L_{\mathcal{G}'} S). \tag{4.32}$$

Substituting the values of $\{W_v = (S_w^v)^{-1} X_v^T S\}_{v=1}^{M}$ into (4.32), it reduces to

$$\mathcal{J}(\{W_v, \omega_v\}_{v=1}^{M}, S) = \sum_{v=1}^{M} \omega_v \left[d - \operatorname{Tr}(2S^T X_v (S_w^v)^{-1} X_v^T S)\right] + \sum_{v=1}^{M} \omega_v \left[\operatorname{Tr}\left\{((S_w^v)^{-1} X_v^T S)^T S_w^v ((S_w^v)^{-1} X_v^T S)\right\} + \gamma \log \omega_v\right] + \beta \operatorname{Tr}(S^T L_{\mathcal{G}'} S). \tag{4.33}$$

After simplifying, (4.33) becomes

$$\mathcal{J}(\{W_v, \omega_v\}_{v=1}^{M}, S) = \sum_{v=1}^{M} \omega_v \left[d + \gamma \log \omega_v\right] - \operatorname{Tr}\left[S^T \left\{\sum_{v=1}^{M} \omega_v X_v (S_w^v)^{-1} X_v^T - \beta L_{\mathcal{G}'}\right\} S\right]. \tag{4.34}$$

Using (4.23), the minimum value of the objective function $\mathcal{J}(\{W_v, \omega_v\}_{v=1}^{M}, S)$ of (4.34) becomes

$$\begin{aligned} \mathcal{J}^* &= d + \gamma \sum_{v=1}^{M} \omega_v \log \omega_v - \operatorname{Tr}(\Lambda) \\ &= d - \sum_{i=1}^{d} \lambda_i + \gamma \sum_{v=1}^{M} \omega_v \log \omega_v. \end{aligned}$$

This concludes the theorem. ∎

Corollary 4.1. *If $\omega_v = 1/M$ for $v = 1, 2, \ldots, M$, then, the minimum value of the objective function of SGR-MCCDA is $\left[d - \frac{1}{M}\left(\sum_{i=1}^{d} \delta_i + \gamma M \log M\right)\right]$, where δ_i is the i-th largest eigenvalue of the matrix $\sum_{v=1}^{M} X_v (S_w^v)^{-1} X_v^T - \beta M L_{\mathcal{G}'}$.*

Proof. From Theorem 4.1, the minimum value of $\mathcal{J}(\{W_v, \omega_v\}_{v=1}^M, S)$ is

$$\mathcal{J}^* = d - \sum_{i=1}^{d} \lambda_i + \gamma \sum_{i=1}^{M} \omega_v \log \omega_v \tag{4.35}$$

where λ_i is the i-th largest eigenvalue of the matrix C of (4.24). Substituting $\omega_v = \frac{1}{M}, \forall v$, into (4.24) and (4.35), the minimum value of $\mathcal{J}(\{W_v, \omega_v\}_{v=1}^M, S)$ becomes

$$\begin{aligned} \bar{\mathcal{J}} &= d - \sum_{i=1}^{d} \lambda_i + \gamma \sum_{i=1}^{M} \frac{1}{M} \log \frac{1}{M} \\ &= d - \sum_{i=1}^{d} \lambda_i - \gamma \log M \end{aligned} \tag{4.36}$$

where λ_i is the i-th largest eigenvalue of the matrix $\sum_{v=1}^M \frac{1}{M} X_v \left(S_w^v\right)^{-1} X_v^T - \beta L_{\mathcal{G}'}$. Now, if λ is an eigenvalue of a matrix $\frac{1}{M}A$, then $M\lambda$ is an eigenvalue of the matrix A. Hence, the expression (4.36) can be written as

$$\bar{\mathcal{J}} = \left[d - \frac{1}{M} \left(\sum_{i=1}^{d} \delta_i + \gamma M \log M \right) \right], \tag{4.37}$$

where $\delta_i = M\lambda_i$ is the i-th largest eigenvalue of the matrix $\sum_{v=1}^M X_v \left(S_w^v\right)^{-1} X_v^T - \beta M L_{\mathcal{G}'}$. ∎

Corollary 4.2. *If $\bar{\mathcal{J}}$ and $\mathcal{J}^*$ be the minimum values of the objective function $\mathcal{J}(\{W_v, \omega_v\}_{v=1}^M, S)$ of SGR-MCCDA for the equal and unequal weights, respectively, then*

$$\mathcal{J}^* - \bar{\mathcal{J}} > \frac{1}{M} \left[\sum_{v=1}^{M} \mathrm{Tr}(A_v) - M \sqrt{\sum_{v=1}^{M} \{\mathrm{Tr}(A_v)\}^2} \right] < 0,$$

where $A_v = X_v \left(S_w^v\right)^{-1} X_v^T$.

Proof. From Theorem 4.1, the minimum value of $\mathcal{J}(\{W_v, \omega_v\}_{v=1}^M, S)$ for unequal weights is $\mathcal{J}^*$, which is given in (4.35). From Corollary 4.1, the minimum value of $\mathcal{J}(\{W_v, \omega_v\}_{v=1}^M, S)$ for equal weights is $\bar{\mathcal{J}}$, which is given in (4.37). Subtracting (4.37) from (4.35), we get

$$\begin{aligned} \mathcal{J}^* - \bar{\mathcal{J}} &= \frac{1}{M} \sum_{i=1}^{d} \delta_i + \gamma \log M - \sum_{i=1}^{d} \lambda_i + \gamma \sum_{i=1}^{M} \omega_v \log \omega_v \\ &= \frac{1}{M} \left(\sum_{i=1}^{d} \delta_i - M \sum_{i=1}^{d} \lambda_i \right) + \gamma \left(\log M + \sum_{i=1}^{M} \omega_v \log \omega_v \right). \end{aligned} \tag{4.38}$$

Since $\gamma\left(\log M + \sum_{i=1}^{M} \omega_v \log \omega_v\right) \geqslant 0$, (4.38) reduces to

$$\mathcal{J}^* - \bar{\mathcal{J}} \geqslant \frac{1}{M}\left(\sum_{i=1}^{d} \delta_i - M \sum_{i=1}^{d} \lambda_i\right). \tag{4.39}$$

Using the relationships between the eigenvalues of a matrix and its trace, the right-hand side of (4.39) reduces to

$$\begin{aligned}
&\frac{1}{M} \operatorname{Tr}\left\{\sum_{v=1}^{M} A_v - \beta M L_{\mathcal{G}'}\right\} - \operatorname{Tr}\left\{\sum_{v=1}^{M} \omega_v A_v - \beta L_{\mathcal{G}'}\right\} \\
&= \frac{1}{M} \sum_{v=1}^{M} \operatorname{Tr}(A_v) - \beta \operatorname{Tr}(L_{\mathcal{G}'}) - \sum_{v=1}^{M} \omega_v \operatorname{Tr}(A_v) + \beta \operatorname{Tr}(L_{\mathcal{G}'}) \\
&= \frac{1}{M} \sum_{v=1}^{M} \operatorname{Tr}(A_v) - \sum_{v=1}^{M} \omega_v \operatorname{Tr}(A_v).
\end{aligned} \tag{4.40}$$

Applying Cauchy-Schwartz inequality on $\omega_v (\geqslant 0)$ and $\operatorname{Tr}(A_v) (\geqslant 0)$, $v = 1, 2, \ldots, M$, the following holds

$$\left\{\sum_{v=1}^{M} \omega_v \operatorname{Tr}(A_v)\right\}^2 \leqslant \sum_{v=1}^{M} {\omega_v}^2 \sum_{v=1}^{M} \{\operatorname{Tr}(A_v)\}^2. \tag{4.41}$$

Moreover,

$$\sum_{v=1}^{M} \omega_v = 1 \Rightarrow \sum_{v=1}^{M} {\omega_v}^2 \leqslant 1. \tag{4.42}$$

Combining (4.41) and (4.42), we get

$$\left\{\sum_{v=1}^{M} \omega_v \operatorname{Tr}(A_v)\right\} \leqslant \sqrt{\sum_{v=1}^{M} \{\operatorname{Tr}(A_v)\}^2}. \tag{4.43}$$

Using (4.39), (4.40), and (4.43), it can be observed that

$$\mathcal{J}^* - \bar{\mathcal{J}} \geqslant \frac{1}{M} \sum_{v=1}^{M} \operatorname{Tr}(A_v) - \sqrt{\sum_{v=1}^{M} \{\operatorname{Tr}(A_v)\}^2}. \tag{4.44}$$

Again,

$$\left(\sum_{v=1}^{M}\mathrm{Tr}(A_v)\right)^2 < \sum_{v=1}^{M}\{\mathrm{Tr}(A_v)\}^2$$
$$\Rightarrow \left(\sum_{v=1}^{M}\mathrm{Tr}(A_v)\right)^2 < M^2\sum_{v=1}^{M}\{\mathrm{Tr}(A_v)\}^2 \text{ (since } M>0)$$
$$\Rightarrow \sum_{v=1}^{M}\mathrm{Tr}(A_v) < M\sqrt{\sum_{v=1}^{M}\{\mathrm{Tr}(A_v)\}^2}. \tag{4.45}$$

Therefore, using (4.45), (4.44) becomes

$$\mathcal{J}^* - \bar{\mathcal{J}} \geqslant \frac{1}{M}\left[\sum_{v=1}^{M}\mathrm{Tr}(A_v) - M\sqrt{\sum_{v=1}^{M}\{\mathrm{Tr}(A_v)\}^2}\right] < 0$$

■

4.4.2 Convergence Analysis

Theorem 4.2. *The iterative rule described in Algorithm 4.2 is guaranteed to converge.*

Proof. Recall the objective function of the SGR-MCCDA, $\mathcal{J}(\{W_v,\omega_v\}_{v=1}^M, S)$, and assume that W_v^t, ω_v^t and S^t are the values of W_v, ω_v and S derived at the t-th iteration. To prove the convergence of the proposed algorithm, the following two conditions must hold:

- the objective function is bounded below, that is, $\mathcal{J}(\{W_v,\omega_v\}_{v=1}^M, S) \geqslant \mathcal{K}$, where $\mathcal{K}$ is some constant;
- the objective function decreases monotonically with the iteration, that is,

$$\mathcal{J}(\{W_v^{t+1},\omega_v^{t+1}\}_{v=1}^M, S^{t+1}) \leqslant \mathcal{J}(\{W_v^t,\omega_v^t\}_{v=1}^M, S^t)$$

.

From the Corollary (4.2), it can be written that

$$\mathcal{J}^* \geqslant \bar{\mathcal{J}} - \sqrt{\sum_{v=1}^{M}\{\mathrm{Tr}(A_v)\}^2} + \frac{1}{M}\sum_{v=1}^{M}\mathrm{Tr}(A_v).$$

Assuming $\mathcal{K} = \bar{\mathcal{J}} - \sqrt{\sum_{v=1}^{M}\{\mathrm{Tr}(A_v)\}^2} + \frac{1}{M}\sum_{v=1}^{M}\mathrm{Tr}(A_v)$, it can be concluded that the objective function $\mathcal{J}(\{W_v,\omega_v\}_{v=1}^M, S)$ is bounded below.

Now, at the $(t+1)$-th iteration, keeping $\{\omega_v^t\}_{v=1}^M$ fixed, the update rules (4.22) and (4.23) together imply that the optimization problem (4.18) can be considered as the subproblem given by

$$\begin{aligned} &\max_{S} \ \operatorname{Tr}\left[S^T\left\{\sum_{v=1}^{M}\omega_v^t X_v(S_w^v)^{-1}X_v^T - \beta L_{\mathcal{G}'}\right\}S\right] \\ &\text{subject to } S^TS = I, \end{aligned} \tag{4.46}$$

where S can be obtained from the eigendecomposition of the matrix $\sum_{v=1}^{M}\omega_v^t X_v(S_w^v)^{-1}X_v^T - \beta L_{\mathcal{G}'}$ and W_v can be obtained from (4.22). Therefore, it can be written that

$$(\{W_v^{t+1}\}_{v=1}^M, S^{t+1}) = \underset{\{W_v\}_{v=1}^M, S}{\arg\min}\ \mathcal{J}(\{W_v\}_{v=1}^M, S, \{\omega_v^t\}_{v=1}^M);$$

which suggests that

$$\mathcal{J}(\{W_v^{t+1}\}_{v=1}^M, S^{t+1}, \{\omega_v^t\}_{v=1}^M) \leqslant \mathcal{J}(\{W_v^t, \omega_v^t\}_{v=1}^M, S^t). \tag{4.47}$$

Once $\{W_v^{t+1}\}_{v=1}^M$ and S^{t+1} are obtained, $\{\omega_v^{t+1}\}_{v=1}^M$ will satisfy the following minimization problem:

$$\{\omega_v^{t+1}\}_{v=1}^M = \underset{\{\omega_v\}_{v=1}^M}{\arg\min}\ \mathcal{J}(\{W_v^{t+1}, \omega_v\}_{v=1}^M, S^{t+1}). \tag{4.48}$$

Thus,

$$\mathcal{J}(\{W_v^{t+1}, \omega_v^{t+1}\}_{v=1}^M, S^{t+1}) \leqslant \mathcal{J}(\{W_v^{t+1}, \omega_v^t\}_{v=1}^M, S^{t+1}). \tag{4.49}$$

Combining (4.47) and (4.49), it can be concluded that

$$\mathcal{J}(\{W_v^{t+1}, \omega_v^{t+1}\}_{v=1}^M, S^{t+1}) \leqslant \mathcal{J}(\{W_v^t, \omega_v^t\}_{v=1}^M, S^t), \tag{4.50}$$

which indicates that the second condition holds true. Therefore, the convergence of the Algorithm 4.2 is guaranteed. ■

4.4.3 Computational Complexity Analysis

This section briefly presents the computational complexity of the proposed SGR-MCCDA algorithm. For convenience, assume that $p = \max_v\{d_v\}$, where d_v is the dimension of the v-th view X_v. The computational cost for executing each step of Algorithm **??** is outlined below.

1. The time cost for the initialization step is $\mathcal{O}(M)$.

2. The computational cost to construct the graph Laplacian $L_{\mathcal{G}'}$ is $\mathcal{O}(n^2)$.

3. In Step 5 of the Algorithm 4.2, the cost for computing the class scatter matrices $\{S_w^v\}_{v=1}^M$ is $\mathcal{O}(Mnp^2)$ and the inverse calculation takes $\mathcal{O}(Mp^3)$ time. To compute the $n \times n$ matrix C, the time complexity is $\mathcal{O}(M(p^2n+n^2p+p^3))+\mathcal{O}(n^2)$. Therefore, the total computational cost for this step is $\mathcal{O}(Mnp^2+Mp^3+Mp^2n+Mn^2p+Mp^3+n^2) \approx \mathcal{O}(p^3+n^2)$, since $M << p, n$.

4. The time complexity of computing d eigenpairs of the matrix C is $\mathcal{O}(dn^2)$.

5. The computational cost for updating $\{W_v\}_{v=1}^M$ is $\mathcal{O}(M(p^3+p^2n+pnd)) \approx \mathcal{O}(p^3+p^2n)$.

6. The computational cost for updating $\{\omega_v\}_{v=1}^M$ is $\mathcal{O}(M(pnd+d^2p+dp^2+d)) \approx \mathcal{O}(d^2p+dp^2)$.

Therefore, the total computational complexity of the proposed algorithm is $\mathcal{O}(M + n^2 + \tau(p^3 + n^2 + dn^2 + p^3 + p^2n + d^2p + dp^2)) \approx \mathcal{O}(\tau(p^3 + p^2n + dn^2))$, where τ is the number of iterations required to converge the iterative procedure. The complexity is expressed in terms of the sample size n, the feature size p of the largest view, and the reduced dimension d. The complexity can also be written as $\mathcal{O}(p^3)$ or $\mathcal{O}(dn^2)$ depending on the condition of $n << p$ or $p << n$. Note that, for the case when feature size p is very high compared to the sample size n, the eigendecomposition in Step 4, which is one of the time consuming step, takes $\mathcal{O}(dn^2)$ time cost. This makes the algorithm efficient for "large p small n" problem.

4.4.4 Comparison of CSP-MvCDA and SGR-MCCDA

This section presents the statistical difference between CSP-MvCDA, proposed in Chapter 3, and SGR-MCCDA, proposed in this chapter. The SGR-MCCDA is different from CSP-MVCDA in terms of the following three aspects.

- The regularizer term, included in the objective function of CSP-MvCDA, captures the local class-structure of the data and aggregates class-geometry from each of the views. On the other hand, the regularizer term, included in SGR-MCCDA, depends only on the class-geometry of the common source space.

- The CSP-MvCDA algorithm considers equal relevance of each view while constructing the common discriminative latent space, whereas the SGR-MCCDA computes the relevance of each view for constructing the common discriminative latent space.

- The SUMCOR criterion of MCCA forms the basis of CSP-MvCDA, while the MAX-VAR criterion of MCCA forms the basis of SGR-MCCDA. When the sample size is very small compared to the feature size, the execution time of SGR-MCCDA is significantly reduced. On the other hand, when the feature size is smaller than the sample size, the CSP-MvCDA executes faster than the SGR-MCCDA.

Table 4.1: Descriptions of Data Sets Used in the Study

Data Sets	Sample	Feature Type(Feature Size)	Class
CRC	261	ω_1: RNA Sequence (13465) ω_2: Protein expression(222) ω_3: miRNA Sequence(236) ω_4: DNA methylation (293526)	2
GBM	213	ω_1: Gene expression (12042) ω_2: miRNA expression (534) ω_3: DNA methylation (21422) ω_4: Copy number segmentation (4070)	5
Reuters	18758	ω_1: English (21531) ω_2: French (24892) ω_3: German (34251) ω_4: Italian (15506) ω_5: Spanish (11547)	6
Caltech-20	2386	ω_1: Gabor (48) ω_2: Wavelet moments (40) ω_3: Cenhist (254) ω_4: Hog (1984) ω_5: Gist (512) ω_6: Local binary patterns (LBP) (928)	20
100Leaves	1600	ω_1: Shape descriptor (64) ω_2: Fine scale margin (64) ω_3: Texture histogram (64)	100

4.5 Experimental Results and Discussion

This section describes the experimental study comprehensively and analyzes the performance of the proposed method, called SGR-MCCDA, with respect to different state-of-the-art multi-view classification algorithms, namely, MAXVAR-MCCA [129], GMA [222], MULDA [239], MvDA [124], MvDA-VC [124], MvCCDA [292], ReDMiCA [164], GMCCA [43], DMvTSVM-AE [278], and SMDDRL [116]. Among these methods, the DMvTSVM-AE and SMDDRL are deep learning algorithms. The experiments are conducted on five cancer data sets and eight benchmark data sets. The cancer data sets are obtained from TCGA. Three cancer data sets, namely, BRCA, CESC, and LGG are already described in Chapter 3, and two new cancer data sets, namely, colorectal carcinoma (CRC) and glioblastoma multiforme (GBM) are added in this chapter to conduct the experiments. On the other hand, out of eight benchmark data sets, four data sets, namely, CiteSeer, Cora, NW-Object, Caltech-101, and ALOI, are already described in Chapter 3, and the three new

data sets, namely, Reuters[1] (multilingual document data set), Caltech-20[2] (image based object data set) and 100Leaves[3] (one-hundred plant species leaves data set) are added in this chapter for the experiments. A brief description of the newly added data sets are given in Table 4.1 and their detailed description can be found in Appendix A.

Similar to the experimental setup of Chapter 3, both the train-test and 10-fold cross validation (CV) setup are used to conduct the experiment. To compare the performance of all the methods, top 25 features are extracted except MvCCDA and two deep learning algorithms, namely, SMDDRL and DMvTSVM-AE. The MvCCDA effectively generates c number of features, where c is number of classes present in the data. For these three methods, the optimal number of features extracted by them, are used to report the classification accuracy. The support vector machine (SVM) with linear kernels is used to compute the classification accuracy. Two significance tests, namely, paired sample t-test and Wilcoxon's signed rank test, are carried out, and the associated p-value is presented, to determine the statistical significance of the proposed technique in 10-fold CV setup.

Table 4.2: Parameter Analysis of SGR-MCCDA on Different Data Sets

	Data	Proposed Parameter Optimization			Exhaustive Search	
	Sets	(γ,β,α)	TrainAcc	TestAcc	TestAcc	(γ,β,α)
10-fold CV	BRCA	(10, 0.7, 0.9)	0.9917	**0.8375**	**0.8375**	(10, 0.7, 0.9)
	CESC	(10, 0.5, 0.1)	0.9843	0.8166	**0.8671**	(100, 0.5, 0.3)
	CRC	(1, 0.7, 0.8)	0.9930	0.9148	**0.9259**	(0.1, 0.7, 0.8)
	LGG	(10, 0.5, 0.1)	0.9942	**0.9842**	**0.9842**	(10, 0.5, 0.1)
	GBM	(100, 0.001, 0.2)	0.9375	**0.8417**	**0.8417**	(100, 0.001, 0.2)
	CiteSeer	(10, 0.01, 0.1)	0.7877	**0.6897**	**0.6897**	(10, 0.01, 0.1)
	CORA	(10, 0.3, 0.1)	0.9278	0.7875	**0.7915**	(10, 1, 0.1)
	Reuters	(100, 1, 0.1)	0.8944	0.8764	**0.8789**	(100, 1, 0.2)
	Caltech-101	(10, 0.0001, 0.1)	0.9531	**0.8201**	**0.8201**	(10, 0.0001, 0.1)
	Caltech-20	(100, 0.001, 0.5)	1.0000	**0.9672**	**0.9672**	(100, 0.001, 0.5)
	NW-Object	(10,0.001,0.2)	0.4561	0.4387	**0.4449**	(10,0.001,0.4)
	ALOI	(10,0.0001,0.4)	0.9998	**0.9933**	**0.9933**	(10,0.0001,0.4)
	100Leaves	(10, 0.1, 0.1)	1.0000	**0.9840**	**0.9840**	(10, 0.1, 0.1)
Train-Test	BRCA	(10, 0.5, 0.8)	1.0000	**0.7273**	**0.7273**	(10, 0.5, 0.8)
	CESC	(100, 0.001, 0.4)	0.9615	0.7884	**0.8216**	(100, 0.01, 0.2)
	CRC	(100, 0.3, 0.8)	0.9924	**0.8692**	**0.8692**	(100, 0.3, 0.8)
	LGG	(10, 0.001, 0.1)	0.9947	**0.9785**	**0.9785**	(10, 0.001, 0.1)
	GBM	(100, 0.7, 0.9)	0.9815	0.7810	**0.7920**	(100, 0.5, 0.9)
	CiteSeer	(100,0.01,0.1)	0.8032	**0.6848**	**0.6848**	(100,0.01,0.1)
	CORA	(100, 0.001, 0.3)	0.9469	**0.7647**	**0.7647**	(100, 0.001, 0.3)
	Reuters	(10, 0.7, 0.1)	0.9496	**0.8655**	**0.8655**	(10, 0.7, 0.1)
	Caltech-101	(100,0.001,0.1)	0.9538	**0.6160**	**0.6160**	(100,0.001,0.1)
	Caltech-20	(100, 0.001, 0.3)	1.0000	**0.9544**	**0.9544**	(100, 0.001, 0.3)
	NW-Object	(100,0.001,0.4)	0.4432	0.4091	**0.4217**	(100,0.001,0.3)
	ALOI	(100, 0.01,0.1)	0.9982	0.9712	**0.9842**	(100, 0.001,0.1)
	100Leaves	(100, 0.0001, 0.5)	1.0000	**0.9713**	**0.9713**	(100, 0.0001, 0.5)

[1] https://github.com/yeqinglee/mvdata
[2] https://github.com/yeqinglee/mvdata
[3] https://archive.ics.uci.edu/ml/datasets/One-hundred+plant+species+leaves+data+set

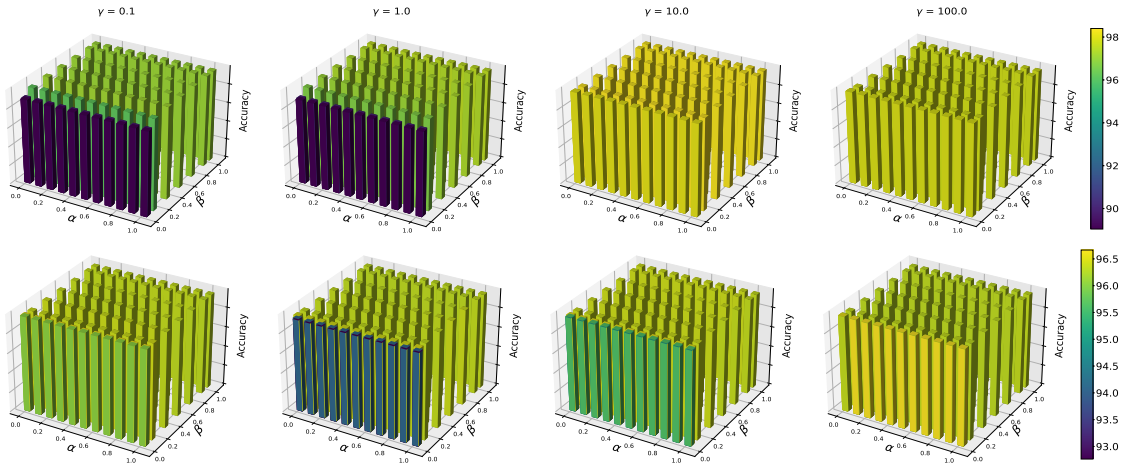


Figure 4.2: 3D bar plot of classification accuracy on the test set over 10-fold cross-validation for the parameters α, β and γ. The lighter shade denotes the better classification accuracy. Top row: 100Leaves; bottom row: Caltech-20.

4.5.1 Parameter Analysis

The proposed SGR-MCCDA model has three parameters, namely, α, γ and β, which are required to be tuned. A grid search strategy is followed to find out the optimal parameters. The parameter α signifies the balance between the within-class and between-class graph information of the data, and is varied from 0.0 to 1.0 having step size of 0.1. The parameter β emphasizes the contribution of the term consisting of structural information into the performance of the proposed model. The value of β belongs to the set $\{0.0001, 0.001, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$. The entropy term responsible for obtaining suitable weight distribution among multiple views is controlled by the parameter γ. The parameter γ is varied in a log-scale and taken from the set $\{0.1, 1.0, 10.0, 100.0\}$, because the weights can increase or decrease exponentially according to (4.29). The optimal parameter is chosen based on best classification accuracy obtained over the training set. The optimal parameters along with the corresponding training accuracy (TrainAcc) and test accuracy (TestAcc) are reported in Table 4.2 for both the experimental setup. The classification accuracy of the test set based on the exhaustive search procedure is also reported in Table 4.2 under the column "ExhaustiveAcc" to evaluate the effectiveness of optimal parameters. The results in Table 4.2 show that the difference between the test accuracy obtained for the optimal parameters and that obtained by exhaustive search process is either zero or not significant enough for most of the cases.

To analyze the sensitivity of the parameters α, β, and γ, the variation of the classification accuracy over the test set of 10-fold CV for these parameters, is presented by $3D$-bar plot in Fig. 4.2. The figure describes that the parameter γ is sensitive to the choice of values in the range of $\{0.1, 1.0, 10.0, 100.0\}$, as the variation of the classification accuracy is comparatively high with respect to the parameter γ. This is justified because the relevance of each view for constructing common latent space is controlled by the parameter γ. On the other hand for a particular value of γ, the classification accuracy does not differ much for the parameters α and β in the range of 0.1 to 1.0. But, it is also worth noting that the

classification accuracy varies a lot if the values of β change in the range of 0.0001 to 0.1. This signifies that the model is less sensitive to the parameters α and β in the range of 0.1 to 1.0 but sensitive to β in the range of 0.0001 to 0.1.

4.5.2 Ablation Study

The proposed SGR-MCCDA method is evaluated under different conditions to justify the role of including graph knowledge of the estimated source vectors and learning proper weight for each view instead of considering equal weights. The ablation study indicating the classification performance of the proposed method over the five cancer data sets and eight benchmark data sets is given in Table 4.3. The results show that when $\beta \neq 0$, that is, when the structural geometry is incorporated into the proposed model, the classification performance of the model increases compared to the model, MCCDA, when $\beta = 0$. Again, the proposed model ($\beta \neq 0$) is examined under two different scenarios: first, equal weights are considered for all the views; second, an appropriate weight for each view is learned while integrating multiple views. From Table 4.3, it is clear that the model, SGR-MCCDA, where the weight for each view is learned iteratively, obtain the best classification accuracy for all the thirteen data sets except for GBM and NW-Object in train-test case. It is also noticeable that for LGG, the performance is same across all different conditions, which suggests that incorporating graph knowledge might not improve the performance but did not degrade the performance as well. The weights learned by the proposed method is given in Table 4.4.

4.5.3 Comparative Performance Analysis

In this section, the classification performance of the proposed method is compared with that of different state-of-the-art algorithms on real cancer and benchmark data sets.

4.5.3.1 Classification Results on Omics Data Sets

To show the efficacy of the proposed model, the classification accuracy for both the 10-fold CV and train-test setup is presented in Table 4.5 for five cancer data sets, namely, BRCA, CESC, CRC, LGG and GBM. The mean, median and standard deviation are reported for 10-fold CV and the p-values corresponding to paired t-test and Wilcoxon signed rank test are presented to prove the statistical significance of these results. From Table 4.5, it is clear that the proposed SGR-MCCDA has achieved highest classification accuracy on all the five cancer data sets for both the experimental setup. For BRCA and CESC, where most of the existing methods have very low classification score, the proposed model has obtained almost 73% and 79% in train-test setup and 84% and 82% in case of 10-fold CV, respectively. The classification performance for LGG data set is also similar. The results signify that considering both the shared correlated structure of multiple views and intra-view discrimination to obtain the common space is more useful than considering only the intra-view and inter-view discrimination, which have been used in MvDA, MvDA-VC and MvCCDA. Taking only the intra-view and inter-view discrimination, and ignoring the correlation among multiple views may encounter undesired information, which may degrade the performance. In case of CRC data set, most of the existing models have performed

Table 4.3: Ablation Study of SGR-MCCDA on Different Data Sets

Data Sets	MCCDA ($\beta = 0$)			
	Accuracy (Train-Test)	Accuracy (10-fold)		
		Mean	Median	StdDev
BRCA	0.697	0.800	0.813	0.104
CESC	0.692	0.750	0.750	0.083
CRC	0.862	0.904	0.914	0.050
LGG	0.968	**0.984**	**0.974**	0.013
GBM	**0.800**	0.804	0.792	0.053
CiteSeer	0.647	0.650	0.653	0.016
Cora	0.740	0.760	0.762	0.018
Reuters	0.856	0.853	0.851	0.009
Caltech-101	0.544	0.808	0.800	0.014
Caltech-20	0.886	0.903	0.901	0.005
NW-Object	**0.419**	0.422	0.422	0.012
ALOI	0.938	0.972	0.974	0.006
100Leaves	0.956	0.967	0.978	0.008

Data Sets	SGR-MCCDA ($\beta \neq 0$)							
	Equal weight				Proposed weight			
	Accuracy (Train-Test)	Accuracy (10-fold)			Accuracy (Train-Test)	Accuracy (10-fold)		
		Mean	Median	StdDev		Mean	Median	StdDev
BRCA	0.712	0.831	0.844	0.089	**0.727**	**0.838**	**0.844**	0.085
CESC	**0.788**	76.667	75.000	8.975	**0.788**	**0.817**	**0.833**	0.090
CRC	0.869	0.911	**0.926**	0.038	**0.877**	**0.915**	**0.926**	0.041
LGG	0.973	**0.984**	**0.974**	0.013	**0.978**	**0.984**	**0.974**	0.013
GBM	0.771	0.825	0.813	0.486	0.781	**0.842**	**0.833**	0.039
CiteSeer	0.683	0.683	0.686	0.028	**0.685**	**0.690**	**0.695**	0.022
Cora	0.764	0.771	0.777	0.023	**0.765**	**0.785**	**0.789**	0.020
Reuters	0.864	0.874	0.874	0.008	**0.866**	**0.876**	**0.874**	0.008
Caltech-101	0.606	0.812	0.812	0.012	**0.616**	**0.820**	**0.820**	0.009
Caltech-20	0.942	0.956	0.957	0.010	**0.954**	**0.967**	**0.967**	0.005
NW-Object	**0.419**	0.431	0.432	0.009	0.409	**0.439**	**0.438**	0.007
ALOI	0.968	0.985	0.985	0.003	**0.971**	**0.993**	**0.993**	0.003
100Leaves	0.961	0.981	0.983	0.006	**0.973**	**0.984**	**0.985**	0.009

well. It happens probably because of the less number of classes ($K = 2$) present in the CRC data. Among all the existing methods, the deep learning method, DMvTSVM-AE, has noticeable performance in all the four cancer data sets. In fact, for BRCA and CRC, it obtains second best classification accuracy for both the train-test and 10-fold CV setup. The ReDMiCA has attained second best classification score on LGG data for the train-test set up. From the table, it can be seen that all the p-values, except for the MvDA-VC and DMvTSVM-AE, are significant. The MvDA-VC and DMvTSVM-AE models correspond non-significant p-value for CRC and the DMvTSVM-AE gives non-significant p-value for GBM with respect to the proposed model. The non-significant p-value indicates that the

Table 4.4: Proposed View Weights for SGR-MCCDA

Data sets	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
BRCA	0.2554	0.2426	0.2588	0.2432	-	-
CESC	0.2519	0.2477	0.2517	0.2487	-	-
CRC	0.2610	0.2387	0.2380	0.2623	-	-
LGG	0.2083	0.1989	0.2058	0.1789	0.2081	-
GBM	0.2508	0.2491	0.2424	0.2577	-	-
CiteSeer	0.2476	0.2498	0.2500	0.2526	-	-
Cora	0.2498	0.2544	0.2508	0.2449	-	-
Reuters	0.1977	0.1868	0.2198	0.1852	0.2105	-
Caltech-101	0.1736	0.1731	0.1729	0.1502	0.1675	0.1627
Caltech-20	0.1879	0.1863	0.1827	0.1268	0.1632	0.1530
NW-Object	0.2000	0.2001	0.2002	0.1998	0.1999	-
ALOI	0.2591	0.2539	0.2412	0.2458	-	-
100Leaves	0.2795	0.3604	0.3601	-	-	-

two models have similar distribution of classification accuracy over 10-fold CV.

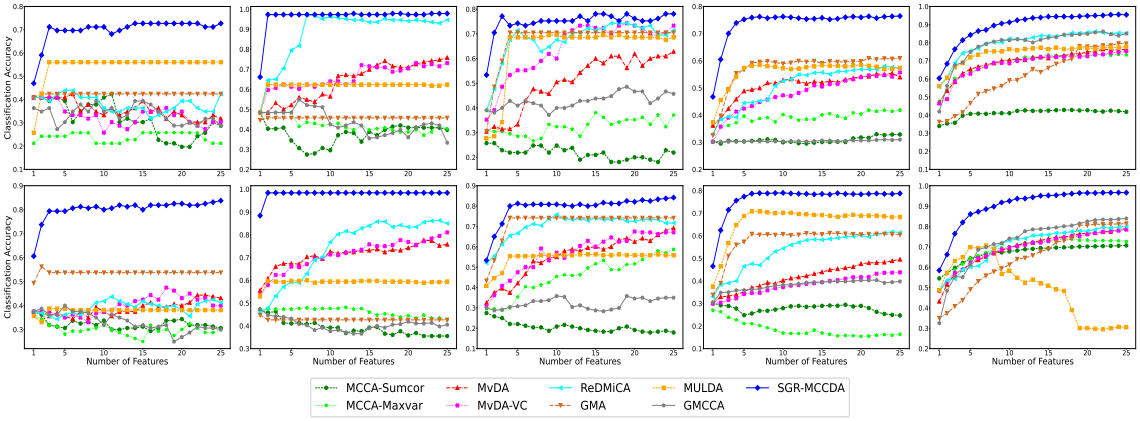


Figure 4.3: Variation of classification accuracy with respect to the number of extracted features for the proposed (SGR-MCCDA) and existing methods on Omics and benchmark data sets (top-row: training-testing, bottom-row: 10-fold CV; left-to-right: BRCA, LGG, GBM, Cora, Caltech-20).

Fig. 4.3 shows the variation of classification accuracy with respect to the number of extracted features for the proposed as well as existing methods. The classical method MvCCDA and two deep learning methods, SMDDRL and DMvTSVM-AE, are excluded from this analysis. The feature versus accuracy curves are shown for three cancer data sets, namely, BRCA, LGG and GBM. Note that the mean classification accuracy is taken to plot the curves for 10-fold CV. Also, the scatter plots for the top two extracted features are given in Fig. 4.4 for LGG data. From Fig. 4.3, it is clear that the proposed model has an increasing curve of classification accuracy with respect to the number of extracted features for both the experimental setup. Moreover, it is worth noting that as the number of features approaching to 25, the curve is getting flat indicating the saturation in the

Table 4.5: Classification Accuracy of Proposed Method (SGR-MCCDA) and Different Algorithms on Cancer Data Sets

Different Algorithms	Data Sets	Accuracy (Train-Test)	Accuracy and Significance Analysis for 10-fold CV					Time (Sec.)
			Mean	Median	StdDev	Paired-t:p	Wilcoxon:p	
MCCA-SUMCOR [106]	BRCA	0.288	0.306	0.319	0.055	**5.50E-08**	**1.95E-03**	31.2
MCCA-MAXVAR [129]		0.212	0.300	0.300	0.097	**3.97E-07**	**1.95E-03**	18.6
GMA [222]		0.424	0.538	0.531	0.109	**4.89E-04**	**3.91E-03**	33.0
MvDA [124]		0.318	0.431	0.375	0.100	**2.08E-05**	**1.95E-03**	14.1
MvDA-VC [124]		0.303	0.400	0.400	0.079	**2.92E-06**	**1.95E-03**	15.2
MULDA [239]		0.561	0.381	0.406	0.123	**1.69E-05**	**1.95E-03**	76.5
MvCCDA [292]		0.455	0.569	0.625	0.095	**2.72E-04**	**1.95E-03**	0.9
ReDMiCA [164]		0.424	0.413	0.400	0.119	**2.34E-06**	**1.95E-03**	5774.5
GMCCA [43]		0.288	0.300	0.319	0.131	**8.47E-06**	**1.95E-03**	49.4
DMvTSVM-AE [278]		0.692	0.615	0.615	0.843	**9.44E-04**	**5.86E-03**	147.7
SMDDRL [116]		0.373	0.474	0.481	0.086	**9.23E-07**	**1.95E-03**	1.8
SGR-MCCDA		**0.727**	**0.838**	**0.844**	0.085	-	-	2.4
MCCA-SUMCOR [106]	CESC	0.385	0.458	0.458	0.137	**3.51E-08**	**1.95E-03**	6.0
MCCA-MAXVAR [129]		0.385	0.300	0.333	0.143	**3.97E-07**	**1.95E-03**	15.9
GMA [222]		0.462	0.558	0.625	0.124	**1.29E-05**	**5.32E-03**	10.4
MvDA [124]		0.423	0.467	0.492	0.153	**2.71E-05**	**1.95E-03**	15.5
MvDA-VC [124]		0.404	0.500	0.467	0.142	**3.29E-06**	**1.95E-03**	21.9
MULDA [239]		0.769	0.483	0.417	0.170	**1.09E-05**	**1.95E-03**	50.7
MvCCDA [292]		0.712	0.833	0.833	0.091	**3.51E-04**	**1.95E-03**	0.4
ReDMiCA [164]		0.500	0.567	0.608	0.203	**1.24E-07**	**1.95E-03**	5162.4
GMCCA [43]		0.423	0.492	0.492	0.144	**5.47E-06**	**1.95E-03**	51.4
DmvTSVM-AE [278]		0.650	0.730	0.700	0.078	**7.21E-04**	**2.16E-04**	147.4
SMDDRL [116]		0.423	0.390	0.381	0.079	**7.13E-08**	**1.95E-03**	1.8
SGR-MCCDA		**0.788**	**0.817**	**0.833**	0.090	-	-	2.3
MCCA-SUMCOR [106]	CRC	0.738	0.607	0.741	0.104	**1.70E-07**	**1.95E-03**	14.7
MCCA-MAXVAR [129]		0.600	0.704	0.737	0.080	**7.54E-06**	**1.95E-03**	16.0
GMA [222]		0.792	0.800	0.796	0.058	**6.22E-04**	**1.95E-03**	32.8
MvDA [124]		0.808	0.848	0.852	0.039	**7.25E-04**	**7.42E-03**	6.2
MvDA-VC [124]		0.831	0.867	0.889	0.060	*8.98E-02*	*1.05E-01*	6.3
MULDA [239]		0.746	0.796	0.796	0.053	**4.33E-04**	**3.91E-03**	635.5
MvCCDA [292]		0.838	0.848	0.852	0.078	**3.86E-03**	**1.12E-02**	0.6
ReDMiCA [164]		0.800	0.878	0.874	0.068	**3.19E-02**	**4.07E-02**	5285.3
GMCCA [43]		0.838	0.785	0.770	0.076	**6.55E-03**	**3.91E-03**	50.5
DMvTSVM-AE [278]		0.865	0.892	0.887	0.037	*1.61E-01*	*2.32E-01*	66.909
SMDDRL [116]		0.733	0.772	0.766	0.027	**1.20E-05**	**1.95E-03**	2.509
SGR-MCCDA		**0.869**	**0.915**	**0.926**	0.041	-	-	7.1
MCCA-SUMCOR [106]	LGG	0.398	0.355	0.342	0.077	**1.96E-09**	**1.95E-03**	14.2
MCCA-MAXVAR [129]		0.403	0.426	0.474	0.072	**2.75E-09**	**1.95E-03**	48.1
GMA [222]		0.457	0.426	0.447	0.082	**3.52E-09**	**1.95E-03**	135.3
MvDA [124]		0.758	0.758	0.763	0.080	**9.52E-06**	**1.95E-03**	16.4
MvDA-VC [124]		0.731	0.811	0.789	0.078	**5.29E-05**	**1.95E-03**	16.9
MULDA [239]		0.624	0.595	0.592	0.090	**4.98E-07**	**1.95E-03**	1183.6
MvCCDA [292]		0.694	0.766	0.776	0.057	**1.23E-06**	**1.95E-03**	0.7
ReDMiCA [164]		0.946	0.850	0.842	0.035	**5.31E-07**	**1.95E-03**	5958.2
GMCCA [43]		0.333	0.405	0.408	0.087	**1.72E-09**	**1.95E-03**	143.1
DMvTSVM-AE [278]		0.743	0.746	0.757	0.052	**3.16E-07**	**1.95E-03**	143.4
SMDDRL [116]		0.861	0.877	0.900	0.055	**1.91E-04**	**1.95E-03**	4.3
SGR-MCCDA		**0.978**	**0.984**	**0.974**	0.013	-	-	3.6
MCCA-SUMCOR [106]	GBM	0.219	0.179	0.200	0.065	**1.20E-08**	**1.95E-03**	25.6
MCCA-MAXVAR [129]		0.371	0.588	0.463	0.101	**9.80E-06**	**1.95E-03**	15.7
GMA [222]		0.705	0.754	0.750	0.044	**1.64E-03**	**7.63E-03**	137.8
MvDA [124]		0.629	0.692	0.588	0.077	**7.38E-03**	**7.70E-03**	6.3
MvDA-VC [124]		0.733	0.671	0.600	0.099	**8.07E-05**	**1.95E-03**	6.6
MULDA [239]		0.686	0.558	0.563	0.088	**1.56E-05**	**1.95E-03**	265.5
MvCCDA [292]		0.543	0.704	0.708	0.071	**4.56E-04**	**1.95E-03**	0.4
ReDMiCA [164]		0.714	0.717	0.729	0.047	**3.42E-04**	**7.58E-03**	2150.7
GMCCA [43]		0.457	0.350	0.313	0.102	**2.18E-06**	**1.95E-03**	44.5
DMvTSVM-AE [278]		0.655	0.795	0.810	0.064	*7.12E-02*	*8.40E-02*	68.861
SMDDRL [116]		0.262	0.370	0.359	0.071	**1.47E-07**	**1.95E-03**	2.626
SGR-MCCDA		**0.781**	**0.842**	**0.833**	0.039	-	-	7.240

performance. Most of the existing methods are not consistent with their performance and have a decreasing nature with respect to the number of extracted features. Among the existing algorithm, the curve of ReDMiCA is comparatively better in case of LGG; the curve of GMA is better and consistent for GBM in both the setup. In Fig. 4.4, the scatter plot of LGG shows that the top two extracted features of the proposed method have almost isolated the three classes of LGG, whereas none of the existing algorithms are capable of doing it. The ReMiCA and MvDA-VC are able to separate one class of the LGG data. The scatter plot signifies that the proposed model forms the common space by considering the graph structure of classes, which is reflected in the picture.

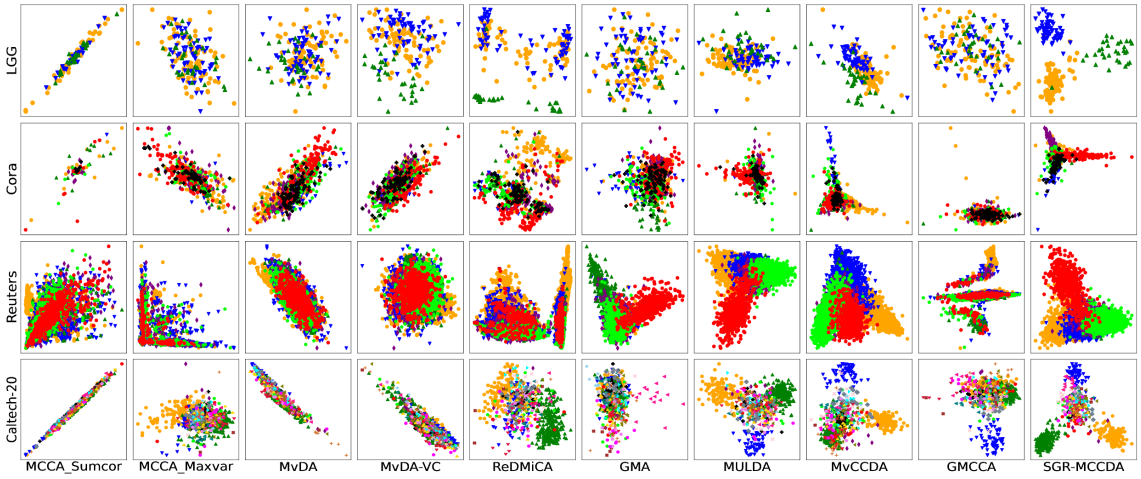


Figure 4.4: Scatter plots for the proposed and different existing methods, based on the first two extracted features.

4.5.3.2 Classification Results on Benchmark Data Sets

To prove the diverse potency of the SGR-MCCDA method, eight benchmark data set, namely, CiteSeer, Cora, Caltech-20, Caltech-101, Reuters, ALOI and 100Leaves with different characteristics have been used in the study. The Reuters and NW-Object are data sets with large number of samples and features. The 100leaves and ALOI have less number of feature than the samples but consists of 100 classes. The Cora and Caltech-20 have 6 and 20 number of classes, respectively, with a reasonable sample size. The Caltech-101 has large number of samples as well as large number of classes. From Table 4.6 and 4.7, it is seen that the proposed method obtains highest classification accuracy over all the eight benchmark data sets, except Reuters, Caltech-101, and ALOI, for both the train-test and 10-fold CV setup. In case of Reuters, the deep learning model SMDDRL has the best accuracy in train-test case. The MvDA and MvDA-VC have performed better than the proposed method for the data set Caltech-101 in 10-fold CV setup. In case of ALOI. the MvDA-VC has obtained height classification accuracy in train-set setup. In all these cases, the proposed method obtains the second best accuracy and has no significant difference. The results suggest that the proposed method can deal with the large as well as small data sets. Moreover, the results on 100Leaves and ALOI data sets prove that the proposed method can distinguish between large number of classes with very low number of features.

Table 4.6: Classification Accuracy of Proposed Method (SGR-MCCDA) and Different Algorithms on Benchmark Data Sets

Different Algorithms	Data Sets	Accuracy (Train-Test)	Accuracy and Significance Analysis for 10-fold CV					Time (Sec.)
			Mean	Median	StdDev	Paired-t:p	Wilcoxon:p	
MCCA-SUMCOR [106]	Cora	0.329	0.247	0.256	0.073	**5.42E-09**	**1.95E-03**	6.1
MCCA-MAXVAR [129]		0.418	0.164	0.227	0.027	**1.50E-14**	**1.95E-03**	3.6
GMA [222]		0.609	0.604	0.608	0.024	**3.53E-08**	**1.95E-03**	57.4
MvDA [124]		0.539	0.495	0.504	0.029	**9.27E-09**	**1.95E-03**	71.0
MvDA-VC [124]		0.557	0.438	0.441	0.020	**1.70E-10**	**1.95E-03**	70.1
MULDA [239]		0.574	0.683	0.687	0.013	**4.92E-08**	**1.95E-03**	31.1
MvCCDA [292]		0.716	0.748	0.753	0.038	**8.92E-03**	**9.77E-03**	15.0
ReDMiCA [164]		0.573	0.615	0.610	0.035	**1.39E-07**	**1.95E-03**	1286.8
GMCCA [43]		0.310	0.397	0.405	0.030	**2.31E-11**	**1.95E-03**	6.8
DMvTSVM-AE [278]		0.704	0.755	0.754	0.023	**8.15E-04**	**1.95E-03**	137.3
SMDDRL [116]		0.613	0.686	0.683	0.021	**1.02E-06**	**1.95E-03**	38.8
SGR-MCCDA		**0.765**	**0.785**	**0.789**	0.020	-	-	11.3
MCCA-SUMCOR [106]	Reuters	0.575	0.657	0.661	0.013	**3.32E-14**	**1.95E-03**	1392.3
MCCA-MAXVAR [129]		0.298	0.278	0.238	0.019	**5.20E-16**	**1.95E-03**	325.5
GMA [222]		0.837	0.821	0.821	0.007	**6.04E-08**	**1.95E-03**	307.7
MvDA [124]		0.560	0.578	0.579	0.012	**5.21E-13**	**1.95E-03**	23.0
MvDA-VC [124]		0.551	0.582	0.588	0.015	**1.00E-12**	**1.95E-03**	42.2
MULDA [239]		0.847	0.856	0.858	0.007	**3.96E-04**	**1.95E-03**	590.4
MvCCDA [292]		0.810	0.815	0.814	0.010	**5.59E-10**	**1.95E-03**	826.0
ReDMiCA [164]		0.662	0.696	0.697	0.008	**9.36E-12**	**1.95E-03**	11434.1
GMCCA [43]		0.685	0.406	0.418	0.042	**1.40E-08**	**1.95E-03**	509.8
DMvTSVM-AE [278]		0.808	0.848	0.848	0.015	**6.26E-03**	**1.37E-02**	656.5
SMDDRL [116]		**0.928**	0.864	0.864	0.003	**2.01E-03**	**3.91E-03**	203.7
SGR-MCCDA		0.866	**0.876**	**0.874**	0.008	-	-	297.8
MCCA-SUMCOR [106]	Caltech-20	0.418	0.707	0.705	0.025	**6.92E-13**	**1.95E-03**	19.8
MCCA-MAXVAR [129]		0.733	0.727	0.484	0.024	**1.55E-10**	**1.95E-03**	22.4
GMA [222]		0.795	0.821	0.817	0.012	**4.73E-10**	**1.95E-03**	46.9
MvDA [124]		0.763	0.788	0.789	0.015	**1.26E-11**	**1.95E-03**	10.8
MvDA-VC [124]		0.753	0.785	0.791	0.020	**4.92E-11**	**1.95E-03**	11.4
MULDA [239]		0.778	0.305	0.303	0.049	**2.59E-11**	**1.95E-03**	27.1
MvCCDA [292]		0.915	0.936	0.939	0.011	**4.72E-05**	**1.95E-03**	27.2
ReDMiCA [164]		0.852	0.801	0.791	0.026	**1.80E-08**	**1.95E-03**	8882.1
GMCCA [43]		0.850	0.839	0.837	0.011	**1.12E-10**	**1.95E-03**	4.9
DMvTSVM-AE [278]		0.866	0.720	0.724	0.011	**2.82E-14**	**1.95E-03**	79.5
SMDDRL [116]		0.761	0.753	0.747	0.017	**6.09E-10**	**1.95E-03**	26.2
SGR-MCCDA		**0.954**	**0.967**	**0.967**	0.007	-	-	7.9
MCCA-SUMCOR [106]	100Leaves	0.803	0.680	0.649	0.074	**4.74E-07**	**1.95E-03**	0.3
MCCA-MAXVAR [129]		0.738	0.853	0.850	0.029	**3.05E-07**	**1.95E-03**	0.4
GMA [222]		0.225	0.286	0.280	0.031	**2.07E-13**	**1.95E-03**	0.2
MvDA [124]		0.924	0.958	0.955	0.019	**1.46E-03**	**1.07E-02**	6.0
MvDA-VC [124]		0.951	0.979	0.983	0.015	*2.79E-01*	*2.57E-01*	6.1
MULDA [239]		0.911	0.855	0.855	0.035	**1.36E-06**	**1.95E-03**	0.2
MvCCDA [292]		0.941	0.974	0.975	0.008	**7.13E-03**	**1.06E-02**	6.1
ReDMiCA [164]		0.885	0.949	0.950	0.021	**1.05E-03**	**3.91E-03**	135.6
GMCCA [43]		0.738	0.675	0.670	0.022	**1.20E-10**	**1.95E-03**	0.4
DMvTSVM-AE [278]		0.238	0.099	0.100	0.002	**1.08E-19**	**1.95E-03**	68.7
SMDDRL [116]		0.163	0.104	0.133	0.011	**2.43E-18**	**1.95E-03**	6.4
SGR-MCCDA		**0.971**	**0.984**	**0.985**	0.009	-	-	0.8

Among the classical methods, MvCCDA performs really well in case of Caltech-20 and obtains the second highest classification accuracy after the proposed method SGR-MCCDA. The p-values for the benchmark data set, reported in Table 4.6 and 4.7, are all significant except for the methods, MvDA-VC with respect to 100Leaves data set and MvCCDA with respect to the data sets CiteSeer and NW-Object. It suggests that the 10-fold CV values of the methods MvDA-VC and MvCCDA are similar to the proposed method for those data sets. Both the deep learning models, DMvTSVM-AE and SMDDRL, perform poorly

Table 4.7: Classification Accuracy of Proposed Method (SGR-MCCDA) and Different Algorithms on Benchmark Data Sets

Different Algorithms	Data Sets	Accuracy (Train-Test)	Accuracy and Significance Analysis for 10-fold CV					Time (Sec.)
			Mean	Median	StdDev	Paired-*t*:p	Wilcoxon:p	
MCCA-SUMCOR [106]	CiteSeer	0.581	0.592	0.599	0.028	**1.72E-14**	**1.95E-03**	0.9
MCCA-MAXVAR [129]		0.567	0.560	0.453	0.010	**8.11E-05**	**1.95E-03**	4.4
GMA [222]		0.652	0.652	0.646	0.025	**2.01E-14**	**1.95E-03**	11.3
MvDA [124]		0.377	0.414	0.416	0.020	**2.13E-14**	**1.95E-03**	30.5
MvDA-VC [124]		0.435	0.477	0.473	0.032	**1.93E-05**	**1.07E-02**	27.2
MULDA [239]		0.666	0.671	0.676	0.026	**1.66E-19**	**1.95E-03**	7.5
MvCCDA [292]		0.685	0.682	0.694	0.026	**1.75E-15**	*6.44E-02*	17.4
ReDMiCA [164]		0.646	0.641	0.643	0.027	**1.43E-14**	**1.95E-03**	447.7
GMCCA [43]		0.234	0.462	0.458	0.040	**1.35E-14**	**1.95E-03**	5.9
DMvTSVM-AE [278]		0.489	0.480	0.480	0.016	**2.54E-09**	**1.95E-03**	132.6
SMDDRL [116]		0.515	0.618	0.625	0.028	**3.23E-05**	**1.95E-03**	64.3
SGR-MCCDA		**0.685**	**0.690**	**0.695**	0.022	-	-	14.9
MCCA-SUMCOR [106]	Caltech-101	0.306	0.345	0.332	0.013	**7.54E-15**	**1.95E-03**	5.2
MCCA-MAXVAR [129]		0.509	0.755	0.759	0.015	**8.01E-08**	**1.95E-03**	43.1
GMA [222]		0.379	0.285	0.286	0.013	**1.29E-15**	**1.95E-03**	4.4
MvDA [124]		0.607	**0.852**	**0.854**	0.091	**1.13E-16**	**1.95E-03**	1.1
MvDA-VC [124]		0.606	0.839	0.839	0.096	**2.16E-07**	**1.95E-03**	4.9
MULDA [239]		0.572	0.493	0.489	0.024	**2.45E-02**	**2.73E-02**	24.2
MvCCDA [292]		0.605	0.775	0.775	0.009	**1.76E-15**	**1.95E-03**	25.3
ReDMiCA [164]		0.391	0.412	0.400	0.014	**2.61E-02**	**2.73E-02**	4531.7
GMCCA [43]		0.509	0.678	0.678	0.016	**2.92E-08**	**1.95E-03**	62.6
DMvTSVM-AE [278]		0.553	0.737	0.736	0.007	**1.10E-09**	**1.95E-03**	164.0
SMDDRL [116]		0.378	0.704	0.715	0.017	**2.55E-08**	**1.95E-03**	128.3
SGR-MCCDA		**0.616**	0.820	0.820	0.009	-	-	35.7
MCCA-SUMCOR [106]	NW-Object	0.303	0.322	0.321	0.005	**5.48E-17**	**1.95E-03**	0.5
MCCA-MAXVAR [129]		0.051	0.054	0.080	0.008	**1.11E-11**	**1.95E-03**	344.4
GMA [222]		0.283	0.296	0.298	0.008	**5.25E-17**	**1.95E-03**	0.2
MvDA [124]		0.290	0.280	0.279	0.008	**5.55E-17**	**1.95E-03**	53.2
MvDA-VC [124]		0.286	0.279	0.281	0.008	**1.18E-10**	**1.95E-03**	73.8
MULDA [239]		0.370	0.314	0.314	0.007	**4.58E-02**	**1.95E-03**	0.9
MvCCDA [292]		0.424	0.431	0.432	0.007	**5.30E-17**	*6.44E-02*	1745.3
ReDMiCA [164]		0.377	0.382	0.382	0.009	**1.29E-17**	**1.95E-03**	1344.7
GMCCA [43]		0.046	0.055	0.055	0.005	**5.48E-17**	**1.95E-03**	1659.2
DMvTSVM-AE [278]		0.387	0.374	0.373	0.027	**8.02E-03**	**1.95E-03**	145.3
SMDDRL [116]		0.346	0.368	0.366	0.008	**9.35E-15**	**1.95E-03**	240.7
SGR-MCCDA		**0.409**	**0.439**	**0.438**	0.007	-	-	783.7
MCCA-SUMCOR [106]	ALOI	0.608	0.736	0.742	0.021	**1.72E-14**	**1.95E-03**	0.1
MCCA-MAXVAR [129]		0.874	0.916	0.917	0.009	**2.16E-15**	**1.95E-03**	55.0
GMA [222]		0.931	0.941	0.940	0.005	**2.31E-08**	**1.95E-03**	0.2
MvDA [124]		0.971	0.986	0.986	0.004	**2.01E-04**	**1.95E-03**	0.5
MvDA-VC [124]		**0.981**	0.992	0.992	0.003	**1.07E-15**	*6.44E-02*	4.9
MULDA [239]		0.906	0.926	0.914	0.007	**1.66E-14**	**1.95E-03**	0.3
MvCCDA [292]		0.977	0.986	0.985	0.002	**5.85E-09**	**1.95E-03**	141.6
ReDMiCA [164]		0.941	0.949	0.950	0.011	**1.66E-14**	**1.95E-03**	436.6
GMCCA [43]		0.913	0.892	0.893	0.009	**3.51E-09**	**1.95E-03**	58.8
DMvTSVM-AE [278]		0.808	0.871	0.881	0.008	**2.04E-14**	**1.95E-03**	133.6
SMDDRL [116]		0.820	0.838	0.839	0.006	**8.84E-13**	**1.95E-03**	60.6
SGR-MCCDA		0.971	**0.993**	**0.993**	0.003	-	-	22.8

on 100Leaves data set. This degradation in the performance is probably caused because of the less number of samples per class in that data, which affects the training of these model. Overall the performance of the proposed model with the reduced 25 dimension is the best among all the existing methods.

The feature versus accuracy curves for the data sets Cora and Caltech-20 are presented in Fig. 4.3 and the scatter plots for the data sets Cora, Reuters, and Caltech-20 are shown in Fig. 4.4 to compare the quality of the extracted features of the proposed model with

the existing ones. In Fig. 4.3, the blue curve of the proposed method stays always on the top of the other existing algorithms and also shows an increasing nature with respect to the number of dimensions. The performance of MULDA is impressive in case of Cora on both the setup. In case of Caltech-20, the yellow curve of MULDA is decreasing with the number of features over 10-fold CV. The scatter plots in Fig. 4.4 show that the proposed method has the best separating patterns of the classes for Cora and Caltech-20 data sets. In case of Reuters, the proposed SGR-MCCDA, MULDA, and MvCCDA have comparable performance in separating the classes. Other than the proposed method, MULDA has performed well in separating the classes of Caltech-20 with the top two extracted features.

Thus, the results in Table 4.5, 4.6, and 4.7, feature versus accuracy curves in Fig. 4.3 and scatter plots in Fig. 4.4 establish that the proposed method outperforms the existing state-of-art algorithms in terms of the classification accuracy in both the setup, separating the classes with the top-two extracted features and the increasing nature of classification accuracy with respect to the number of features. Moreover, the proposed model is scalable to the data sets with large number of samples and features. The proposed method can deal with the data sets having large number of classes as well.

4.5.3.3 Execution Time

Finally, the execution time reported in Table 4.5, Table 4.6 and Table 4.7 shows that the proposed method has comparatively lower execution time than most of the existing algorithms. In this regard, it should be noted that the deep learning models are executed on 16 GB NVIDIA RTX-4000 GPU, while all other models are evaluated with the following CPU configuration: RAM - 32 GB and Processor - Intel® Core™ i7-4770 CPU @ 3.40GHz × 8. The lower execution time of the proposed method is due to the quick convergence of the iterative algorithm. Fig. 4.5 shows the normalized deviation $\mathcal{D}$ of the objective function

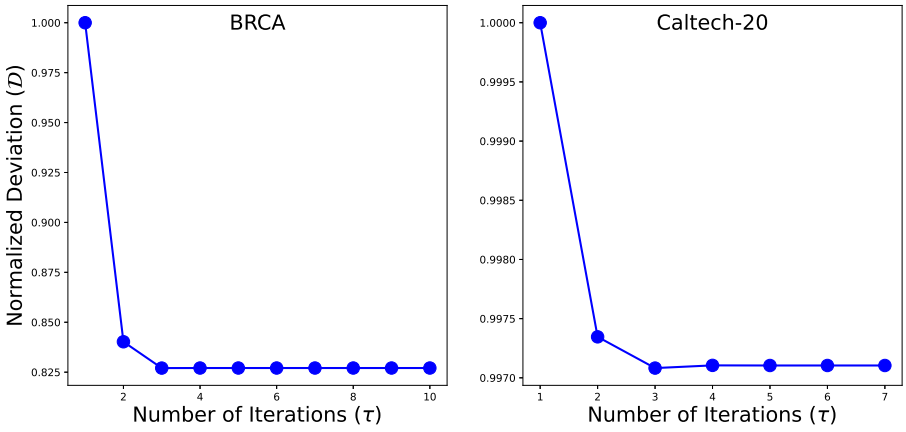


Figure 4.5: Variation of $\mathcal{D}$ with respect to the number of iterations τ.

of the SGR-MCCDA from its theoretical lower bound in each iteration, which is given by:

$$\mathcal{D} = \frac{\mathcal{J}(t) - \mathcal{K}}{\mathcal{J}(0) - \mathcal{K}}, \tag{4.51}$$

where $\mathcal{K}$ and $\mathcal{J}(t)$ are the theoretical lower bound and the value of the objective function of the SGR-MCCDA at the t-th iteration. The results are reported for two data sets, namely, BRCA and Caltech-20, as examples. The figure shows that the normalized deviation decreases rapidly with the increase of the number of iterations and stabilizes after few iterations for both the data sets, which establishes the quick convergence of the proposed SGR-MCCDA.

4.5.4 Comparison of CSP-MvCDA and SGR-MCCDA

To show the effectiveness of CSP-MvCDA and SGR-MCCDA, the performance of the two algorithms has been evaluated for all five cancer and eight benchmark data sets. The performance of the two algorithms, in terms of classification accuracy for both the train-test and 10-fold setup, is given in Table 4.8. The results show that, in case of train-test setup, the CSP-MvCDA attains significantly better classification accuracy for BRCA, CESC, GBM and NW-Object, while SGR-MCCDA obtains higher classification accuracy for the data sets CRC, Caltech-20 and 100Leaves. In case of 10-fold setup, the CSP-MvCDA performs better on CESC, LGG, Cora, Caltech-101 and NW-Object, while SGR-MCCDA performs significantly better on CRC, Reuters, Caltech-20 and 100Leaves. The possible reason can be observed from the Table 4.4. It shows that the weights of each view, learned by SGR-MCCDA, are not equally distributed, which promotes view relevance and justifies the higher accuracy value of SGR-MCCDA for these data sets. For LGG, Citeseer, Cora, Caltech-101 and ALOI the performance of the two algorithms are comparable in case of both the setups.

Table 4.8: Comparative Study of CSP-MvCDA and SGR-MCCDA on Different Data Sets

Data Sets	CSP-MvCDA				SGR-MCCDA			
	Accuracy (Train-Test)	Accuracy (10-fold)			Accuracy (Train-Test)	Accuracy (10-fold)		
		Mean	Median	StdDev		Mean	Median	StdDev
BRCA	**0.803**	**0.838**	0.813	0.085	0.727	**0.838**	**0.844**	0.085
CESC	**0.865**	**0.942**	**1.000**	0.092	0.788	0.817	0.833	0.090
CRC	0.862	0.909	0.912	0.061	**0.877**	**0.915**	**0.926**	0.041
LGG	**0.984**	**0.984**	**0.974**	0.012	0.978	**0.984**	**0.974**	0.013
GBM	**0.829**	0.820	0.815	0.060	0.781	**0.842**	**0.833**	0.039
CiteSeer	**0.698**	0.687	**0.701**	0.025	0.685	**0.690**	**0.695**	0.022
Cora	**0.775**	**0.785**	**0.801**	0.019	0.765	**0.785**	0.789	0.020
Reuters	0.859	0.775	0.774	0.008	**0.866**	**0.876**	**0.874**	0.008
Caltech-101	0.610	**0.823**	**0.826**	0.011	**0.616**	0.820	0.820	0.009
Caltech-20	0.905	0.882	0.882	0.018	**0.940**	**0.964**	**0.963**	0.005
NW-Object	**0.439**	**0.445**	**0.446**	0.008	0.409	0.439	0.438	0.007
ALOI	**0.981**	**0.993**	0.992	0.003	0.971	**0.993**	**0.993**	0.003
100Leaves	0.945	0.896	0.903	0.020	**0.971**	**0.984**	**0.985**	0.009

The time comparison of the two algorithms is also shown by the bar plot in Fig. 4.6. In case of the five cancer data sets, namely, BRCA, CESC, CRC, LGG, and GBM, the sample size is very small compared to the feature size. Since the SGR-MCCDA is an

iterative process and requires eigendecomposition of a sample-by-sample matrix, the time complexity is greatly reduced in SGR-MCCDA for the cancer data sets. On the other hand, the rest of the eight benchmark data sets, except CiteSeer and Cora, have a small number of features compared to the number of samples. Since the CSP-MvCDA requires eigendecomposition of a feature-by-feature matrix, the time to execute the CSP-MvCDA algorithm for these data sets is much shorter than the SGR-MCCDA. For the data sets, CiteSeer and Cora, the sample and feature size are almost equal. In this case, the SGR-MCCDA takes longer time to execute, because, it possibly took large number of iterations to compute the eigenvalues of the sample-by-sample matrix.

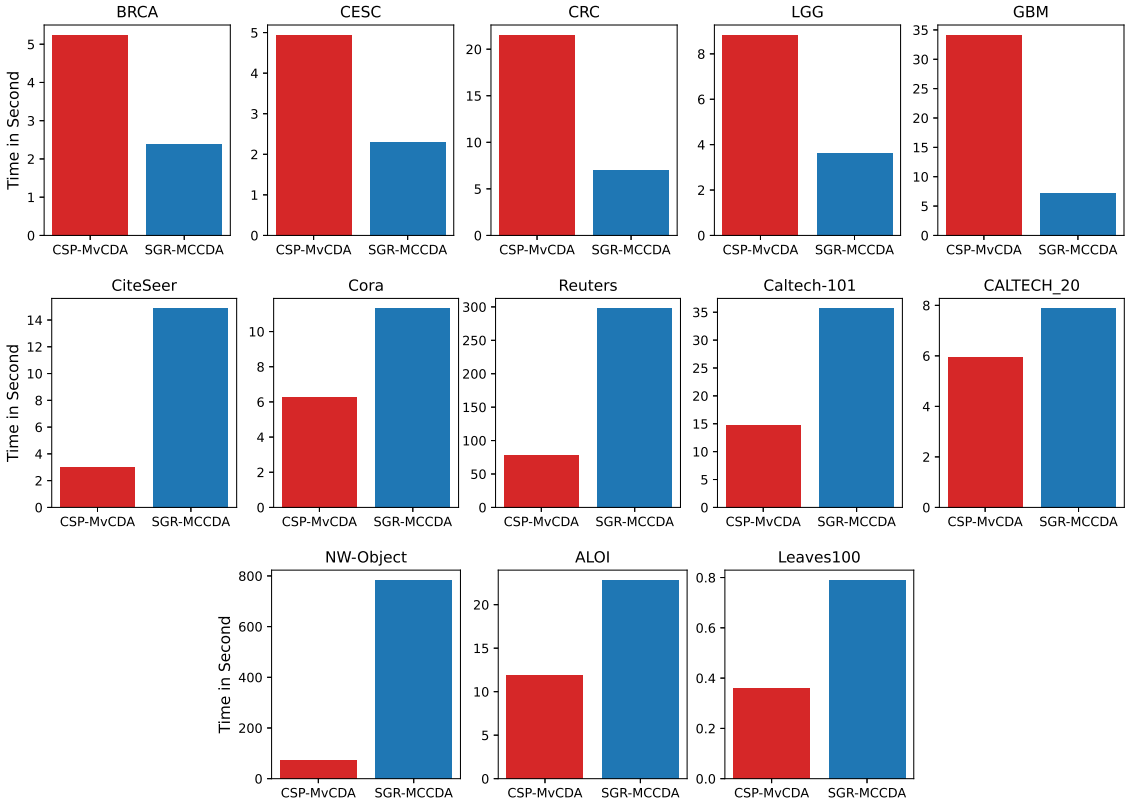


Figure 4.6: Time comparison of CSP-MvCDA and SGR-MCCDA on different data sets.

4.6 Conclusion

This chapter introduces a new supervised graph regularized multi-view learning algorithm, termed as SGR-MCCDA, based on the unsupervised subspace learning algorithm MCCA-MAXVAR. The method utilizes the prior knowledge of the data in terms of the class label and the structural geometry of the source vector. The class labels are used to encode the structural geometry of the source vectors by constructing the within-class and between-class graphs, which have been invoked as the regularizer term into the model. Due to the supervised graph information, the common subspace learned by the proposed method has the ability to preserve the class structure of the data. An iterative solution of the

proposed formulation has been presented. The convergence of the iterative rule and the computational complexity of the method have been studied. To deal with the data vectors whose dimension is very high, the low rank approximation technique is used to find inverse of the rank deficient class-covariance matrices. Finally, to prove the diverse potency of the proposed method, several cancer and benchmark data sets with different characteristics have been used in this study. The results justify the performance of the proposed method by giving significantly higher classification accuracy compared to the state-of-the-art algorithms.

In Chapter 3 and Chapter 4, two MVL algorithms, namely, CSP-MvCDA and SGR-MCCDA, respectively, are successfully applied to extract features from the multiple omics data for cancer stratification. In the next chapter, both the methods are applied in the domain of imaging genetics study for the identification of imaging and genetic biomarkers to classify neurodegenerative disease subtypes.

Chapter 5

Multi-View Data Analysis in Imaging Genetics Studies

5.1 Introduction

The neurodegenerative disorders, such as Alzheimer, Parkinson, and Huntington, are very common nowadays, and sometimes become serious or even life-threatening, as they cause brain deterioration over the time [235]. The risk of getting affected by these disorders is increasing day-by-day. These disorders occur when the nerve cells get damaged and lose their functions. A person having one of these disorders may suffer from forgetfulness, inability to move, or may face balance problem. An early diagnosis may help in improving some of the mental and physical symptoms associated with it. Therefore, it is necessary to develop new approaches for proper diagnosis of the disease and better understanding about the cause of the disease [99, 112].

The neurodegenerative disorders are complex diseases and considered to be caused by the interplay of a number of genetic factors such as change of gene regulation, alteration of mRNA and single nucleotide polymorphism (SNP). The genetic factors play an important role in causing these diseases. Both genetic variants and brain region abnormalities are recognized as important factors for such complex diseases. Therefore, brain function study and genetic variant study are two most popular approaches to diagnose these diseases. There exist many neuroimaging technologies such as magnetic resonance imaging (MRI), functional MRI (fMRI) and positron emission topography (PET), which give information about the brain function from different perspectives. On the other hand, the gene or genetic variants, SNP and copy number variation (CNV) are responsible for the disease to occur. Imaging genetics is an integrative study of the imaging (MRI, fMRI, PET) and genetic (SNP, CNV) modalities, to understand the impact of genetic variations over the brain functions and structure. It provides a comprehensive information to diagnose the neurodegenerative disease [74]. The importance of imaging genetics study is illustrated in Fig. 5.1. Recent large-scale initiative from the UK Biobank (https://www.ukbiobank.ac.uk/), CHARGE consortium (https://www.hgsc.bcm.edu/human/charge-consortium), and ENIGMA consortium (https://enigma.ini.usc.edu/) have gathered large amount of data containing genome-wide genetic information and brain imaging in order to investigate the relationship between spe-

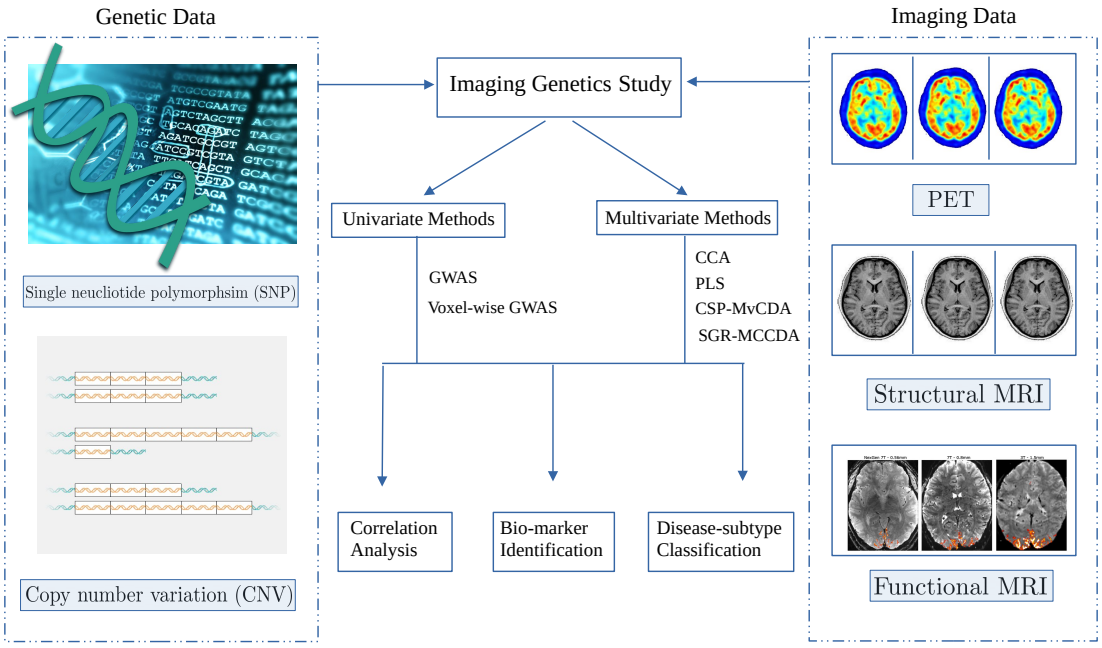


Figure 5.1: Illustration of imaging genetics studies.

cific nucleotide variants in the genome and differences in brain structure. Due to increase in the imaging genetics data availability and collection, more thorough studies have been possible for the practical importance of these discoveries in connection to human behavior and illness.

The main challenge in integrating the imaging and genetic modalities is that both of them are represented by a large number of features ($\sim 10^6$) for a small set of samples. Moreover, these two modalities are heterogeneous in nature. The mass univariate and voxel-wise methods are two popular univariate approaches to study the imaging genetics association. The genome-wide association study (GWAS) is a very powerful and widely used framework, which uses univariate marginal regression for identifying genetic variants that are associated with a given phenotype [261]. In Oxford Brain Imaging Genetics Server-BIG40, the results from GWAS studies across 40,000 subjects have been presented associating 17,103,079 SNPs to almost 4,000 imaging-derived phenotypes [68]. But, the univariate methods simply look for the association between each pair of voxel and genetic marker, and treat them independently. On the other hand, the voxel-wise approaches [99, 112] include multiple genetic markers in each model additionally. However, they overlook the joint relationships between the genetic marker and imaging quantitative traits (QTs), and thus fail to model the data explicitly. These methods also require massive univariate analysis ($\sim 10^{12}$), which leads to multiple testing problems.

In order to address the problem of univariate methods, several multivariate methods have been introduced to study the association between imaging and genetic data jointly, and to model the data explicitly. The most popular and successful methods of multivariate analysis for imaging genetics study are canonical correlation analysis (CCA) [107] and partial least square methods (PLS) [274]. The CCA finds a projective space where the

two data blocks are maximally correlated, whereas the PLS finds it by maximizing the covariance between the two data sets. However, due to the "large features-small samples" nature of the imaging genetics data, both CCA and PLS undergo severe overfitting issues. To overcome this problem, sparse CCA (SCCA) [31,194] and sparse PLS (SPLS) [50] have been introduced. Both of them have the ability to select a small set of features, employing a sparse penalty to the model, and are found to be very effective in identifying imaging genetic association [31,50]. In [205], a multivariate method, deploying sparse reduce rank regression (SRRR), has been formulated to improve the statistical performance for large scale imaging genetics data. An application to large scale UK Biobank data has been given to justify its performance on large scale data. In recent time, polygenic risk score (PRS) [199] models have been developed to understand the role of genetic variants, in risk of developing a disease. Recently developed algorithm, namely, batch screening iterative lasso [204], has been applied to individual-level genotype phenotype data from UK Biobank to find the significant PRS models. However, given multi-class imaging genetics data, the SCCA or SPLS models cannot utilize the label information of the data. Incorporation of the diagnostic status of the disease into the model would help in better feature selection and provide an in-depth analysis of the imaging genetics data.

Collaborative regression (CoRe) [86] and cooperative learning (CoopLe) [59] are two supervised multi-view learning algorithms, which have been introduced for analyzing multiple modalities. The CoRe is a form of supervised canonical correlation analysis under a regression framework, whereas CoopLe combines the concept of squared error loss of prediction with an agreement penalty to encourage the prediction of different data views to agree. In the context of imaging genetics, CoRe and CoopLe are two useful approaches for identifying characteristics shared by phenotypic and genotypic data, and they produce encouraging outcomes [16,315].

In this regard, two supervised MVL algorithms, namely, CSP-MvCDA and SGR-MvCDA, proposed in Chapter 3 and Chapter 4, respectively, are successfully used in this chapter to extract the features for imaging genetics studies for the identification of the genotype-phenotype association and disease-subtype classification. A real neuroimaging data set, obtained from Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort, has been used for imaging genetics analysis by using the AD-related genetic variants (SNPs) and brain imaging (fMRI) modalities. This chapter presents a comparative performance analysis of two proposed MVL methods, namely, CSP-MvCDA and SGR-MCCDA, with that of the existing MVL algorithms to show the effectiveness of the proposed methods in terms of classifying disease subtypes in imaging genetics studies. The results establish that the classification performance of proposed CSP-MvCDA and SGR-MCCDA on the ADNI data set is better than that of the existing MVL models. Moreover, these two approaches perform well with respect to capturing the correlation between imaging and genetic data, although they are not capable of identifying the disease-specific imaging genetics association.

The remaining sections of this chapter are as follows: Section 5.2 presents the data acquisition and preprocessing of the real neuroimaging genetic data. The performance of CSP-MvCDA and SGR-MCCDA in terms of feature extraction and biomarker selection with respect to the existing algorithms in the domain of imaging genetics study is given in Section 5.3. The chapter is concluded in Section 5.4.

5.2 Real Neuroimaging and Genetic Data

This section presents a brief description of the real neuroimaging and genetic data along with their preprocessing.

5.2.1 Data Source

The data used in the current study were obtained from the ADNI database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD[1]. The primary goal of ADNI is to detect early stage of Alzheimer's disease (AD) and to track the progression of the disease by developing clinical, imaging, genetic and bio-chemical biomarkers. Different types of neuroimaging data like PET scan, MRI and resting state fMRI, and different types of genetic modalities such as gene expression, DNA methylation profiling, and SNP data, are available in this cohort. A total of 120 participants are included in the current study whose resting state fMRI and SNP data are collected in their raw form from the ADNI cohort. The participants belong to four diagnostic groups, namely, control normal (CN), early mild cognitive impairment (EMCI), late mild cognitive impairment (LMCI) and AD. The basic characteristics of these participants are given in Table 5.1.

Table 5.1: Characteristic of the Participants

	CN	EMCI	LMCI	AD
Number	30	43	23	24
Gender (M/F)	15/15	18/25	15/8	11/13
Handedness (R/L)	28/2	43/0	23/0	23/1
Age (mean $\pm$ std)	76.27 $\pm$ 6.28	70.83 $\pm$ 6.72	70.68 $\pm$ 7.37	72.43 $\pm$ 7.67
Education (mean $\pm$ std)	16.30 $\pm$ 2.11	15.48 $\pm$ 2.66	16.39 $\pm$ 2.68	14.92 $\pm$ 2.66

5.2.2 fMRI Data Acquisition and Preprocessing

The fMRI images of the 120 participants were acquired by resting the participants with closed eyes and with no given task. The fMRI scans were captured by using Philips Medical Systems 3T scanner. The scanning protocol is as follows: Field Strength = 3.0 Tesla; Flip Angle = 80.0 degree; Acquisition Matrix = 64×64 pixels; Pixel Dimension = 3.3×3.3 mm^2; Pulse Sequence = GR; Slices = 6720.0; Slice Thickness = 3.3 mm; TE = 30.0 ms; and TR = 3000.0 ms. For quality control, the fMRI data is preprocessed using DPARSF toolbox [283] within MATLAB environment following the standard preprocessing steps. The steps involved in preprocessing of the fMRI data are given as follows:

- Conversion of the original file format to NIfTI file format. There were 6720 2-D slices of dimension 64×64 before conversion. After conversion, it gives 140, 3D volume of dimension $64 \times 64 \times 48$.

[1]The investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data, but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

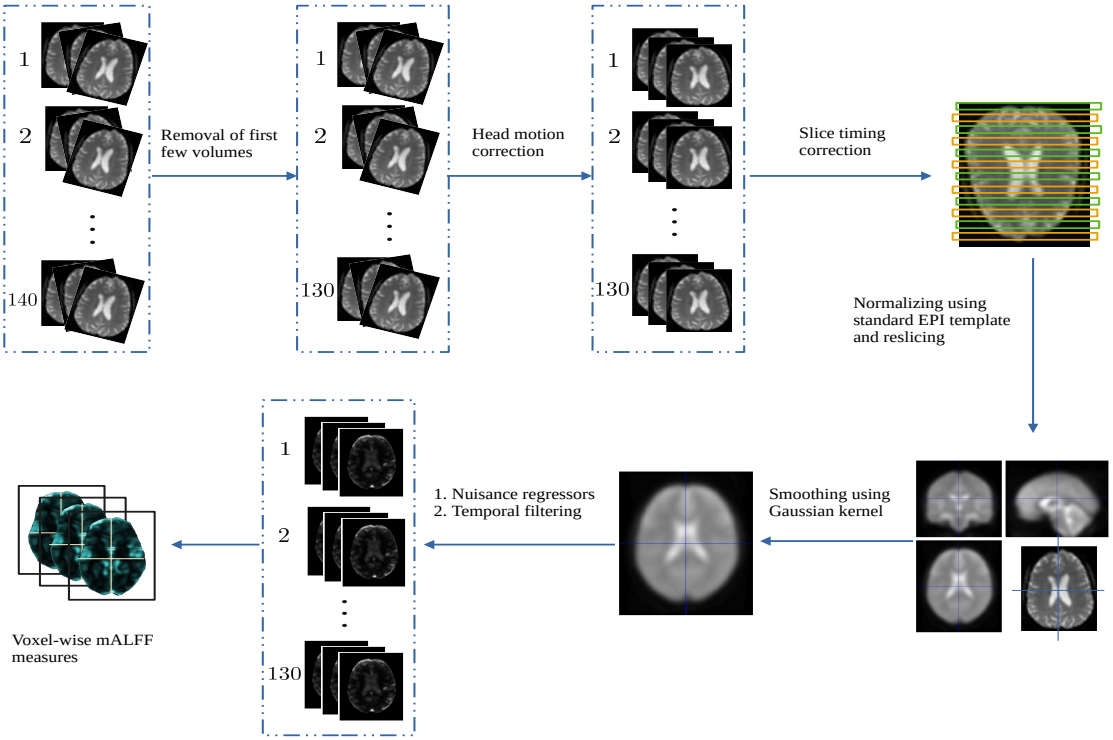


Figure 5.2: Preprocessing of resting state fMRI.

- First 10 volumes of the fMRI scans are discarded to make sure that the magnetic gradient field of the scanner is stable. After this step, 130 volumes are considered for further preprocessing.
- Slice timing correction is performed on the remaining slices to temporally align all slices with reference to a time-point.
- Head motion correction is required to remove the effect of head movement of each participant during scanning procedure.
- Normalization using EPI template is done and is resliced to $2 \times 2 \times 2$ mm^3.
- Images are smoothed to improve the signal-to-noise ratio by applying a Gaussian kernel with the full width at half maximum (FWHM) of 6 mm.
- Nuisance covariates including white matter and global signal are regressed out by using a linear model.
- Temporal filtering in the range of 0.01Hz to 0.08Hz is performed to retain the resting state low frequency signal.

After the standard preprocessing steps, there are 130, 3-D volumes of dimension $91 \times 109 \times 91$ left for further analysis. Considering these 130 3-D volumes, mean amplitude low frequency fluctuation (mALFF) measure of each voxel is calculated, which gives 9,02,629

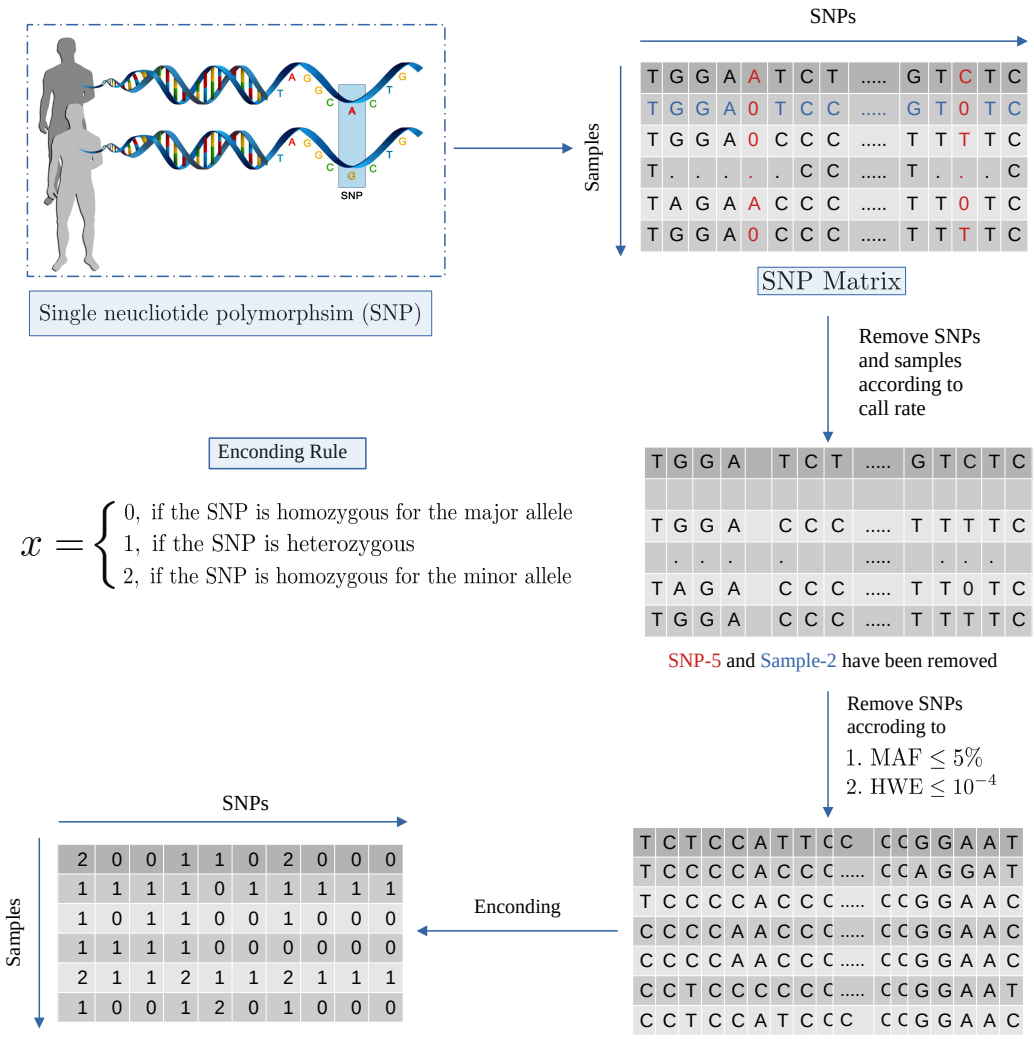


Figure 5.3: Preprocessing of SNP.

voxel-wise mALFF measures. The whole preprocessing of the resting state fMRI are represented pictorially in Fig. 5.2. These voxel-wise measures are further grouped according to the 52 ROIs, which are related to previously identified 7 resting state networks (RSN) including the default mode network (DMN), fronto-parietal attention network (FPAN), visual network, auditory network, sensori-motor network (SMN), basal ganglia network (BGN), and frontal network [228]. To reduce the time complexity, 52 ROIs are taken as a small spherical region centered at a seed voxel whose MNI coordinates are obtained from [228]. The spherical ROIs are created by taking 5 mm radius around the seed voxel using MarsBaR (MARSeille Boîte À Région d'Intérêt) toolbox in SPM, which provides routines for ROI analysis [33]. After grouping the voxels into 52 ROIs, it results in 3411 voxels for further analysis. Moreover, the imaging measurements were preadjusted to exclude the influence of age, gender, handedness and education.

5.2.3 SNP Data Acquisition and Preprocessing

The SNP data of same set of 120 samples is obtained through the Illumina Human 610-Quad or Illumina Human OmniExpress BeadChip. There are 3,34,343 SNPs at the beginning of the experiment. The SNP data is then preprocessed using PLINK tool (http://pngu.mgh.harvard.edu/purcell/plink/) [201] for quality purpose. The preprocessing steps reported in [23] are followed in this study. The raw SNP data goes through the following preprocessing steps.

- The SNPs from chromosome 1 to 22 are extracted.
- Sample call rate threshold is set to 95% to retain the quality of the genetic data.
- Minor allele frequency (MAF) is set to 5%.
- The SNP call rate threshold is set to 99.9%.
- Hardy-Weinberg equilibrium test (HWE) threshold is at 1×10^{-4}.

After the processing of SNP data, 2,64,031 SNPs remain to continue the experiment. A pictorial representation of the preprocessing steps is also given in Fig. 5.3. To reduce the number of SNPs further, all the SNPs are grouped into genes, and the genes, whose SNP counts are greater than a predefined threshold, are retained. After that, the genes were sorted according to the SNP counts and the top 50 genes are selected to reduce the computational cost. A total of 2500 SNPs, grouped into 50 genes, are considered to continue the experiment. The top 50 genes with their corresponding SNP count are reported in Table 5.2.

Following the preprocessing steps, finally, there are two blocks of data, fMRI and SNP, having dimensions of 120×3411 and 120×2500, respectively, to continue the experiment.

5.3 Comparative Performance Analysis

In this section, the performance of two supervised MVL algorithms, namely, CSP-MvCDA and SGR-MCCDA, proposed in Chapter 3 and Chapter 4, respectively, is compared with that of three primary MVL models in the domain of imaging genetics, namely, CCA, CoRe, and CoopLe. Since the objective of imaging genetics study is to find the genotype-phenotype association by identifying the disease related biomarkers, the first pair of principal components are obtained for the analysis of fMRI and SNP data [83, 137]. The canonical weights are sorted according to their absolute values, and the SNPs and voxels are selected by specifying a certain threshold. A five-fold cross-validation strategy is used to find the optimal parameters of the proposed models. The set of parameters, for which the average of canonical correlation coefficient (CCC) and classification accuracy (one versus all) is highest, is considered as the optimal parameter set.

5.3.1 Classification Accuracy

The classification performance of the proposed algorithms is compared with that of the three existing methods, and the corresponding results are reported in Table 5.3. The

Table 5.2: Sorted Genes according to SNP Count

Sl. No.	Gene	SNP Count	Sl. No.	Gene	SNP Count
1	CSMD1	768	26	NPAS3	176
2	RBFOX1	576	27	NELL1	172
3	PTPRD	460	28	SOX5	167
4	MACROD2	302	29	PCDH15	164
5	CDH13	294	30	DAB1	163
6	FHIT	293	31	NTM	162
7	MAGI2	252	32	ZNF385D	159
8	CNTNAP2	238	33	LINGO2	157
9	WWOX	226	34	DPP6	155
10	CNTN5	221	35	NAV2	155
11	OPCML	220	36	RYR3	154
12	TENM3	216	37	SORCS2	154
13	PRKN	215	38	ERBB4	153
14	SGCZ	212	39	ROBO2	152
15	CNTN4	212	40	CAMTA1	152
16	NRXN3	207	41	NRXN1	150
17	DLG2	200	42	ASTN2	146
18	CTNNA2	198	43	CTNNA3	146
19	TMEM132D	188	44	DLGAP2	146
20	PRKG1	182	45	CACNA2D3	145
21	ASIC2	182	46	PDZD2	144
22	ADAMTSL1	181	47	FRMD4A	143
23	PTPRT	179	48	ADARB2	142
24	LRP1B	179	49	PDE4D	142
25	KAZN	178	50	KCNIP4	140

classification accuracy is obtained by using the support vector machine (SVM). The absolute values of the canonical wights corresponding to SNPs and imaging QTs are sorted in ascending order, and the top ten SNPs and imaging QTs are considered, and then concatenated to generate 20 features. These 20 features are then used to calculate the one versus all classification accuracy for each diagnostic group. The classification results reported in Table 5.3 show that the proposed CSP-MvCDA and SGR-MCCDA have comparatively better classification accuracy than the other three algorithms. Particularly, the algorithm SGR-MCCDA proposed in Chapter 4 attains highest classification accuracy for all the four groups, namely, CN, EMCI, LMCI and AD. It signifies that the features selected by the proposed methods contain more discriminative information than the other existing methods with respect to the four classes.

5.3.2 Imaging Genetic Association

The association between genetic and imaging data has been assessed by reporting the canonical correlation coefficient (CCC) between the SNP and fMRI data. The performance of the proposed methods with respect to mean testing CCC is compared with that of the three existing methods, and the corresponding results are reported in Table 5.3. Higher value of CCC indicates the stronger imaging genetic association. It is seen from the results reported in Table 5.3 that both the CCA and CoRe attain similar correlation values of 0.1189 and 0.1110, respectively, but the standard deviation of CCC in case of

Table 5.3: Canonical Correlation Coefficient and Classification Accuracy (mean ± std)

Metric	Algorithms	CN	EMCI	LMCI	AD
Canonical Correlation Coefficient	CCA [107]	0.1189 ± 0.1070	0.1189 ± 0.1070	0.1189 ± 0.1070	0.1189 ± 0.1070
	CoRe [86]	0.1110 ± 0.2063	0.1110 ± 0.2063	0.1110 ± 0.2063	0.1110 ± 0.2063
	CoopLe [59]	0.1307 ± 0.2524	0.1307 ± 0.2524	0.1307 ± 0.2524	0.1307 ± 0.2524
	CSP-MvCDA	0.1353 ± 0.1477	0.1353 ± 0.1477	0.1353 ± 0.1477	0.1353 ± 0.1477
	SGR-MCCDA	**0.1869 ± 0.1533**	**0.1869 ± 0.1533**	**0.1869 ± 0.1533**	**0.1869 ± 0.1533**
Classification Accuracy	CCA [107]	0.5417 ± 0.1054	0.5417 ± 0.0833	0.5667 ± 0.0971	0.5916 ± 0.1196
	CoRe [86]	0.5000 ± 0.1490	0.5583 ± 0.0424	0.4416 ± 0.0772	0.5583 ± 0.0424
	CoopLe [59]	0.5000 ± 0.0874	0.5250 ± 0.0857	0.5500 ± 0.0964	0.4416 ± 0.0857
	CSP-MvCDA	0.5550 ± 0.1224	0.5550 ± 0.1280	0.5916 ± 0.0967	0.5500 ± 0.0928
	SGR-MCCDA	**0.5833 ± 0.0950**	**0.5916 ± 0.0927**	**0.5916 ± 0.0889**	**0.6000 ± 0.0857**

CoRe over five folds are high compared to CCA. On the other hand, although the CoopLe and proposed CSP-MvCDA methods have similar correlation values, the proposed CSP-MvCDA has lower standard deviation compared to CoopLe. The proposed SGR-MCCDA attains highest CCC value of 0.1869 with the standard deviation of 0.1533. All these results signify that although the proposed CSP-MvCDA and SGR-MCCDA algorithms capture better correlation values than the other existing methods, the standard deviation is quite high for all the methods. The possible reason is that the imaging genetics data are noisy and the MVL models used in this study are not designed to handle such noise, which makes these model non-robust with respect to noise of the fMRI and SNP data.

Table 5.4: Top Ten ROIs and Their Weights, Selected by Different Methods.

CCA		CoRe		CoopLe		CSP-MvCDA		SGR-MCCDA	
ROI	weight	ROI	weight	ROI	weight	ROI	weight	ROI	weight
AMPFC	0.025	MCC	0.419	RMCC	0.147	RSFG	0.026	RSFG	0.025
AMPFC	0.020	RSMG	0.173	BCG	0.134	RSFG	0.025	RSFG	0.020
RC	0.020	MCC	0.163	RIPL	0.130	RSFG	0.024	RITG	0.016
RIFG	0.019	RIPL	0.139	RSMG	0.083	LDPFC	0.021	RSFG	0.015
AMPFC	0.018	RIPL	0.121	LLPC	0.076	RSFG	0.021	AMPFC	0.015
VMPFC	0.018	RMCC	0.109	RMFG	0.069	LSFG	0.020	RPHG	0.015
VMPFC	0.018	MCC	0.094	RSMG	0.067	LDPFC	0.020	AMPFC	0.014
LITG	0.017	MCC	0.088	BCG	0.066	LSFG	0.020	RSFG	0.013
AMPFC	0.016	MCC	0.086	LS1	0.065	VMPFC	0.018	LSFG	0.012
RLPC	0.016	MCC	0.079	RSMG	0.063	RITG	0.018	RPHG	0.012

5.3.3 Interpretation of Selected Brain Regions

The selected voxels from the predefined ROIs, belonging to the seven resting state networks, are shown in the right part of the heatmap in Fig. 5.4 and the ROIs selected according to absolute values of the voxel weights are given in Table 5.4. The top ROIs selected by the proposed SGR-MCCDA method are left superior frontal gyrus(LSFG), right superior frontal gyrus (RISG), right inferior temporal gyrus (RITG), anterior medial prefrontal cortex (AMPFC), and right parahippocampal gyrus (RPHG), which are part of the default mode network (DMN) [228]. Among the so-called resting-state networks, the DMN has been linked to self-directed cognitive functions as autobiographical mem-

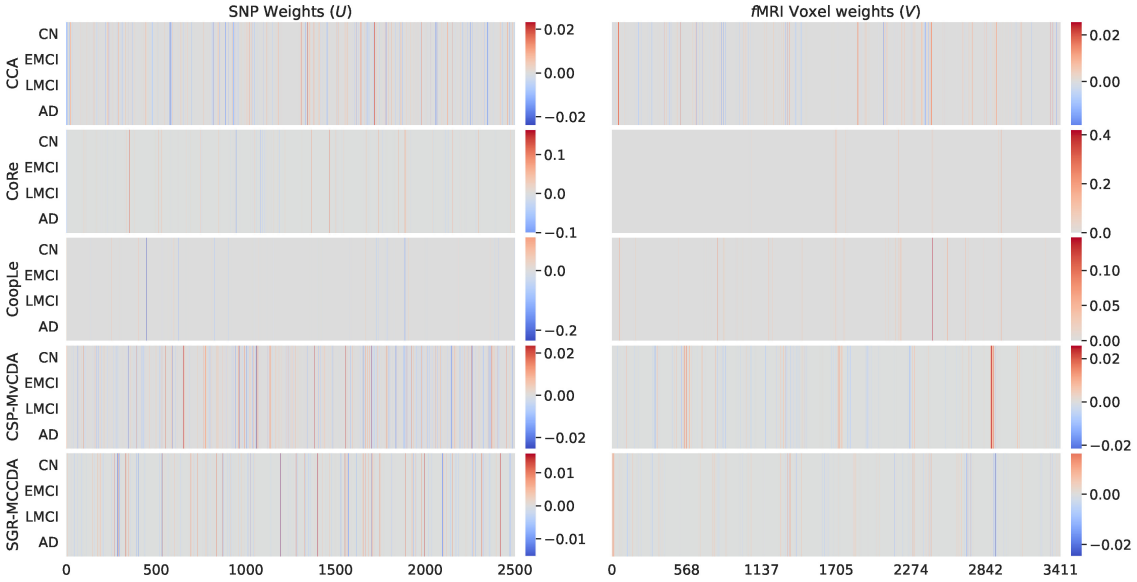


Figure 5.4: Mean canonical weights of SNPs (U: left column) and voxels (V: right column) obtained over five-fold cross-validation. Rows 1 to 5 correspond to the canonical weights for the five methods, namely, CCA, CoRe, CoopLe, CSP-MvCDA and SGR-MCCDA.

ory and introspection. Mental health appears to depend on the integrity of the DMN. Individuals with Alzheimer's disease disrupts the functional connectivity within the DMN brain areas [69, 84]. Thus, the selection of the ROIs, by SGR-MCCDA, is significant with respect to AD. The ROIs selected by the proposed CSP-MvCDA method are right superior frontal gyrus (RSFG), LSFG, RITG, left dorsolateral prefrontal cortex (LDPFC), and ventral medial prefrontal cortex (VMPFC). All these ROIs except LDPFC are part of the DMN, and hence are significant towards the progression of AD. The LDPFC belongs to the fronto-parietal network (FPN), which is involved in executive function and goal-oriented cognitively demanding tasks [255]. Research indicates that in AD, the fronto-parietal network (FPN) and dorsal attention network (DAN) are affected after the primary affected network, DMN [203]. Thus, the selection of these ROIs in case of CSP-MvCDA is also significant with respect to progression of AD. The CCA, apart from selecting ROIs belonging to DMN, selects right claustrum (RC) [258] and right inferior frontal gyrus (RIFG) [127] belonging to frontal network (FN). All the methods select ROIs more or less from the DMN, but most of ROIs selected by the proposed CSP-MvCDA and SGR-MCCDA algorithms are from the primary affected network, DMN, of AD, which supports their higher correlation value in Table 5.3.

5.3.4 Interpretation of Selected SNPs and Corresponding Genes

The selected SNPs are shown by the heatmaps in the left column of Fig. 5.4. To get a clear view of the SNPs shown in the heatmap, the top 10 SNPs are selected according to their absolute value of the canonical weights, and the corresponding genes are shown in Table 5.5. The SNPs selected by the proposed SGR-MCCDA belong to genes NPAS3,

Table 5.5: Top Ten SNPs, along with corresponding Genes and Weights, Selected by Different Methods

CCA		CoRe		CoopLe	
SNP(Gene)	weight	SNP(Gene)	weight	SNP(Gene)	weight
rs4454114 (PRKN)	0.024	rs11218323 (CNTN5)	0.163	rs12421122 (CNTN5)	0.235
rs11946827 (SORCS2)	0.023	rs6464749 (CNTNAP2)	0.125	rs2135732 (CTNNA3)	0.180
rs11749056 (PDZD2)	0.023	rs6016752 (PTPRT)	0.111	rs7931452 (NAV2)	0.146
rs7937542 (NELL1)	0.022	rs7951332 (NTM)	0.102	rs6571585 (NPAS3)	0.116
rs11646321 (CDH13)	0.022	rs2891126 (ADAMTSL1)	0.100	rs7948049 (CNTN5)	0.110
rs3846424 (SORCS2)	0.021	rs7073786 (PCDH15)	0.079	rs10002408 (TENM3)	0.100
rs577876 (PRKN)	0.021	rs790350 (DLG2)	0.078	rs1439281 (TENM3)	0.087
rs4641570 (SOX5)	0.020	rs7931452 (NAV2)	0.076	rs2037197 (DLG2)	0.083
rs11990503 (CSMD1)	0.020	rs7121400 (NELL1)	0.076	rs7657904 (SORCS2)	0.074
rs3828472 (GRM7)	0.019	rs10483889 (NRXN3)	0.076	rs13376837 (CTNNA3)	0.067

CSP-MvCDA		SGR-MCCDA	
SNP(Gene)	weight	SNP(Gene)	weight
rs6016732 (PTPRT)	0.025	rs10151755 (NPAS3)	0.016
rs6072654 (PTPRT)	0.023	rs1911355 (CTNNA3)	0.015
rs12892137 (NRXN3)	0.023	rs1435110 (RYR3)	0.014
rs27723 (PDE4D)	0.023	rs6565062 (CDH13)	0.014
rs10833428 (NELL1)	0.022	rs9922483 (WWOX)	0.013
rs800621 (ZNF385D)	0.021	rs6450853 (PDZD2)	0.013
rs12152570 (SORCS2)	0.019	rs1104979 (NRXN1)	0.012
rs653127 (RBFOX1)	0.019	rs1497366 (ASIC2)	0.012
rs3807669 (MAGI2)	0.018	rs12495770 (ZNF385D)	0.012
rs4837530 (ASTN2)	0.018	rs1465221 (MAGI2)	0.012

CTNNA3, RYR3, CDH13, WWOX, PDZD2, NRXN1, ASIC2, ZNF385D, and MAGI2. Out of these ten genes, eight genes, namely, CTNNA3 [28, 171], RYR3 [82], CDH13 [156], WWOX [108], PDZD2 [9], NRXN1 [167], ASIC2 [227], MAGI2 [302] are all proved to be associated with the progression of AD. There is no significant evidence of the other two genes, NPAS3 and ZNF385D, to be associated with AD. In case of the proposed CSP-MvCDA method, the six genes, namely, NRXN3 [100], PDE4D [276], NELL1 [117], RBFOX1 [206], MAGI2 [302], and ASTN2 [271], are found to be associated with AD, while the three genes have no significant effect on the progression of AD. The SGR-MCCDA method has selected eight AD-related genes, whereas CSP-MvCDA has identified six genes, which justifies the higher correlation value of SGR-MCCDA. Among all the genes, the genes, namely, NAV2, CTNNA3, NRXN1, NRXN3, RBFOX1, RYR3, WWOX, and CSMD1, are found to be common and significant with respect to AD. Most of these genes are selected by the proposed SGR-MCCDA method, which supports the results of getting the highest correlation value among all the methods for imaging genetic association in Table 5.3. The AD-related genes selected by CCA are CSMD1, NELL1, CDH13, and PDZD2. Both methods, CoRe and CoopLe, have selected NAV2, which is a risk gene in AD and is highly expressed in

the brain and involved in nervous system development, according to Wang et al. [264]. All these results justify the role of genetic biomarkers in imaging genetic associations for AD.

The classification results, reported in Table 5.3, indicate that although it is very difficult to separate different diagnostic groups for the real imaging genetic data using resting state fMRI and SNP data, the proposed SGR-MCCDA method has performed best among all the methods. It can also be seen that the correlation obtained by the proposed method, by the selection of meaningful genes and ROIs, is significant. Overall, the proposed SGR-MCCDA method shows a promising result in terms of classification and correlation, and the performance of the proposed CSP-MvCDA method is comparatively better than that of the existing models in multi-class imaging genetics study.

5.4 Conclusion

In imaging genetics research, the main objective is to investigate the complex genotype-phenotype association for the disease under study. To understand the impact of genetic variations over the brain functions and structure, the genotypic data such as SNP is integrated with the phenotypic data such as imaging quantitative traits. In this chapter, two new MVL algorithms, namely, CSP-MvCDA and SGR-MCCDA, proposed in Chapter 3 and Chapter 4, respectively, are successfully used to extract features for imaging genetics correlation and select important biomarkers for the disease subtype classification, which highlights the key characteristics of the imaging genetics study. A comparative performance analysis with the three primary MVL algorithms of imaging genetics domain, namely, CCA, CoRe and CoopLe, is presented in this chapter. A real imaging genetics data set, obtained from ADNI cohort, is used for the analysis by using the AD related genetic variants (SNPs) and brain imaging (fMRI) modalities. The results establish that the classification performance of the proposed CSP-MvCDA and SGR-MCCDA on the ADNI data set is better than that of the existing MVL models. Moreover, these two approaches perform well with respect to capturing the correlation between imaging and genetic data, although they are not capable of identifying the disease-specific imaging genetics association, which necessitates sparse multi-task framework in this field.

Since the number of features in genotypic and/or phenotypic data is significantly higher as compared to the number of samples, the sparse models, based on CCA, are popular in this area to find the complex bi-multivariate genotype-phenotype association. However, the sparse CCA based methods are, in general, unsupervised in nature, and fail to identify the diagnose-specific features those play an important role for the diagnosis and prognosis of the disease under study. In this regard, a new supervised model is proposed in next chapter to study the complex genotype-phenotype association, by judiciously integrating the merits of CCA, linear discriminant analysis (LDA) and multi-task learning.

Chapter 6

Multi-Task Learning and Sparse Discriminant Canonical Correlation Analysis for Imaging Genetics Study

6.1 Introduction

In imaging genetics studies, identification of diagnose-specific imaging genetic association may provide an in-depth investigation of a targeted group and can help in personalized treatments. In this regard, the algorithms, which can select a small set of features for each of the diagnostic groups, are of great interest. In [74], Fang et al. proposed joint SCCA (JSCCA), to study the imaging genetic association by applying the SCCA model in each diagnostic group. However, it has the limitation of getting many undesirable imaging genetics associations, which dominate the associations of interest, because of applying the SCCA within a single diagnostic group. Moreover, it is not suitable for diagnose-specific feature selection, although it is supervised in nature. In [66], another SCCA model, termed as multi-task SCCA (MT-SCCA), has been proposed to study the genetic associations with multiple imaging phenotypes by applying multiple SCCA model jointly. Similar to most of the SCCA models, the MT-SCCA is also unsupervised and cannot capture the diagnose-specific genotype-phenotype association. Also, the MT-SCCA model uses the group structure of the SNP, but is unable to use the group structure of brain imaging modality to learn the canonical weights. A supervised technique, termed as MT-SCCALR, has been proposed in [63], which is the combination of MT-SCCA and multi-task logistic regression. It has the capability of identifying the characteristic of a diagnostic group by selecting the class-specific and class-consistent features. Although it is supervised in nature and has the ability to select diagnose-specific features, it is computationally very expensive and is not suitable for large scale imaging genetic associations. In recent time, two more supervised methods, namely, collaborative regression (CoRe) [86] and cooperative learning (CoopLe) [59], have been introduced for analyzing multiple modalities. Though these methods are supervised, they are not suitable for diagnose specific feature selection.

In this regard, a novel multi-task bi-multivariate learning method, termed as multi-task

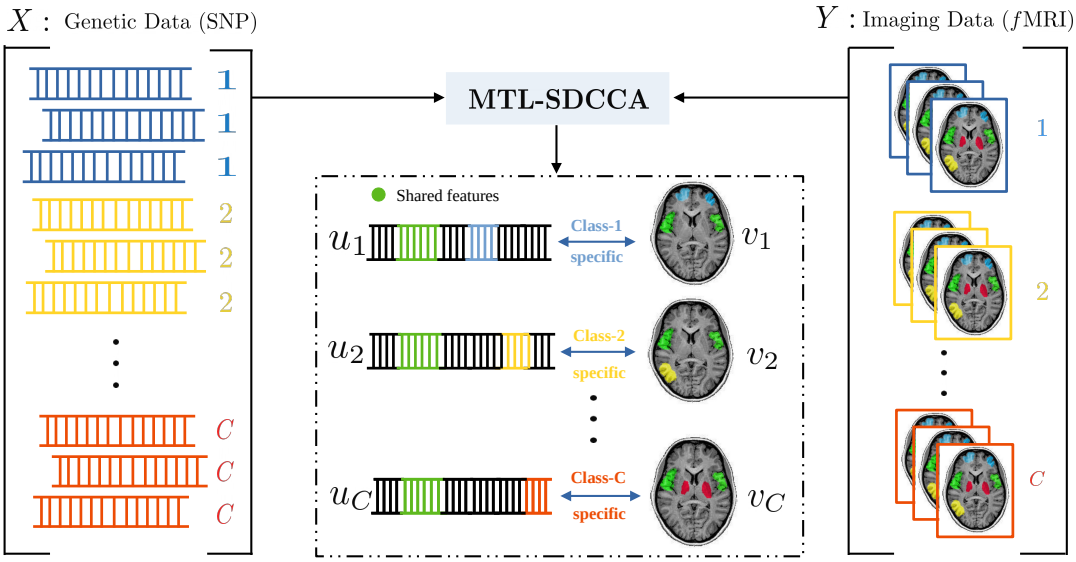


Figure 6.1: An illustration of the MT-SDCCA method for the identification of diagnose-specific imaging genetic association.

learning and sparse discriminant canonical correlation analysis (MTL-SDCCA) is proposed in this chapter to identify the diagnostic group specific imaging genetic patterns as illustrated in Fig. 6.1. It judiciously integrates the merits of CCA, linear discriminant analysis (LDA) and multi-task learning. Unlike most of the existing SCCA models, it learns two canonical weight matrices, one for the genetic data and another for the imaging data. Each column of the weight matrices is associated with a class-specific task. The proposed method not only considers the group structure of the genetic variants, but also considers the group structure of the brain imaging modality that is generally overlooked by most of the existing methods. In the proposed model, the group structure is modeled with the group $l_{2,1}$, that is, $G_{2,1}$ regularization. Also, the joint selection of features through all the classes is modeled using $l_{2,1}$ regularization, and the $l_{1,1}$ norm is used to sparsify the features. The proposed method is applied on several simulated data sets and one real imaging genetic data obtained from ADNI cohort. The current study integrates the SNP and resting state fMRI data to find the bi-multivariate associations with respect to Alzheimer disease. All the results show that the proposed method outperforms the existing algorithms. It can capture the class-specific and class-consistent features more accurately than any other methods, and yields better canonical correlation over all the existing algorithms. The class-specific and class-consistent genetic variants associated with the resting state network can provide better information for calculating polygenic scores (PGs) for Alzheimer disease. The proposed method also provides higher classification accuracy as compared to other existing methods. Moreover, the proposed method is computationally less expensive than the existing diagnosis-specific feature selection methods, such as MT-SCCALR. Some of the results of this chapter can be found in [177].

The remaining sections of this chapter are as follows: Section 6.2 presents the proposed methodology. Computational complexity analysis of the proposed MTL-SDCCA

with respect to different algorithms is also given in this section. The efficacy of the proposed method, in terms of selecting biomarkers for imaging genetics association and disease subtype classification, is shown by comparing the performance of it with the other state-of-the-art algorithms over two simulated and a real neuroimaging data in Section 6.3. The chapter is concluded in Section 6.4.

6.2 Proposed Method

In this section, a multi-task sparse method is introduced to identify the association between the imaging and genetic data in such a way that the disease specific genetic markers and brain voxels can be found out accurately.

6.2.1 Formulation of the Proposed Method

Let $X \in \mathbb{R}^{n\times p}$ be the genetic data and $Y \in \mathbb{R}^{n\times q}$ represents the brain imaging phenotype data, where n is the number of participants, p and q are the number of genetic markers and imaging QT measures, respectively. Let us also assume that there are C number of diagnostic groups. Then, two discriminative spaces for two modalities can be constructed by the objective function of the LDA as follows:

$$\max_{w_x} \frac{w_x^T S_{T,x} w_x}{w_x^T S_{W,x} w_x}; \qquad \max_{w_y} \frac{w_y^T S_{T,y} w_y}{w_y^T S_{W,y} w_y}; \tag{6.1}$$

where w_x and w_y are the projection vectors corresponding to the data sets X and Y, respectively. On the other hand, $S_{T,x}$ and $S_{W,x}$ represent the total scatter matrix and within-class scatter matrix of the data X, respectively.

Similarly, a common space, where two data sets are maximally correlated, can be found by the CCA objective function as follows:

$$\max_{w_x,w_y} \frac{w_x^T \Sigma_{xy} w_y}{\sqrt{w_x^T \Sigma_{xx} w_x}\sqrt{w_y^T \Sigma_{yy} w_y}}; \tag{6.2}$$

where w_x and w_y are the direction vectors along which two data sets are projected.

Combining (6.1) and (6.2), the proposed method finds a new projective space where the correlation between the two data sets as well as the class separability of individual data set are simultaneously maximized. Hence, the objective function of the proposed method can be formulated as follows:

$$\max_{w_x,w_y} \frac{w_x^T \Sigma_{xy} w_y}{\sqrt{w_x^T \Sigma_{xx} w_x}\sqrt{w_y^T \Sigma_{yy} w_y}} \sqrt{\frac{w_x^T S_{T,x} w_x}{w_x^T S_{W,x} w_x}} \sqrt{\frac{w_y^T S_{T,y} w_y}{w_y^T S_{W,y} w_y}}.$$

Since the total scatter of the data sets X and Y equal to

$$S_{T,x} = \Sigma_{xx}; \quad \text{and} \quad S_{T,y} = \Sigma_{yy}; \tag{6.3}$$

the objective function of the proposed method reduces to

$$\max_{w_x,w_y} \frac{w_x^T\Sigma_{xy}w_y}{\sqrt{w_x^T\Sigma_{xx}w_x}\sqrt{w_y^T\Sigma_{yy}w_y}}\sqrt{\frac{w_x^T\Sigma_{xx}w_x}{w_x^TS_{W,x}w_x}}\sqrt{\frac{w_y^T\Sigma_{yy}w_y}{w_y^TS_{W,y}w_y}}$$

$$= \max_{w_x,w_y} \frac{w_x^T\Sigma_{xy}w_y}{\sqrt{w_x^TS_{W,x}w_x}\sqrt{w_y^TS_{W,y}w_y}}. \tag{6.4}$$

As the rescaling of w_x and w_y cannot affect the solutions of the problem (6.4), the above maximization problem can also be expressed as a constraint optimization problem as follows:

$$\begin{aligned} &\max_{w_x,w_y} \quad w_x^T\Sigma_{xy}w_y \\ &\text{subject to} \quad w_x^TS_{W,x}w_x = 1, \quad w_y^TS_{W,y}w_y = 1. \end{aligned} \tag{6.5}$$

Based on the above objective function, a novel multi-task model is proposed to find the disease specific genetic markers and imaging QTs while maximizing the correlation between them. The formulation of the objective function as a minimization problem is given by

$$\begin{aligned} &\min_{u_c,v_c} \quad -\sum_{c=1}^{C} u_c^T\Sigma_{xy}v_c \\ &\text{subject to} \quad u_c^TS_{W,x}^cu_c = 1, \quad v_c^TS_{W,y}^cv_c = 1, \\ &\qquad P_1(U) < t_1, \quad P_2(V) < t_2; \end{aligned} \tag{6.6}$$

where u_c and v_c are analogous to w_x and w_y of (6.5) for each class c, U is the canonical weight matrix associated with X, consisting of the class-specific weights u_c of the genetic markers and V is the canonical weight matrix associated with Y, consisting of the class-specific weights v_c, of the imaging QTs. Here, $P_1(U)$ and $P_2(V)$ are two penalty functions. The within-class variance matrices for a specific class c, namely, $S_{W,x}^c$ and $S_{W,y}^c$ are defined as follows:

$$\begin{aligned} S_{W,x}^c &= \sum_{i\in c}(x_i-\mu_c)(x_i-\mu_c)^T + \sum_{i\notin c}(x_i-\bar{\mu}_c)(x_i-\bar{\mu}_c)^T; \\ S_{W,y}^c &= \sum_{i\in c}(y_i-\nu_c)(y_i-\nu_c)^T + \sum_{i\notin c}(y_i-\bar{\nu}_c)(y_i-\bar{\nu}_c)^T; \end{aligned}$$

where μ_c and ν_c are the means of the samples belonging to the c-th class of the data X and Y, respectively, and $\bar{\mu}_c$ and $\bar{\nu}_c$ are the means of the samples those do not belong to the c-th class of X and Y, respectively. Two scatter matrices $S_{W,x}^c$ and $S_{W,y}^c$ help in class binarization to set up a multi-class classification task by using the one-versus-all decomposition technique. If the classes in the data set are not balanced, the class balancing technique such as random oversampling can be used to overcome this issue, while computing

the scatter matrices $S^c_{W,x}$ and $S^c_{W,y}$ [198].

6.2.2 Regularization for Class-Specific and Class-Consistent Feature Selection

Since the genetic markers SNPs can be grouped according to their corresponding genes or linkage disequilibrium or tag SNPs and the voxel-wise measures in different neuroimaging modalities can be grouped according to different region of interests (ROIs) in the brain, a realistic modeling should include these information into the model. The group lasso penalty can be used to incorporate the group structure of SNPs or voxel-wise measures. But, since it can be applied to a vector, it cannot be used directly to penalize the canonical weight matrices U and V of the proposed method. To penalize the weight matrices U and V, $G_{2,1}$ norm, proposed in [263], is used, which is defined for U and V, respectively, as follows:

$$\|U\|_{G_{2,1}} = \sum_{k=1}^{K} \|U^k\|_F = \sum_{k=1}^{K} \sqrt{\sum_{i \in p_k} \sum_{j=1}^{C} u_{ij}^2};$$
$$\|V\|_{G_{2,1}} = \sum_{l=1}^{L} \|V^l\|_F = \sum_{l=1}^{L} \sqrt{\sum_{i \in q_l} \sum_{j=1}^{C} v_{ij}^2}.$$

Here, SNPs are assumed to be partitioned into K subgroups, $\{p_k\}_{k=1}^K$, and voxel-wise imaging measures are partitioned into L ROIs, $\{q_l\}_{l=1}^L$. Applying this regularization would consider the SNPs or voxels belonging to the same group as a whole, and would give a similar weight to all the SNPs and voxels. Also, this regularization would penalize the canonical coefficients of a group of variables across multiple task jointly, which practically helps to select task or class-consistent features.

Although the $G_{2,1}$ norm helps to select the SNPs and voxels according to their group structure, it is unable to select features at an individual level shared through all the task. There might be the case that some SNPs within a specific group shared by all the task is relevant to the imaging QTs, but the remaining ones are not. In that case, an individual level feature selection is required. In the proposed model, this is obtained via $l_{2,1}$ norm, the modified lasso regularization for multi-task feature selection. It is defined for U and V, respectively, as follows:

$$\|U\|_{l_{2,1}} = \sum_{i=1}^{p} \|u_i\|_2 = \sum_{i=1}^{p} \sqrt{\sum_{j=1}^{C} u_{ij}^2};$$
$$\|V\|_{l_{2,1}} = \sum_{i=1}^{q} \|v_i\|_2 = \sum_{i=1}^{q} \sqrt{\sum_{j=1}^{C} v_{ij}^2}.$$

So, $G_{2,1}$ and $l_{2,1}$ regularization norms together are capable of selecting the task consistent features at the group level as well as at an individual level.

Besides the class-consistent features, there may exist some features (SNPs or imaging

QTs) those are relevant to a specific class. Then, the element-wise sparsity across all the classes is necessary to select the class-specific features. To select these features, the proposed method applies $l_{1,1}$ penalty, which is similar to l_1 but is applicable for the multi-task learning. This is defined for U and V, respectively, as follows:

$$\|U\|_{l_{1,1}} = \sum_{i=1}^{p} \|u_i\|_1 = \sum_{i=1}^{p} \sum_{j=1}^{C} |u_{ij}|;$$
$$\|V\|_{l_{1,1}} = \sum_{i=1}^{q} \|v_i\|_1 = \sum_{i=1}^{q} \sum_{j=1}^{C} |v_{ij}|.$$

Although the norm $l_{1,1}$ is non-smooth, it can be optimized since it is convex. Applying the $G_{2,1}$-norm, $l_{2,1}$-norm and $l_{1,1}$-norm regularizations for the canonical weights U and V, the proposed objective function can be expressed as follows:

$$\begin{aligned} &\min_{u_c, v_c} \quad -\sum_{c=1}^{C} u_c^T \Sigma_{xy} v_c \\ &\text{subject to} \quad u_c^T S_{W,x}^c u_c = 1, \quad v_c^T S_{W,y}^c v_c = 1, \\ &\|U\|_{G_{2,1}} < a_1, \quad \|U\|_{l_{2,1}} < b_1, \quad \|U\|_{l_{1,1}} < c_1, \\ &\|V\|_{G_{2,1}} < a_2, \quad \|V\|_{l_{2,1}} < b_2, \quad \|V\|_{l_{2,1}} < c_2; \end{aligned} \tag{6.7}$$

where $a_1, a_2, b_1, b_2, c_1, c_2$ are constants. The necessity of using different norms for the regularization of class-consistent and class-specific features is illustrated in Fig. 6.2

In this regard, it should be noted that unlike the MT-SCCALR model [63], the proposed method integrates judiciously the merits of CCA and LDA. In effect, the proposed model simultaneously has the power of maximizing correlation and can classify the diagnostic groups properly. Both $S_{W,x}^c$ and $S_{W,y}^c$ play an important role to identify the diagnose-specific features. On the other hand, in the MT-SCCALR [63], the multi-class logistic regression is combined with the MT-SCCA model [66], to identify the diagnose-specific features. However, the proposed method is easy to implement and has lower computational cost as compared to the MT-SCCALR model. Also, unlike the MT-SCCALR, the proposed model is capable of dealing with large number of features.

6.2.3 Analytical Solution

To solve the minimization problem of (6.7), the following Lagrangian function is considered:

$$\begin{aligned} L(U, V) = \sum_{c=1}^{C} \Big(&- u_c^T \Sigma_{xy} v_c + \lambda_1 (u_c^T S_{W,x}^c u_c - 1) + \lambda_2 (v_c^T S_{W,y}^c v_c - 1) \Big) \\ &+ \alpha_1 (\|U\|_{G_{2,1}} - a_1) + \alpha_2 (\|U\|_{l_{2,1}} - b_1) + \alpha_3 (\|U\|_{l_{1,1}} - c_1) \\ &+ \beta_1 (\|V\|_{G_{2,1}} - a_2) + \beta_2 (\|V\|_{l_{2,1}} - b_2) + \beta_3 (\|V\|_{l_{2,1}} - c_2); \end{aligned} \tag{6.8}$$

where λ_1, λ_2, α_1, α_2, α_3, β_1 β_2, and β_3 are the Lagrange multipliers of the model to be tuned. The parameters α_1 and α_2, respectively, control the group-level and feature-level

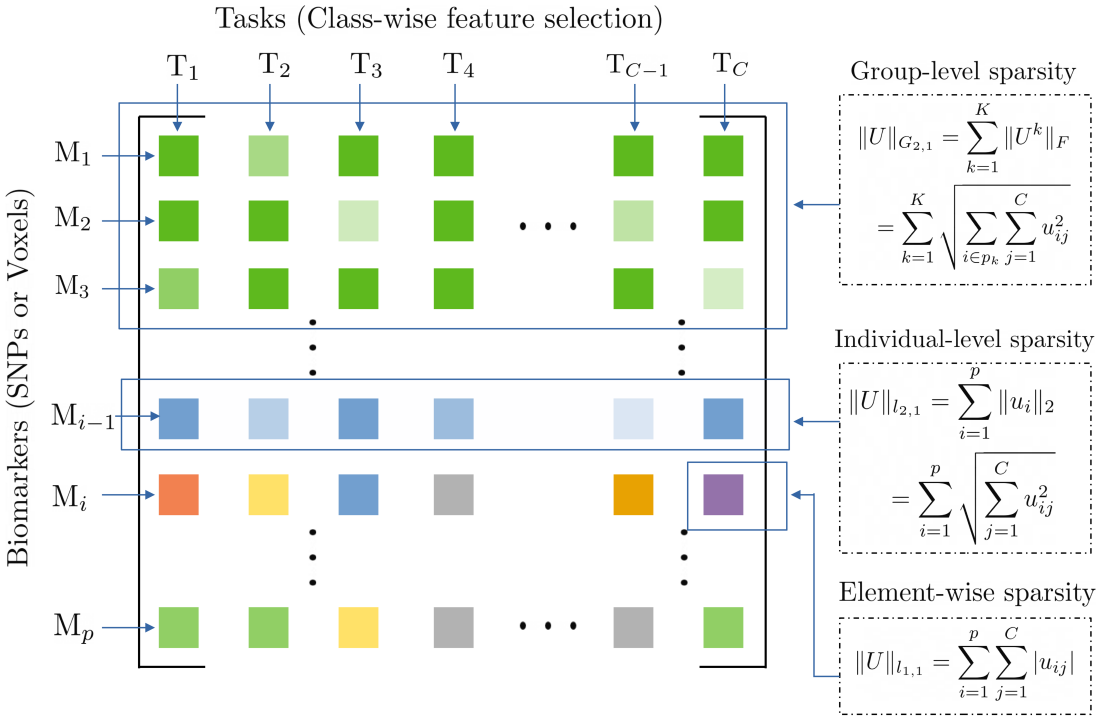


Figure 6.2: Illustration of different sparsity norms, included in the proposed method.

sparsity of the canonical weights U, while β_1 and β_2 are the group-level and feature-level sparsity parameters of the canonical weights V. On the other hand, α_3 and β_3 control the element-wise sparsity across all the classes for the canonical weights U and V, respectively.

The function $L(U, V)$ in (6.8) is a biconvex function, that is, convex in U when V is fixed and convex in V, when U is fixed. So, the problem can be solved using alternative optimization rule. Taking the derivative of $L(U, V)$ with respect to U, when V is fixed, and equating it to 0 gives

$$\begin{aligned} & X^T\mathbb{Y} - \lambda_1 S_{W,x}^c U - 2\alpha_1 D_{u0} U - 2\alpha_2 D_{u1} U - 2\alpha_3 D_{u2} U = 0 \\ \Rightarrow & \left(\frac{\lambda_1}{2} S_{W,x}^c + \alpha_1 D_{u0} + \alpha_2 D_{u1} + \alpha_3 D_{u2} \right) U = \frac{1}{2} X^T \mathbb{Y} \end{aligned}$$

$$U = \frac{1}{2} \left(\frac{\lambda_1}{2} S_{W,x}^c + \alpha_1 D_{u0} + \alpha_2 D_{u1} + \alpha_3 D_{u2} \right)^{-1} X^T \mathbb{Y}. \tag{6.9}$$

So, the updation rule for U is given by (6.9), where $\mathbb{Y} = [Yv_1\ Yv_2\ \ldots\ Yv_c]$, $2D_{u0}U$ is the subgradient of $\|U\|_{G_{2,1}}$ with respect to U and $2D_{u1}U$ is the subgradient of $\|U\|_{l_{2,1}}$ with respect to U. The matrix D_{u0} is a block diagonal matrix whose entries are $\frac{1}{2\|U^k\|_F} I_k$, $(k = 1, 2, \ldots K)$, where I_k is an identity matrix of size equal to the k-th subgroup of genetic marker. The matrix D_{u1} is a diagonal matrix whose entries are $\frac{1}{2\|u_i\|_2}$ $(i = 1, 2, \ldots p)$. Also,

the matrix D_{u2} is a diagonal matrix whose diagonal entries are $\frac{1}{|u_{ij}|}$ ($i = 1, 2, \ldots, p$ and $j = 1, 2, \ldots, C$). When $\|U^k\|_F = 0$, the k-th diagonal block of D_{u0} can be regularized as $\frac{1}{2\sqrt{\|U^k\|_F^2+\eta}}$, where η is a very small positive number. Similarly, D_{u1} and D_{u2} can be regularized.

Similarly, taking the derivative of $L(U, V)$ with respect to V, when U is fixed, and equating it to 0, gives the updation rule of V as follows:

$$V = \frac{1}{2}\left(\frac{\lambda_2}{2}S^c_{W,y} + \beta_1 D_{v0} + \beta_2 D_{v1} + \beta_3 D_{v2}\right)^{-1} Y^T\mathbb{X}; \tag{6.10}$$

where $\mathbb{X} = [Xu_1\ Xu_2\ \ldots\ Xu_c]$, $2D_{v0}V$ is the subgradient of $\|V\|_{G_{2,1}}$ with respect to V and $2D_{v1}V$ is the subgradient of $\|V\|_{l_{2,1}}$ with respect to V. The matrix D_{v0} is a block diagonal matrix whose entries are $\frac{1}{2\|V^l\|_F}I_l$, ($l = 1, 2, \ldots L$), where I_l is an identity matrix of size equal to the l-th ROI. The matrix D_{v1} is a diagonal matrix whose diagonal entries are $\frac{1}{2\|v_i\|_2}$, ($i = 1, 2, \ldots q_l$). Also, D_{v2} is a diagonal matrix whose diagonal entries are $\frac{1}{|v_{ij}|}$ ($i = 1, 2, \ldots, q$ and $j = 1, 2, \ldots, C$). When $\|V^l\|_F = 0$, the l-th diagonal block of D_{v0} can be regularized as $\frac{1}{2\sqrt{\|V^l\|_F^2+\eta}}$, where η is a very small positive number. Similarly, D_{v1} and D_{v2} can be regularized.

Both U and V can be computed efficiently by solving the following linear system of equations:

$$\left(\frac{\lambda_1}{2}S^c_{W,x} + \alpha_1 D_{u0} + \alpha_2 D_{u1} + \alpha_3 D_{u2}\right)U = \frac{1}{2}X^T\mathbb{Y} \tag{6.11}$$

$$\left(\frac{\lambda_1}{2}S^c_{W,y} + \beta_1 D_{v0} + \beta_2 D_{v1} + \beta_3 D_{v2}\right)V = \frac{1}{2}Y^T\mathbb{X}, \tag{6.12}$$

without involving the matrix inversion which is computationally more expensive [263]. The solution of the proposed optimization problem can now be obtained through iteratively and alternatively optimizing U and V. The basic steps are outlined in Algorithm 6.1, which converge to a local optimum depending on the initialization of U and V.

6.2.4 Computational Complexity Analysis

This section establishes that the proposed method has low computational complexity with respect to the number of features p and q of two data sets X and Y, respectively, and the number of common samples n. The computational complexity to initialize $p \times C$ canonical weight matrix U is $\mathcal{O}(pC)$, where C is the number of classes; while that to initialize $q \times C$ canonical weight matrix V is $\mathcal{O}(qC)$. The time complexity for both class binarization and class balancing is $\mathcal{O}(n)$. This step is to be executed C times. Therefore, the total complexity of Steps 2 and 3 is $\mathcal{O}((p + q + n)C)$.

In Step 5, one $p \times p$ block diagonal matrix D_{u0} and two $p \times p$ diagonal matrices D_{u1} and D_{u2} need to be computed with a total computational complexity of $\mathcal{O}(pC)$. In Step 6, the complexity to update U according to (6.11) is $\mathcal{O}(p(p^2 + qnC))$. Similarly, in Step 7, one $q \times q$ block diagonal matrix D_{v0} and two $q \times q$ diagonal matrices D_{v1} and D_{v2} need to be computed with a total computational complexity of $\mathcal{O}(qC)$. In Step 8, the complexity

Algorithm 6.1 Proposed Algorithm

1: **INPUT:** $X \in \mathbb{R}^{n\times p}$, $Y \in \mathbb{R}^{n\times q}$, λ_1, λ_2, α_1, α_2, α_3, β_1, β_2, β_3.
2: **Initialize** the canonical weights $U = [u_1, u_2, \ldots, u_c] \in \mathbb{R}^{p\times C}$, $V = [v_1, v_2, \ldots, v_c] \in \mathbb{R}^{q\times C}$.
3: Do class binarization and class balancing.
4: **while** not converge **do**
5: Calculate D_{u0}, D_{u1}, and D_{u2}.
6: Solve U according to (6.11) and normalize each u_c $(c = 1, 2, \ldots, C)$ such that $u_c^T S_{W,x}^c u_c = 1$.
7: Calculate D_{v0}, D_{v1}, and D_{v2}.
8: Solve V according to (6.12) and normalize each v_c $(c = 1, 2, \ldots, C)$ such that $v_c^T S_{W,y}^c v_c = 1$.
9: **end while**
10: Sort each u_c and v_c $(c = 1, 2, \ldots, C)$ in descending order according to the absolute values of their elements.

to update V according to (6.12) is $\mathcal{O}(q(q^2 + pnC))$. The Steps 5 to 8 are executed τ times, where τ is the number of iterations required to converge the proposed algorithm. So, the overall complexity of Steps 5 to 8 is $\mathcal{O}(\tau(pC + p(p^2 + qnC) + qC + q(q^2 + pnC)))$ $\approx \mathcal{O}(\tau(p^3 + q^3 + pqnC))$. Finally, Step 10 has $\mathcal{O}((p^2 + q^2)C)$ complexity. Hence, the overall computational complexity of the proposed multi-task sparse discriminant CCA is $\mathcal{O}((p+q+n)C + \tau(p^3+q^3+pqnC) + (p^2+q^2)C) \approx \mathcal{O}(\tau p^3)$, assuming $p > q$ and $n, C \ll p$.

In table 6.1, the computational complexity of different existing algorithms and the proposed method is compared. In this table, n represents the number of samples and M is the number of imaging modalities in case of MT-SCCAR and JCB-SCCA.

6.3 Experimental Results and Discussion

The performance of the proposed method is extensively compared with that of several related existing algorithms, namely, CoRe [86], CoopLe [59] , JSCCA [74], JCB-SCCA (joint-connectivity-based SCCA) [132], MT-SCCALR [63], and MT-SCCAR (multi-task SCCA and regression) [128]. Since the objective of the proposed method is not only to classify samples, but also to identify the important class-specific features, the one versus all classification model is considered for the comparative performance analysis. A five-fold cross-validation strategy is used to find the optimal parameters of the proposed model. The set of parameters, for which the average of mean canonical correlation coefficient (CCC) and classification accuracy (one versus all) is highest, is considered as the optimal parameter set [63]. Two different metrics, namely, CCC and classification accuracy, are used here to find the optimal parameter for the model, as CCC selects the features which maximize the correlation while classification accuracy emphasizes to select the class specific features with maximum correlation. In the current experiment, the mean CCC is computed as follows:

$$\frac{1}{5}\sum_{k=1}^{5}\sum_{c=1}^{C}\rho(Xu_c, Yv_c),$$

Table 6.1: Complexity Analysis of Proposed and Existing Algorithms

Different Method	Computational Complexity
CoRe [86]	$\mathcal{O}(\tau((p^3+p^2n)+(q^3+q^2n)))$
CoopLe [59]	$\mathcal{O}(\tau((p^3+p^2n)+(q^3+q^2n)))$
JSCCA [74]	$\mathcal{O}(\tau(p^3+pn\log(n)+npq^2))$
JCB-SCCA [132]	$\mathcal{O}(\tau((nq^2+p)p+(np^2+q)qM+qn\log(n)))$
MT-SCCALR [63]	$\mathcal{O}(\tau(n^3q^3+np^2q^2+p^3))$
MT-SCCAR [128]	$\mathcal{O}(\tau(p^3+(q^3+n^2pq)M))$
Proposed	$\mathcal{O}(\tau(p^3+q^3))$

where u_c and v_c are the c-th class-specific features. The proposed method has eight parameters, namely, $\lambda_1, \lambda_2, \alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3$. The fine tuning of this set of eight parameters together is computationally intensive. To reduce the computational cost, the parameters λ_1 and λ_2 are set to 2.0, since they mainly effect the magnitude of the weight matrices U and V, respectively. The remaining six parameters are tuned by the grid search strategy over the set $\{10^{-2}, 10^{-1}, 1, 10, 10^2\}$. The stopping criteria of all the methods are considered as $\|U^{\tau+1} - U^{\tau}\| \leqslant \epsilon$ and $\|V^{\tau+1} - V^{\tau}\| \leqslant \epsilon$, where τ is the iteration number and ϵ is the tolerance level, which is set to 10^{-5} empirically.

6.3.1 Simulation Study

This section presents the comparative performance analysis of different methods on six simulated data sets.

6.3.1.1 Data Generation

A set of six simulated data is build to show the strength of the proposed method in identifying the bi-multivariate association of the imaging genetic data by selecting the diagnose-specific features. In all the simulations, three diagnostic groups are considered. Let us assume that each data set consists of n samples, and n_i is the number of samples in the i-th class such that $n_1 + n_2 + n_3 = n$. First, a latent variable model, similar to [194], is used to correlate the SNP and fMRI data. Let the latent vector be $l \in \mathbb{R}^n$, which is generated from the normal distribution $\mathcal{N}(0, 1)$. Then, two sparse canonical weight matrices, namely, $U = [u_1, u_2, u_3] \in \mathbb{R}^{p\times 3}$ and $V = [v_1, v_2, v_3] \in \mathbb{R}^{q\times 3}$ are generated. Based on the latent vector l and the canonical weights U and V, four pairs of data sets $X_c \sim \mathcal{N}(l_i u_c, \sigma_x I_{p\times p})$ and $Y_c \sim \mathcal{N}(l_i v_c, \sigma_y I_{q\times q})$ are generated for each class c, and further concatenated to form the SNP data X and fMRI data Y, where σ_x and σ_y are the noise strength to X and Y, respectively. The SNP data X is further recoded into $\{0, 1, 2\}$ using a binomial distribution $B(2, \text{logit}^{-1}(X + \text{logit}(\kappa)))$, where $\text{logit}(\kappa) = \log(\frac{\kappa}{1-\kappa})$ and κ is the minor allele frequency of SNP, generated from the uniform distribution $U(0.2, 0.4)$.

DS-A: Following the above procedure, a data set namely, DS-A, is generated where the SNP data X and Imaging data Y have dimensions $(150, 300)$ and $(150, 450)$. The DS-A is simulated in such a way that both the class-consistent and class-specific features are being present in the data. Among 150 samples, each of the three classes has 50 samples.

Also, the group structures for both the SNP and fMRI data are given as prior information. The SNPs in the data are grouped into 15 genes, each having size of 20, and each voxel from the fMRI data is grouped into 15 ROIs, each having size of 30. The noise strength, while generating DS-A, is considered as $\sigma_x = \sigma_y = 0.5$. To show the importance of class-consistent and class-specific features separately, three more data sets, namely, DS-A1, DS-A2 and DS-A3, are generated following the same procedure. The descriptions of these data sets and the corresponding results are given in Section I and Section II, respectively, of the supplementary material of [177].

DS-B: Another data set, namely, DS-B is generated to mimic the real high dimensional data in the field. The data set is simulated to reflect the class-specific and class-consistence feature across three classes. The dimensions of SNP data X and imaging data Y are set to $(300, 20000)$ and $(300, 3000)$. Each class has 100 samples. The SNPs are grouped into 40 genes, each having size of 500 and the voxels in the fMRI data are grouped into 30 ROIs, each having size of 1000. In this case the following three properties hold:

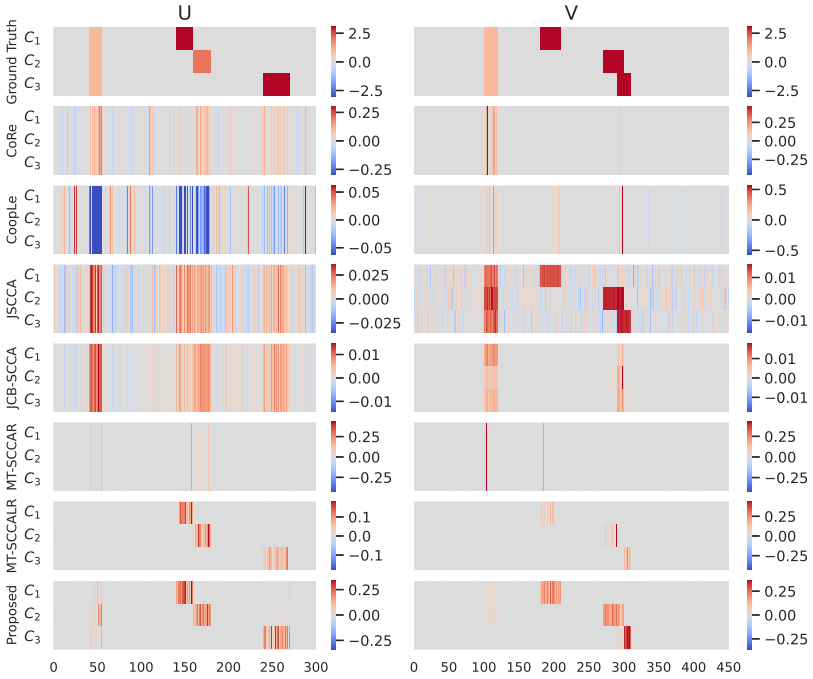
- To mimic the group effect of the real genetic and imaging data, two covariance matrices Σ_{xx} and Σ_{yy} are generated. The correlation between the correlated variables belonging to the same group is set to be from the uniform distribution $U(0.1, 0.3)$. Then, for each class c, $X_c \sim \mathcal{N}(l_i u_c, \Sigma_{xx})$ and $Y_c \sim \mathcal{N}(l_i v_c, \Sigma_{yy})$ are generated and concatenated to form the SNP and voxel data.
- X_c and Y_c are generated from c independent normal distributions to keep the labels independent to each other.
- The minor allele frequency (κ) is generated from the uniform distribution $U(0.01, 0.1)$ to analyze the rare variant of SNPs, while recoding the SNP data into $\{0, 1, 2\}$.

Following the same procedure, another data set, namely, DS-B1, is generated. The details of this data set and the corresponding results are given in Section I and Section II, respectively, of the supplementary material of [177].

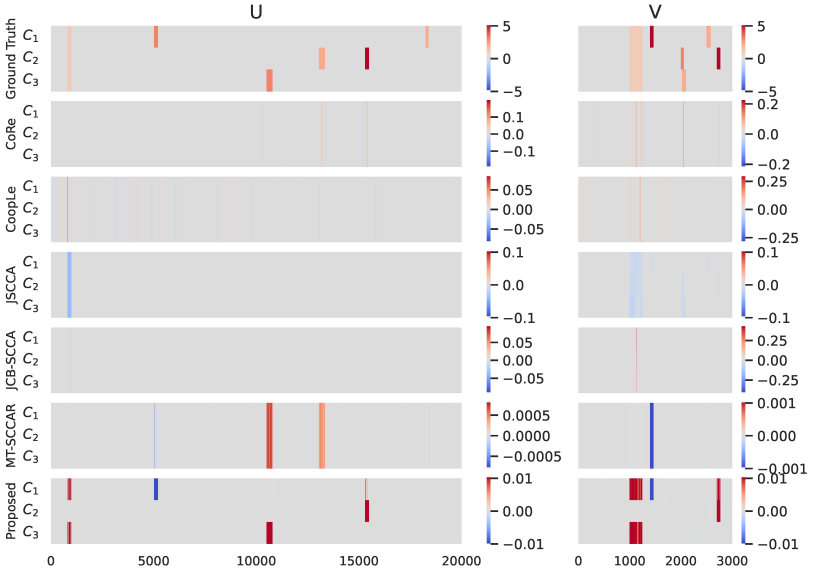
6.3.1.2 Accuracy and Correlation Analysis

To evaluate the performance of the proposed method with respect to the existing algorithms, the feature selection ability is shown through heatmaps in Fig. 6.3, while mean testing canonical correlation coefficient (CCC) and mean testing classification accuracy (one versus all) are computed and represented by bar graphs in Fig. 6.4. To better understand the difference between the existing algorithms and the proposed method, the receiver operating characteristic (ROC) curve is presented in Fig. 6.5 by varying the cut-off of the canonical loadings.

In Fig. 6.3, the canonical weights U and V are represented by the heatmaps indicating the importance of features. Since both JSCCA and JCB-SCCA generate one canonical weight vector u for all the task and C canonical weights $\{v_c\}$ for C tasks, u is stacked C times to make its heatmaps comparable. Also, as CoRe, CoopLe and MT-SCCAR generate one canonical weight vector for all the tasks, both u and v are stacked C times to draw the heatmaps. From Fig. 6.3, it is seen that the proposed method is able to identify the features which have higher degree of similarity with the ground truth than any other existing methods.



(a) DS-A



(b) DS-B

Figure 6.3: Heatmaps showing comparison of the canonical weights U and V for two data set, namely, DS-A and DS-B. The rows 1 to 8 depict the results corresponding to the ground truth, CoRe, CoopLe, JSCCA, JCB-SCCA, MT-SCCAR, MT-SCCALR, and the proposed method, while the columns 1 and 2 represent the canonical weights U and V, respectively. Each row contains three task (C_1, C_2, C_3) specific features. The features are shown by taking the mean computed over five-fold cross-validation.

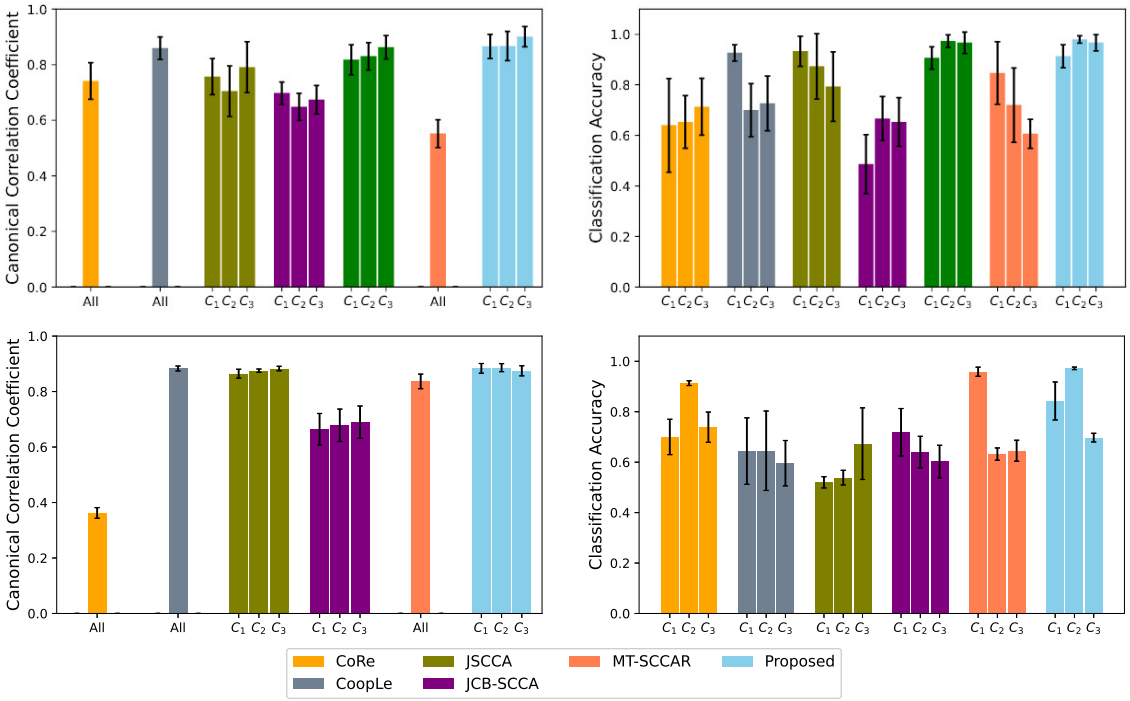


Figure 6.4: Comparative performance analysis of different methods with respect to mean testing canonical correlation coefficient (left-column) and mean classification accuracy (right-column) on two data sets, namely, DS-A (top row) and DS-B (bottom row) obtained over five-fold cross-validation.

For the data set DS-A, the MT-SCCALR has identified only the class-specific features and overlooked the shared information, while the proposed method has captured both the class-specific and shared information across all the classes. For the data set DS-B, MT-SCCALR is out of its program limit due to the large number of SNPs and voxels. In this case JSCCA has captured both the shared and class-specific information for the canonical weight V, but captured only the shared information for canonical weight U. All other methods are capable of selecting a small amount of shared information only, but unable to select the class-specific information across the three class. But, the proposed method has performed better in case of selecting the class-consistent and class-specific features for both the canonical weights U and V. From all the results reported in Fig. 6.3, it can be concluded that the proposed method has more diverse feature selection ability than other six existing algorithms.

In Fig. 6.4, the mean testing CCCs and mean classification accuracy are shown to signify the association between the SNPs and voxels. The support vector machine (SVM), of LIBSVM software package, is used to find the classification accuracy of all the methods. The classification accuracy is obtained by selecting the top ten (10) generated features of imaging and genetic data, and then concatenating them. In top row of Fig. 6.4, the graphs of test CCCs and classification accuracy signify that the proposed method attains the highest correlation and comparatively better classification accuracy among all the methods in case of DS-A. Besides the proposed method, MT-SCCALR has also performed well in

case of DS-A. On the other hand, for DS-B data set, the proposed method as well as JSCCA have comparatively performed better in capturing the canonical correlation coefficient. But, JSCCA has performed poor in case of obtaining classification accuracy. This happens because the feature selection for U, by JSSCA is highly biased towards class-consistent features. In effect, the top 20 features of SNP data come from class-consistent features. In

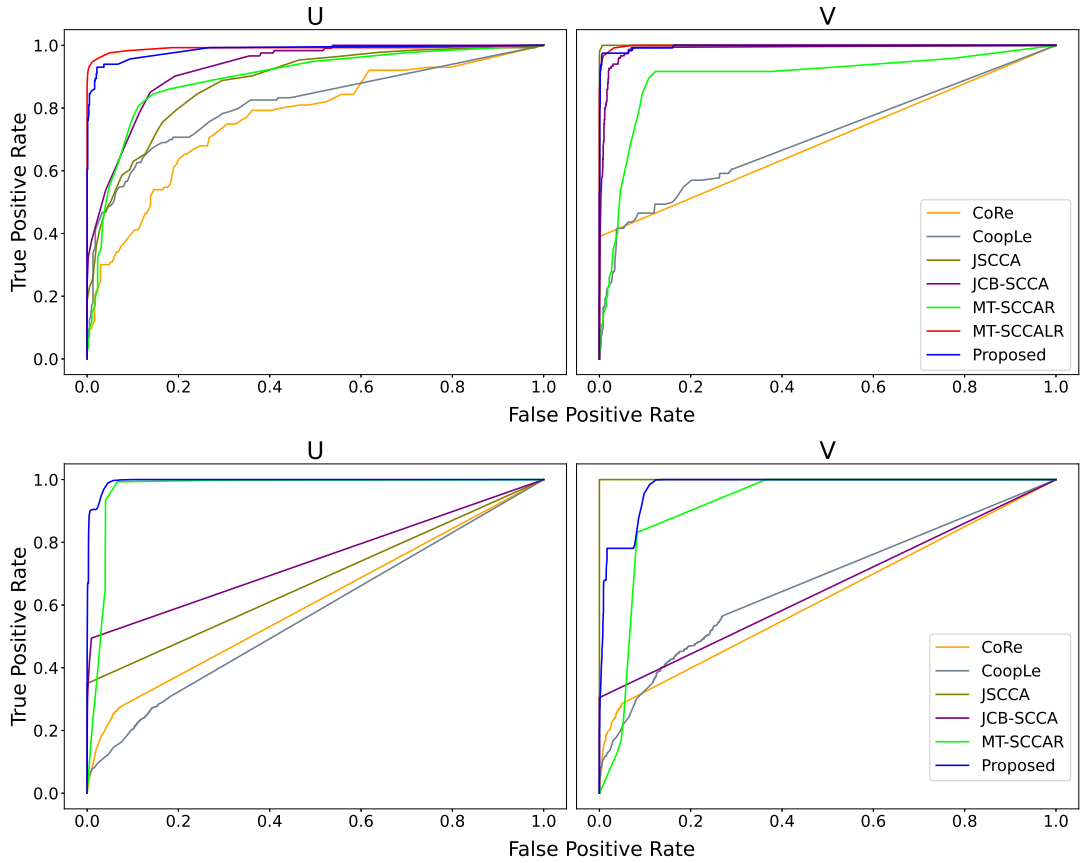


Figure 6.5: ROC curves of different methods on two data sets, namely, DS-A (top row) and DS-B (bottom row). Column 1 depicts FPR versus TPR on detection of canonical SNPs (U), and column 2 depicts FPR versus TPR on detection of canonical Voxels (V).

top row of Fig. 6.5, the graph of receiver operating characteristics (ROC) curves for DS-A shows that MT-SCCALR has highest true positive rate (TPR) than all other methods in case of selecting the canonical weight U, whereas JSCCA has the best TPR values for selecting the canonical weight V, with lower false positive rate (FPR). But, in both the cases, MT-SCCALR and the proposed methods are consistent enough to have the descent TPR value and JSCCA perform poorly in case of the canonical weight U. On the other hand the ROC curve for DS-B presented in the bottom row of Fig. 6.5, shows that the proposed method has highest TPR values in case of the canonical wight U, but in case of V, JSCCA has the best TPR values with low FPR values. All other methods have poor TPR values as the FPR increases.

All the results demonstrate that although there exist several existing methods, such as MT-SCCALR and JSCCA, which provide similar performance to that of the proposed

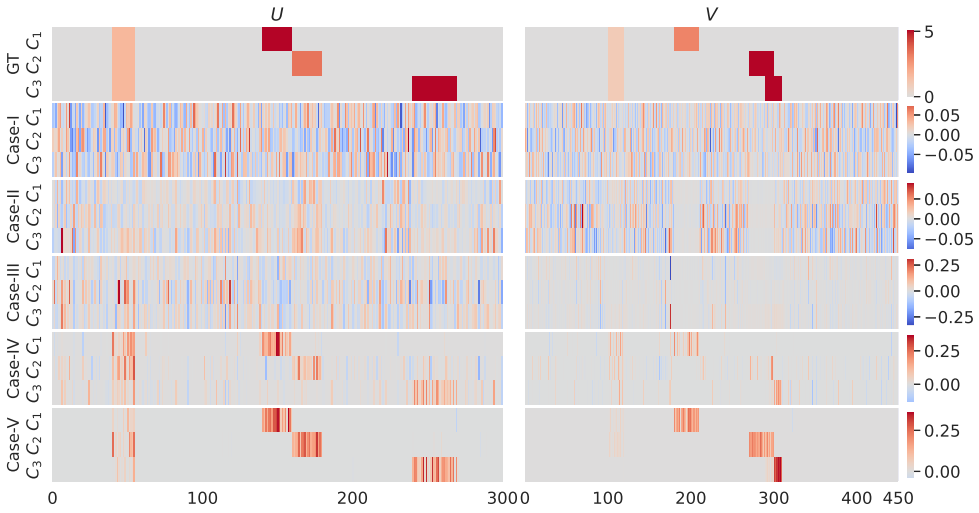
method, with respect to CCCs, classification accuracy and ROC curves, the proposed method has the better class-specific feature selection ability than all other methods.

6.3.1.3 Ablation Study

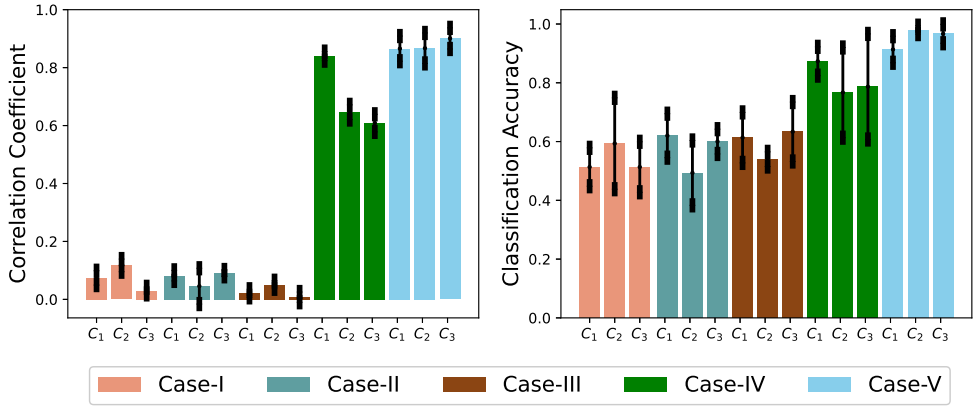
In this section, the importance of including different regularizer terms into the objective function of the proposed method is evaluated by the capability of selecting class-specific and class-consistent features, mean canonical correlation coefficient and mean classification accuracy. The corresponding results are shown in Fig. 6.6. Depending on the three regularizer terms, the experiments are conducted for the following five cases:

- Case-I: None of the three regularizer terms is present in the objective function; that is, α_1,α_2, α_3, β_1, β_2, and β_3 all are set to zero.
- Case-II: The regularizer term, which controls the group-level sparsity, is only present, while the other regularizer terms are absent, that is, $\alpha_1 \neq 0$ and $\beta_1 \neq 0$, while all the other hyperparameters, α_2, α_3, β_2, and β_3, are set to zero.
- Case-III: The regularizer term, which controls the individual-level sparsity, is only present, while the other regularizer terms are absent, that is, $\alpha_2 \neq 0$ and $\beta_2 \neq 0$, while all the other hyperparameters α_1, α_3, β_1, and β_3 are set to zero.
- Case-IV: The regularizer term, which controls the element-wise sparsity, is only present, while the other regularizer terms are absent, that is, $\alpha_3 \neq 0$ and $\beta_3 \neq 0$, while all the other hyperparameters, α_1, α_2, β_1, and β_1, are zero.
- Case-V: All three regularizer terms are present in the objective function; that is, all of the hyperparameters, α_1, α_2, α_3, β_1, β_2, and β_3, are nonzero. This is the proposed formulation.

It is observed from Fig. 6.6 that when there are no regularizer terms involved in the objective function of the proposed method, that is, in Case-I, the proposed method is unable to identify the class-consistent and class-specific features, which is reflected in the heatmap of the canonical weights U and V in Fig. 6.6a. For Case-II and Case-III, where either the group-level sparsity or individual-level sparsity is present in the objective function, respectively, the proposed model still cannot identify the ground truth features for the canonical weights U and V, which is reflected in Fig. 6.6a. These result in poor correlation coefficient and classification accuracy for the cases I, II, and III, which can be observed in Fig. 6.6b. On the other hand, for Case-IV, when the element-wise sparsity is included in the proposed model, the model is able to identify the class-consistent and class-specific features, but it has captured some unwanted noisy features also. At last, for Case-V, when all the regularizer terms are present in the objective function, the performance of the proposed method, in terms of selecting the class-consistent and class-specific features, capturing correlation, and classifying one class from the others, has increased significantly, which can be shown in Fig. 6.6a and Fig. 6.6b.



(a)

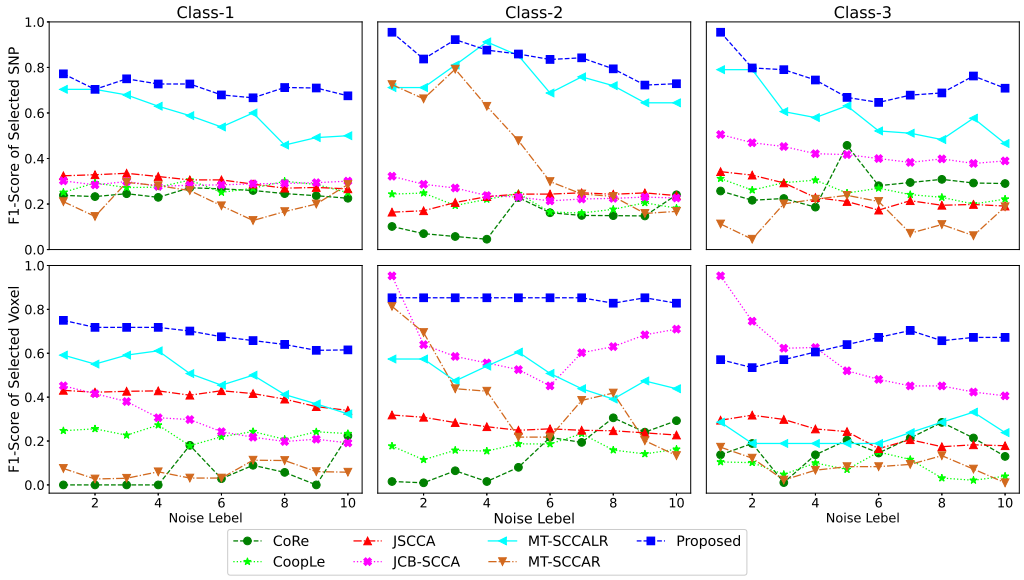


(b)

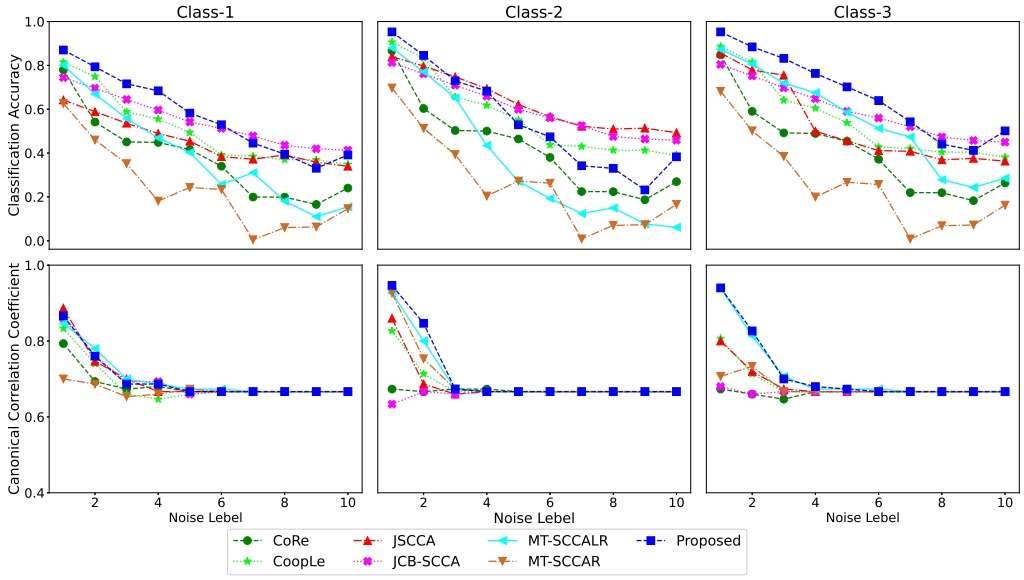
Figure 6.6: The performance of the proposed method on DS-A data set under different regularizer terms. (a) Heatmaps showing comparison of the canonical weights U and V. The rows 1 to 6 depict the results corresponding to the ground truth (GT), and different combinations of the proposed method presented through Case-I to Case-V respectively. (b) Comparison with respect to mean testing canonical coefficient (left-column) and mean classification accuracy (right-column).

6.3.1.4 Noise Sensitivity Analysis

To analyze the noise sensitivity of the proposed method over the existing methods, the noise strengths σ_x and σ_y are varied from 1 to 10 for the synthetic data DS-A. The F1-scores, along with mean CCC and mean classification accuracy, are used to evaluate the performance of different methods. The variation of F1-score with respect to noise level is studied in Fig. 6.7(a). It indicates that the F1-score of the proposed method deceases slowly with the increase in noise level. However, the proposed method attains a higher F1-score as compared to all the existing methods, irrespective of the noise levels, modalities and



(a)



(b)

Figure 6.7: Comparative performance analysis of different methods on DS-A data set under different noise levels, with respect to (a) F1 score; and (b) CCCs and classification accuracy. In (a), the 1st row represents the results corresponding to the selected SNPs, while the 2nd row shows that of selected canonical voxels.

classes. Finally, Fig. 6.7(b) shows the variation of both CCCs and classification accuracy with respect to noise level. The same optimal parameters, obtained over the data set DS-A, has been used to compute the CCCs and classification accuracy of the proposed method for all the noisy synthetic data sets. All the results reported here confirm that both CCCs and accuracy of all the methods decrease with the increase in noise level. The proposed

method achieves better performance with respect to CCCs for classes 1 and 3, while both JSCCA and JCB-SCCA provide better performance for class 2, in most of the noise levels. However, in case of the proposed method, the computed CCCs are more relevant since the CCCs are captured by selecting the class-specific features, whereas the CCCs captured by other models, namely, JSCCA, JCB-SCCA and MT-SCCAR, are highly similar for all the classes. The performance of the proposed method and MT-SCCALR, with respect to classification accuracy, is quite similar to each other, and after the noise level of 5, it becomes similar for all the models.

6.3.2 Performance on Real Neuroimaging Genetic Data

This section presents the performance of different methods on real neuroimaging and genetic data, obtained from the ADNI database (adni.loni.usc.edu). A total of 120 participants are included in the current study whose resting state fMRI and SNP data are collected from the ADNI cohort. The data acquisition and preprocessing are described in Section 5.2 of Chapter 5. The performance of the proposed method is compared with that of CoRe, CoopLe, JSCCA, JCB-SCCA and MT-SCCAR. The MT-SCCALR is excluded from the comparative analysis since the number of voxels and SNP counts are large enough to exceed the program limit of the MT-SCCALR. In this regard, it should be noted that the proposed method is capable of handling large number of features, consumes less memory and computationally less expensive than the MT-SCCALR.

6.3.2.1 Imaging Genetic Association and Classification

The performance of the proposed method is extensively compared with that of several existing methods, with respect to mean testing CCCs and classification accuracy, and the corresponding results are reported in Table 6.2. Higher value of CCC indicates the stronger imaging genetic association. It is seen from the results reported in Table 6.2 that both JSCCA and MT-SCCALR attain the correlation values near to zero. On the other hand, the JCB-SCCA achieves a correlation value of 0.15, 0.14, 0.18 and 0.15 with respect to CN, EMCI, LMCI and AD, respectively, but the standard deviation of the correlation computed over five-fold cross-validation is quite high. Similar results are observed in case of CoRe and CoopLe. However, the proposed method achieves the best correlation value with low standard deviation, irrespective of the diagnostic groups. This result also suggests that the correlation is significant for the imaging genetic association.

The classification accuracy reported in Table 6.2 is obtained by using the SVM. The top ten SNPs and imaging QTs are considered, and then concatenated to generate 20 features. All the results reported in Table 6.2 show that the proposed method attains highest classification accuracy (one versus all) for the three groups, namely, EMCI, LMCI and AD. It signifies that the class-wise feature selection, which can distinguish one class over the other, of the proposed method is comparatively better than the feature selection of all other existing methods. The classification results also indicate that it is very difficult to separate different diagnostic groups for the real imaging genetic data using resting state fMRI and SNP data. However, it can also be seen that the correlation obtained by the proposed method, by the selection of meaningful genes and ROIs, is significant. Overall, the proposed method shows a promising result in multi-class imaging genetic problem.

Table 6.2: Canonical Correlation Coefficient and Classification Accuracy (mean ± std) for Real Data

Metric	Algorithms	CN	EMCI	LMCI	AD
Canonical Correlation Coefficient	CoRe [86]	0.1110 ± 0.2063	0.1110 ± 0.2063	0.1110 ± 0.2063	0.1110 ± 0.2063
	CoopLe [59]	0.1307 ± 0.2524	0.1307 ± 0.2524	0.1307 ± 0.2524	0.1307 ± 0.2524
	JSCCA [74]	0.0243 ± 0.2996	0.0163 ± 0.1624	0.0100 ± 0.2872	0.0309 ± 0.1720
	JCB-SCCA [132]	0.1517 ± 0.2621	0.1438 ± 0.2517	0.1804 ± 0.2094	0.1519 ± 0.2656
	MT-SCCAR [128]	0.0957 ± 0.1058	0.0957 ± 0.1058	0.0957 ± 0.1058	0.0957 ± 0.1058
	Proposed	**0.2556 ± 0.0968**	**0.1694 ± 0.1011**	**0.1747 ± 0.0797**	**0.2339 ± 0.0476**
Classification Accuracy	CoRe [86]	0.5000 ± 0.1490	0.5583 ± 0.0424	0.4416 ± 0.0772	0.5583 ± 0.0424
	CoopLe [59]	0.5000 ± 0.0874	0.5250 ± 0.0857	0.5500 ± 0.0964	0.4416 ± 0.0857
	JSCCA [74]	**0.6750 ± 0.0485**	0.5250 ± 0.0772	0.4333 ± 0.0971	0.5667 ± 0.1307
	JCB-SCCA [132]	0.5250 ± 0.0677	0.5250 ± 0.0971	0.5500 ± 0.0849	0.5917 ± 0.1000
	MT-SCCAR [128]	0.5333 ± 0.1034	0.5083 ± 0.0716	0.5416 ± 0.1020	0.5750 ± 0.0807
	Proposed	0.5833 ± 0.0950	**0.5916 ± 0.0927**	**0.5916 ± 0.0889**	**0.6000 ± 0.0857**

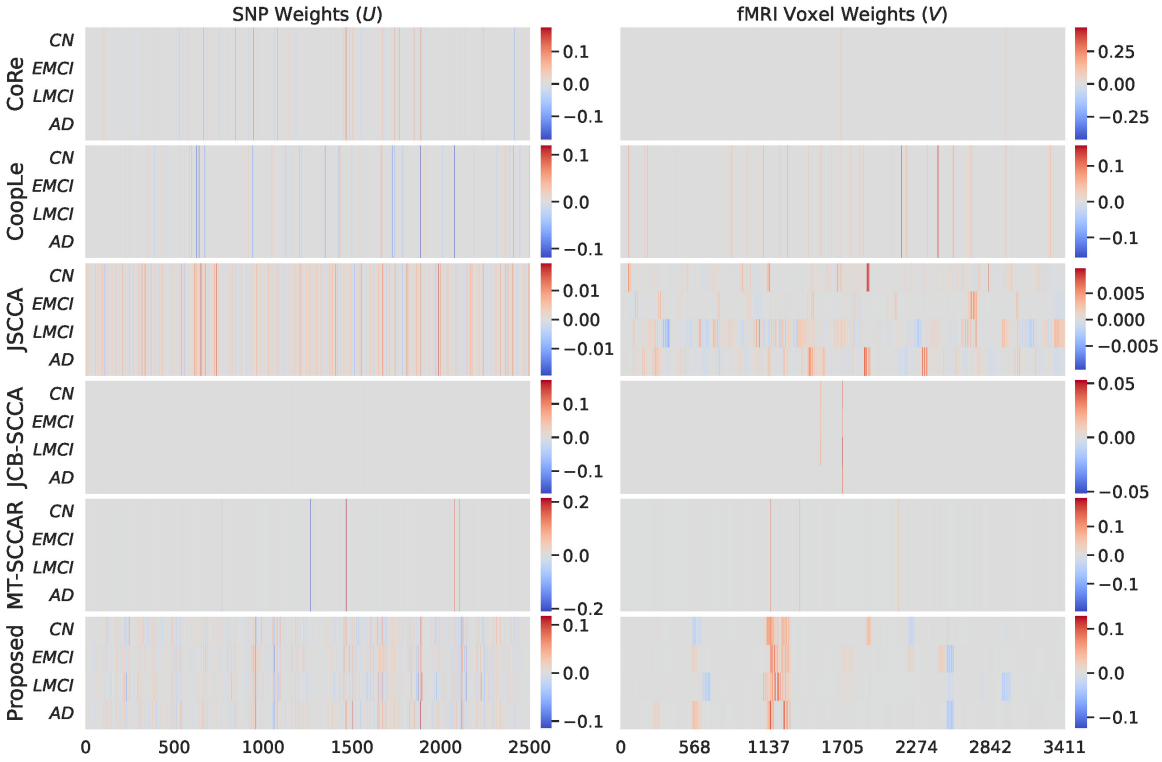
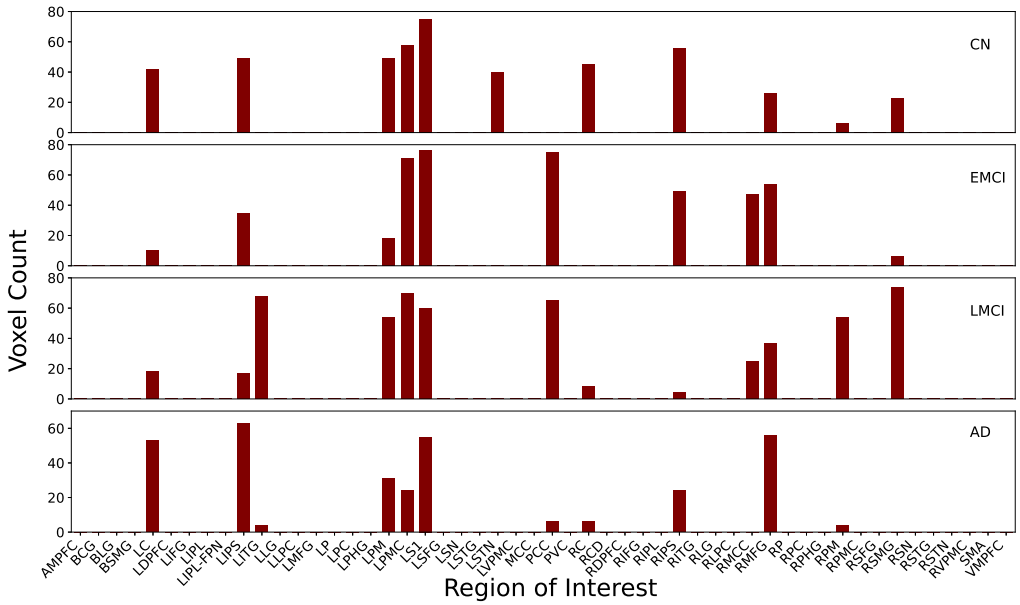


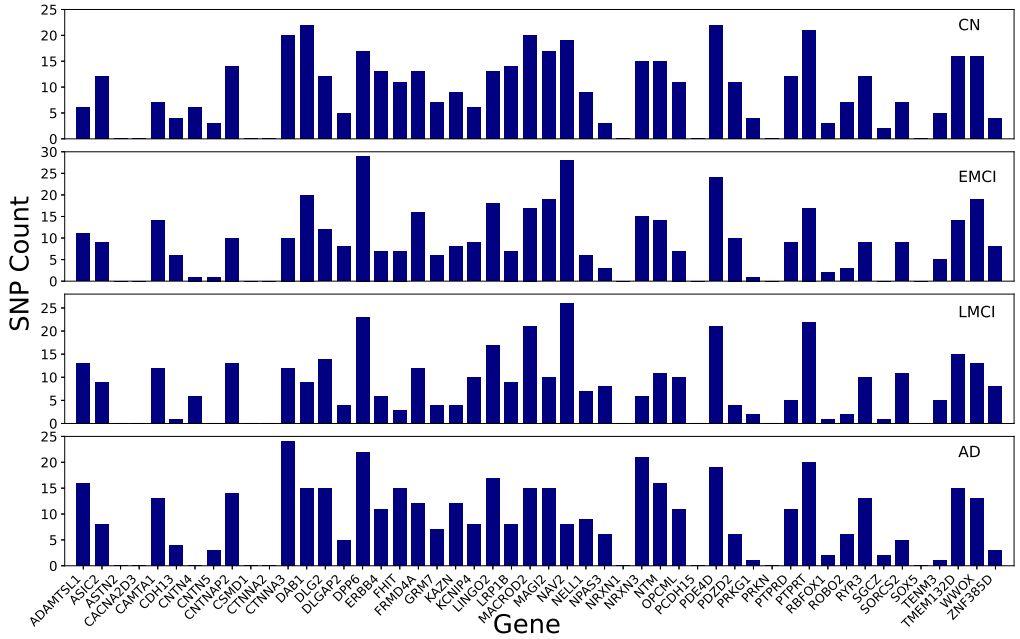
Figure 6.8: Mean canonical weights of SNPs (U: left column) and voxels (V: right column) obtained over five-fold cross-validation. Rows 1 to 5 correspond to the canonical weights for the five existing methods, namely, CoRe, CoopLe, JSCCA, JCB-SCCA and MT-SCCAR, and proposed method.

6.3.2.2 Interpretation of Selected Brain Regions

The imaging QTs, selected by the proposed method, are shown to be correlated significantly with the progression of AD. The selected brain regions are shown in the right part of the heatmap in Fig. 6.8. Most of the brain regions identified by the proposed method are class-



(a)



(b)

Figure 6.9: (a) Number of voxels in 52 brain regions included in RSN. Row:1-4 frequency of different ROIs in CN, EMCI, LMCI, AD group respectively. The higher frequency of the ROI is, the more contribution to the correlation between the group and ROI.; (b) Top fifty genes with corresponding SNP count. Row:1-4 frequency of different genes in CN, EMCI, LMCI, AD group respectively. Higher the frequency is, the more contribution to correlation of the gene associated with the respective group.

consistent and significant with respect to AD. The brain regions selected by the existing methods are not significant enough as evident from Table 6.2. The mean CCC near to zero means the selected ROIs and SNPs are not correlated. The frequency of different ROIs belonging to RSN for each diagnostic group is also shown in Fig. 6.9(a). The common ROIs selected by the proposed method are left/right intraparietal sulcus (LIPS, RIPS), left putamen (LPM), left primary motor cortex (LPMC), left primary somatosensory cortex (LS1), and right middle frontal gyrus (RMFG). It has been shown in the existing works that IPS [211] [149], LPM [160] and LPMC [244] are associated with the progression of AD. In [236], it has been suggested that LS1 is affected early in the progression of AD and may have some consequence on behavioral and functional measures. The functional activity in RMFG is also associated with the AD, which has been shown in [290]. The most affected and significant brain region in AD is posterior singulate cortex (PCC), which is specifically selected here for the diagnostic group EMCI and also occur in both LMCI and AD. It is the evidence of starting cognitive declination in EMCI from the healthy group [220, 291, 293]. The PCC is a central part in DMN and plays an important role in internally directed cognition [34, 207]. The reduced functional connectivity in PCC leads to occur AD [35]. So, the selection of PCC is significant. The right middle singulate cortex (RMCC) is selected in EMCI group and different from ROIs of CN group. It is considered to be one the affected brain regions in AD and justifies the cognitive degradation in EMCI [41]. The RMCC is also selected in LMCI group. The left inferior temporal gyrus (LITG) is selected in case of LMCI and AD only, which is also significant according to [219]. There are two more regions, namely, left and right claustrum (LC, RC), selected in LMCI and AD groups, which are reported to be affected by AD connected with the entorhinal cortex [180, 258]. Overall, it can be concluded that although there are not many class-specific ROIs found and most of the selected features are shared by all the classes; but they are proved to be associated with the progression of AD.

6.3.2.3 Interpretation of Selected SNPs and Corresponding Genes

The selected SNPs are shown by the heatmaps in the left sides of Fig. 6.8. To get a clear view of the diagnose-specific SNPs, the top 20 SNPs, selected according to their canonical weights, and the corresponding genes are shown in Table 6.3. It shows that the AD specific SNP rs7931452 gets the highest canonical weight and it comes from the gene neuron navigator 2 (NAV2). There are three more SNPs, namely, rs11025246, rs10766573 and rs11025105, which are selected from NAV2 in the top 20 SNPs. According to [264], NAV2 gene is highly expressed in brain and involved in the nervous system development. It is a risk gene in AD. The frequency of occurrence of different genes in four diagnostic groups is shown in Fig. 6.9(b). The most frequent genes occurring in AD group are CTNNA3 [171], DPP6 [36], NRXN3 [313], PDE4D [253], LINGO2 [170], ADAMTSL1 [22], NTM [193], DAB1 [77, 182], DLG2 [200], and MACROD2 [136], which are directly or indirectly related with AD. In case of LMCI, the most frequent genes are NAV2, DPP6 [36], PTPRT, MACROD2, PDE4D, LINGO2, TMEM132D, DLG2, ADAMTSL1, and CNTNAP2. From Fig. 6.9(b), it can also be seen that all the above mentioned genes are selected more or less with respect to all the diagnostic groups. It suggests that although these genes cannot give an insight to distinguish the four groups by their selection, but they are responsible for the progression of AD and are significant towards class-specific correlation, which is shown in

Table 6.3: Top Twenty SNPs, along with corresponding Genes and Weights, Selected by the Proposed Method for Four Diagnostic Groups

CN			**EMCI**		
GENE	SNP	Weight	GENE	SNP	Weight
PDE4D	rs6865647	0.078804	PTPRT	rs6016732	-0.090525
TMEM132D	rs1872712	0.075821	PDE4D	rs6865647	0.08488
PDE4D	rs10068809	0.070544	NAV2	rs10766573	0.083776
NAV2	rs10766573	0.069887	TMEM132D	rs1872712	0.082469
NAV2	rs7931452	0.06943	PTPRT	rs6016752	0.072905
MACROD2	rs444594	0.068855	NAV2	rs2012651	0.068071
CTNNA3	rs10509244	-0.057769	DAB1	rs155294	0.06444
PTPRT	rs6016752	0.057724	DAB1	rs17115430	0.063024
MACROD2	rs6042778	-0.054714	NAV2	rs7931824	0.061202
DPP6	rs6464378	0.053829	PDE4D	rs1078368	-0.058396
NTM	rs7951332	0.052471	DAB1	rs12132898	0.054784
TMEM132D	rs2292723	-0.05208	PDE4D	rs27184	-0.054779
NAV2	rs1364792	0.051959	PDE4D	rs27183	-0.053174
MACROD2	rs6034011	0.051011	PDE4D	rs2910641	-0.053147
FRMD4A	rs4748054	-0.047646	NTM	rs10894417	0.052807
DAB1	rs17115430	-0.047507	PDE4D	rs10068809	0.051554
DAB1	rs10493218	-0.046891	PDE4D	rs17721878	-0.051046
PTPRT	rs6016732	-0.046427	NAV2	rs11025246	0.050735
WWOX	rs6564559	0.046128	NAV2	rs7931452	0.050538
PTPRT	rs6102795	0.043723	LINGO2	rs10968246	0.050032
LMCI			**AD**		
GENE	SNP	Weight	GENE	SNP	Weight
NAV2	rs10766573	0.11348	NAV2	rs7931452	0.11662
PDE4D	rs6865647	0.107959	LINGO2	rs10968246	0.08933
NAV2	rs1364792	0.08444	PDE4D	rs10068809	0.087625
TMEM132D	rs1872712	0.080466	TMEM132D	rs1872712	0.083483
MACROD2	rs444594	0.074877	PTPRT	rs6016752	0.07748
MACROD2	rs6074737	-0.074707	LINGO2	rs7018499	0.074052
NAV2	rs7931824	0.073344	MACROD2	rs444594	0.072993
FRMD4A	rs7095537	-0.071283	TMEM132D	rs1386216	0.067443
PDE4D	rs27184	-0.071114	NAV2	rs11025246	0.067382
PDE4D	rs27183	-0.069814	NTM	rs7951332	0.065785
NAV2	rs2012651	0.063346	PTPRT	rs6016732	-0.065705
CTNNA3	rs13376837	-0.059706	MAGI2	rs3779312	-0.059982
NAV2	rs11025239	0.056449	PDE4D	rs2112957	0.056813
PTPRT	rs6016752	0.055904	NRXN3	rs3861630	0.055988
NAV2	rs11025158	-0.055325	TMEM132D	rs2292723	-0.055084
NAV2	rs4757842	-0.051023	NAV2	rs10766573	0.05445
MACROD2	rs6079391	-0.050938	DAB1	rs10489468	0.050205
PDE4D	rs10068809	0.050515	DPP6	rs6464378	0.050178
RYR3	rs2442463	-0.050287	MACROD2	rs10485771	0.050164
DAB1	rs10493220	-0.049653	NAV2	rs11025105	0.049452

Table 6.3.

6.3.3 Execution Time and Memory Usage

Finally, the performance of different methods is compared with respect to execution time and memory consumption. The corresponding results are reported in Table 6.4. The

Table 6.4: Execution Time and Memory Usage of Different Algorithms

Different	Execution Time (in second)			Memory usage (in Gigabytes)		
Algorithms	DS-A	DS-B	ADNI	DS-A	DS-B	ADNI
CoRe [86]	16.80	1398.10	53.14	0.51	0.83	0.57
CoopLe [59]	15.31	1220.00	71.40	0.51	0.83	0.57
JSCCA [74]	9.78	145.30	114.01	0.57	1.66	0.71
JCB-SCCA [132]	0.21	675.45	74.74	0.57	7.16	0.99
MT-SCCALR [63]	316.12	-	-	2.68	-	-
MT-SCCAR [128]	1.66	1582.60	172.38	0.71	12.16	1.15
Proposed	7.34	3186.50	550.98	0.77	21.24	1.53

reported execution time of proposed method is given for the optimal parameter to make the result consistent with all the existing algorithms. From the results, it can be seen that although JCB-SCCA and MT-SCCAR need lower time than the proposed method for convergence, both of them perform poorly, irrespective of the data sets used. Due to the lower computational cost, the proposed method needs significantly lower execution time than the MT-SCCALR. In fact, it is not possible to execute the MT-SCCALR on DS-B and ADNI data sets, as the number of voxels and SNP counts are large enough to exceed the program limit of the MT-SCCALR. Though the methods JSCCA, CoRe, and CoopLe take more time than the proposed method to execute on DS-A, but their execution time on DS-B and ADNI is very less compare to the proposed method. The reason behind this is that in the algorithm of JSCCA, main steps are to execute the fused lasso solver for updating V in each iteration and for the initialization of U, singular value decomposition is used once, while in case of proposed method the main step is to solve the linear system of equation twice for updating U and V respectively. The complexity to solve fused lasso is $\mathcal{O}(pn\log(n))$ [101] and the complexity to solve linear system of equation $\mathcal{O}(p^3)$. For the data set DS-A, the sample and feature size is quite similar for which JSCCA took 2 sec more time than the proposed method for execution but as the feature size increases, time to solve the linear system of equation get much higher than the time to solve fused lasso. Similarly for CoRe and CoopLe, the only step which is to be executed iteratively is a lasso solver. The time to solve lasso is faster compare to solve a system of linear equation for high dimension. Although these methods take less time to execute, they do not perform well in class-specific feature selection.

From the table, it is seen that the memory consumption of MT-SCCALR is the highest among all the algorithms for the data set DS-A and is not applicable for DS-B and ADNI. The proposed method performs better than MT-SCCALR in this respect. The MT-SCCAR consumes less memory than the propose method because it calculates one canonical weight vector u and v for all the classes, where as the proposed method has to calculate the canonical weight u_c and v_c for each class c, so it require class specific weight updation, which consumes more memory.

6.4 Conclusion

In the field of imaging genetics, identification of the diagnose-specific imaging measures and genetic markers are of great importance in personalized treatments. Most of the models in this domain are based on sparse CCA, and thus unsupervised in nature. Though few supervised models have been developed in the recent years, they cannot identify the diagnose-specific biomarkers. The MT-SCCALR, which can identify the diagnose-specific and diagnose-consistent biomarkers, is not capable of handling large number of features, and is also computationally very expensive. In this regard, a supervised multi-task model, termed as MTL-SDCCA, has been proposed by judiciously integrating the merits of both CCA and LDA. The proposed method can identify the diagnose-specific as well as diagnose-consistent features. It is able to handle thousands of features and less computationally expensive than MT-SCCALR. The proposed model has also included different regularizations to sparsify the solution and to select the important features.

The performance of the proposed method is compared with that of several existing supervised models, namely, CoRe, CoopLe, JSCCA, JCB-SCCA, MT-SCCALR and MT-SCCAR, considering four sets of synthetic data and one real imaging genetic data obtained from ADNI cohort. An important finding is that the proposed method can successfully identify the class-specific and class-consistent features, irrespective of the data sets used. It also has better feature selection capability than several existing methods. For real brain imaging data set, the correlation value for the proposed method is best among all the existing methods compared. It signifies that the selected SNPs and imaging QTs are consistent to give associated genes and ROIs with respect to AD. The results also show that the proposed method has a diverse feature selection ability, can handle large number of features and has significantly lower computational complexity than MT-SCCALR. In future, the proposed model may be extended to incorporate the information of multiple imaging modalities to find the bi-multivariate association with respect to the multi-class problem.

The MTL-SDCCA proposed in this chapter is applicable for two modalities only. In the next chapter, a more general sparse multi-task model is proposed to identify the complex relationships among multimodal imaging genetics data having number of modalities more than 2.

Chapter 7

Multi-View Multi-Task Sparse Canonical Correlation Analysis for Imaging Genetics Study

7.1 Introduction

Imaging genetics, one of the most popular research fields in brain science, primarily focuses on studying neurodegenerative disorders, such as Alzheimer disease, to identify the complex connections between genetic variations and brain functions for the disease under consideration [96]. Due to the advancements in imaging technologies, several neuroimaging modalities, such as magnetic resonance imaging (MRI), functional MRI (fMRI), and positron emission tomography (PET), are now available to capture different quantitative traits (QTs) of the brain, in terms of its structure, function, and metabolism, respectively. In imaging genetics study, these multimodal brain imaging data can provide consensus and complementary information of the underlying disease when integrated with genetic data, such as single nucleotide polymorphism (SNP). This can provide a brief overview of the genetic basis for the abnormal brain regions with respect to the underlying disease. Additionally, a wealth of diagnostic and cognitive assessment data from several angles facilitates the identification of diagnosis-specific biomarkers, which can help to diagnose and prevent the disease a priori [128]. Moreover, diagnosis-specific biomarkers can be used to calculate the polygenic risk score (PRS) to follow the disease progression for personalized treatments [49]. Thus, it is crucial to analyze multiple neuroimaging, genotype, and clinical diagnostic data simultaneously to have a thorough understanding of the disease.

In order to find the complex relationships between high-dimensional multivariate genetic and imaging data, multivariate learning methods explicitly model the data to capture the joint influence of the multiple variables of one data set over the other. The sparse canonical correlation analysis (SCCA) [272,273] is a popular multivariate method in imaging genetics study, which is the sparse variant of canonical correlation analysis (CCA) [107]. It includes a ℓ_1 norm as a constraint to the CCA model and can be applied as a variable selection method to identify the important variables for the genetic and imaging data for which the

data sets are maximally correlated. There exist several SCCA based approaches, which rely on different sparsity-inducing norms, in order to explore the unique structure and prior group information of the imaging genetics data [65, 67, 93, 95, 110, 155]. However, most of the SCCA based approaches do not consider the discriminant information between different diagnostic groups when selecting the features for imaging genetics association. Also, the SCCA models are applicable only for unimodal imaging and genetic data. Since multiple brain imaging modalities can be useful for their complementary and shared information over single brain imaging modality, the integration of multiple imaging modalities in association with the genetic data is of great interest. Some attempts have been made to accommodate more than two modalities by replacing two-view SCCA with its multi-view extension, called multi-view SCCA (mSCCA) [273]. The three-way SCCA was introduced by Hao et al. [93] to examine the connections between SNPs, imaging QTs, and diagnostic status. However, the mSCCA based methods include all the pairwise correlation information among the modalities to find the association. So, it is not possible to uncover plausible genetic loci, unless all the SNPs and imaging modalities are highly correlated with each other, which is too strict to assume.

As mentioned in Chapter 6, the multi-task learning (MTL) is a framework where multiple tasks are learned simultaneously, leveraging the shared information across the tasks, to improve the overall performance [39]. The MTL framework has been adopted for the SCCA models, where multiple SCCA models are learned simultaneously, in order to find the complex imaging genetics association by identifying the task-specific and task-consistent features [128]. The multiple tasks correspond to the selection of features specific for either disease subtypes or multiple modalities explaining different perspective of the disease. The multi-task SCCA (MTSCCA) [66] builds several SCCA tasks concurrently and associates SNPs with imaging QTs of one modality, in order to study the multimodal imaging genetics problem. However, the multiple tasks in MTSCCA focus only on modality-wise feature selection. So, the MTSCCA is unable to perform task-specific and task-consistent feature selection separately, for the multimodal imaging QTs. This limitations can be overcome by the dirty MTSCCA (DMTSCCA) [64], which makes use of parameter decomposition and MTL in order to identify not only the shared imaging QTs and genetic loci across multiple modalities, but also the modality-specific imaging QTs and genetic loci. However, both MTSCCA and DMTSCCA do not use the diagnostic status of the samples, which may give important insight in finding the complex relationships among SNPs and multimodal imaging QTs.

The label-guided MTSCCA (LGMTSCCA) [94] applies parameter decomposition and sparse regression analysis, along with the label information of the data, to obtain modality-consistent and modality-specific weight matrices. The label information is used in sparse regression analysis to retain the relevant and noise-free features. In [265], a multimodality discriminant SCCA (MD-SCCA) algorithm has been introduced, which utilizes the discriminant information between different diagnostic groups, and explores the relationships among multiple brain imaging modalities, to identify the complex genotype-phenotype associations. In [128], multi-task SCCA and regression (MT-SCCAR) has been introduced, which uses genetic and multimodal neuroimaging data, along with cognitive measures, to identify the high-risk brain regions and genetic risk factors associated with Alzheimer disease. The joint connectivity based SCCA (JCB-SCCA) [132] introduces a connectivity penalty to incorporate the prior biological connectivity measures such as brain pathway

and joint association of genetic markers. The joint sparse canonical correlation analysis (JSCCA) [74] uses the label information for jointly estimating the multiple SCCA models, to obtain canonical vectors with common and class-specific patterns using a generalized fused lasso penalty. In [63], Du et al. proposed a multi-task bi-multivariate approach, termed as multi-task SCCA and logistic regression (MT-SCCALR), which selects diagnose-specific features to find genotype-phenotype patterns particular to each diagnosis. It uses MTSCCA and multi-task logistic regression simultaneously to select a set of relevant features for each diagnostic group. In order to obtain common and unique characteristics from various disease classes, Song et al. [229] proposed a joint sparse collaborative regression (JSCoReg) model, based on sparse regression and CCA.

However, the MTL based SCCA models described above focus either on the diagnose-specific or modality-specific feature selection. Most of them, such as MTSCCA, DMTSCCA, LGMTSCCA, MD-SCCA, MT-SCCALR, MT-SCCAR, and JSCoReg, are also computationally very expensive since all these methods involve computing inverse of a very high-dimensional covariance matrix of SNP or imaging data, or require to solve a large system of linear equations. The complexity of these methods is thus $\mathcal{O}(p^3)$, where p is the number of features of SNP or imaging data. Moreover, most of them ignore the available information of cognitive assessment or clinical data of the patients, which may help in identifying the relevant features and making the model robust to noise.

In this regard, a multi-view multi-task sparse canonical correlation analysis (MvMt-SCCA) is proposed in this chapter, as illustrated in Fig. 7.1, which utilizes multiple imaging modalities, genetic data, cognitive assessment and label information simultaneously, in order to identify the complex relationships between genetic and multimodal imaging data. While the multiple brain imaging modalities provide the complementary and shared information of different imaging QTs that are intrinsically connected to the genetic variations, the cognitive assessment data, along with the label information, helps to select the features those are relevant and noise-free with respect to disease-subtypes. In effect, the proposed method has the ability to find modality-wise diagnose-specific and diagnose-consistent features, which help in finding the SNPs-QTs association with respect to the disease-subtypes. The proposed method utilizes both lasso and fused lasso penalties, which together guide the modality-wise, class-specific and class-consistent variable selection. An iterative algorithm, using block coordinate descent, is introduced to solve the optimization problem of MvMt-SCCCA. It reduces the complexity of the proposed method to $\mathcal{O}(p^2 \log p)$, and makes the model applicable for large-scale imaging genetics study compared to the state-of-the-art methods. The performance of the proposed method is studied on both simulated and real-world data sets. The results on the simulated data signify that the proposed method obtains a higher correlation value, has a better feature selection capability with respect to each diagnostic group, and is robust to noise. The study using real ADNI data finds a group of risk genes and modality-wise abnormal brain regions corresponding to each diagnosis group, which may contribute to a more comprehensive understanding of neurodegenerative disease and may help in personalized treatment. Some of the results of this chapter can be found in [178].

The remaining sections of this chapter are as follows: Section 7.2 presents a brief overview of the sparse canonical correlation analysis, based on which the proposed methodology is built in Section 7.3. Computational complexity analysis of the proposed MvMt-SCCA is also given in this section. The efficacy of the proposed method, in terms of identi-

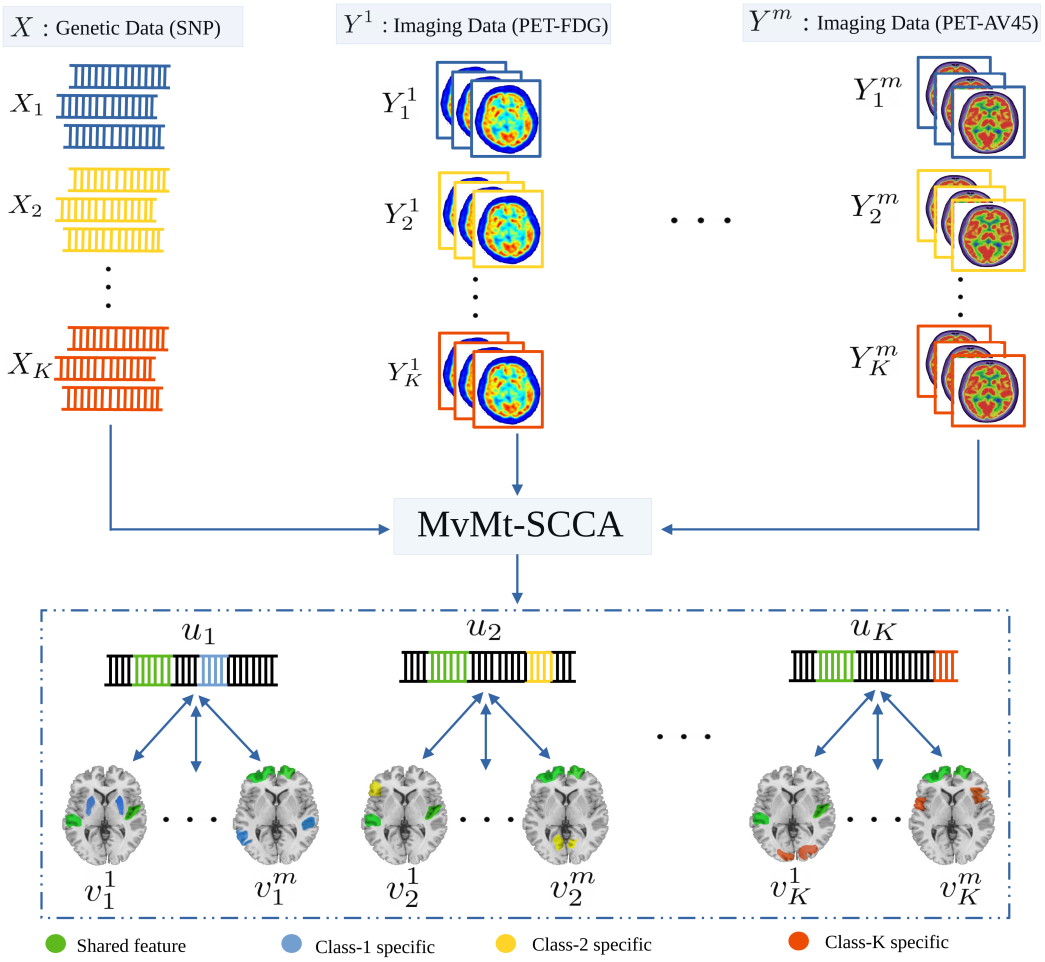


Figure 7.1: An illustration of the MvMt-SCCA method for the identification of diagnose-specific imaging genetic association with respect to multimodal imaging genetics data.

fying important biomarkers for multi-modal imaging genetics study, is shown by comparing the performance of it with the other state-of-the-art algorithms over a simulated and a real neuroimaging data in Section 7.4 and Section 7.5, respectively. The chapter is concluded in Section 6.4.

7.2 Sparse Canonical Correlation Analysis

The SCCA model [272, 273] is the sparse variant of the CCA model. Suppose $X \in \mathbb{R}^{n \times p}$ and $Y \in \mathbb{R}^{n \times q}$ are two data sets, where n denotes the number of common samples, and p and q denote the number of features in X and Y, respectively. The SCCA aims to find the association between X and Y by obtaining two sparse canonical vectors $u^{p \times 1}$ and $v^{q \times 1}$.

The objective function of the SCCA is given by

$$\begin{aligned} \min_{u,v} \quad & -u^T X^T Y v + \lambda_u \|u\|_1 + \lambda_v \|v\|_1 \\ \text{subject to} \quad & \|u\|_2 \leqslant 1, \|v\|_2 \leqslant 1, \end{aligned} \tag{7.1}$$

where λ_u and λ_v are the regularization parameters of sparsity. The ℓ_1 penalty is used to make the model sparse and to prevent the overfitting problem for $p, q >> n$. In the above optimization problem, the data variance matrices are taken as the identity matrices, that is, $X^T X = Y^T Y = I_n$ [272], where I_n is the identity matrix of order n.

The JSCCA model attempts to fit multiple SCCA models jointly for multiple tasks [74]. Assume that the data X and Y can be divided into K classes of normalized data $X_k \in \mathbb{R}^{n_k \times p}$ and $Y_k \in \mathbb{R}^{n_k \times q}$ according to the available information of K classes, where n_k is the number of samples in the k-th class. The JSCCA estimates one shared canonical vector u for X and K related sparse canonical vectors $\{v_k\}_{k=1}^K$ for Y, respectively, by jointly applying multiple SCCA models for multiple classes of normalized data X_k and Y_k. The objective of the JSCCA is defined as follows:

$$\begin{aligned} \min_{u,V} \quad & -\sum_{k=1}^{K} \frac{1}{n_k} u^T X_k^T Y_k v_k + \lambda_u \|u\|_1 + \lambda_v \sum_{k=1}^{K} \|v_k\|_1 + \tau \sum_{k<k'} \|v_k - v_{k'}\|_1 \\ \text{subject to} \quad & \|u\|_2^2 = \|V\|_F^2 = 1, \end{aligned} \tag{7.2}$$

where $V = [v_1, v_2, \ldots, v_K]$, and τ is the regularization parameter. The general fused lasso penalty [52] is applied to V to make the K canonical vectors sparse and to find similar structures across different classes. However, the JSCCA is limited to apply for two modalities only. Moreover, it obtains task specific canonical vectors for the modality Y only, not for the modality X.

7.3 Proposed Method

In this section, a multi-view multi-task sparse CCA (MvMt-SCCA) model is introduced for imaging genetics study. Let $X \in \mathbb{R}^{n \times p}$ be the SNP data set, where p denotes the number of SNPs, and $Y^m \in \mathbb{R}^{n \times q_m}$ represents the neuroimaging data of the m-th modality, where q_m denotes the number of neuroimaging features present in the m-th modality $(m = 1, 2, \ldots M)$. Also, assume that the cognitive assessment data, $Z \in \mathbb{R}^{n \times r}$, with r number of features is available for the n samples. Let the samples be classified into K number of diagnosis groups. Consider dividing K classes of normalized data, $X_k \in \mathbb{R}^{n_k \times p}$ for the SNP data X, $Y_k^m \in \mathbb{R}^{n_k \times q_m}$ for the m-th neuroimaging modality Y^m, and $Z_k \in \mathbb{R}^{n_k \times r}$ for the cognitive assessment data Z, where n_k is the number of samples present in the k-th diagnosis group. The proposed model overcomes the limitations of JSCCA by incorporating the information of prior cognitive measures into the JSCCA model and selecting the class-wise features not only for neuroimaging modalities but also for genetic modality. The objective is to find diagnose-specific correspondence between the genetic data and multiple neuroimaging modalities by jointly fitting multiple SCCA models. The proposed

formulation is as follows:

$$\min_{W,U,V^m} -\sum_{m=1}^{M}\sum_{k=1}^{K}\frac{1}{n_k}(u_k^T X_k^T + w_k^T Z_k^T)Y_k^m v_k^m + \lambda_1\sum_{k=1}^{K}\|w_k\|_1 + \lambda_2\sum_{k=1}^{K}\|u_k\|_1$$
$$+\tau_1\sum_{k<k'}\|u_k - u_{k'}\|_1 + \sum_{m=1}^{M}\lambda_3^m\sum_{k=1}^{K}\|v_k^m\|_1 + \sum_{m=1}^{M}\tau_2^m\sum_{k<k'}\|v_k^m - v_{k'}^m\|_1$$
$$\text{subject to } \|W\|_F^2 = 1, \|U\|_F^2 = 1, \sum_{m=1}^{M}\|V^m\|_F^2 = 1, \tag{7.3}$$

where $U = [u_1, u_2, \ldots, u_K]$, $V^m = [v_1^m, v_2^m, \ldots, v_K^m]$ for $m = 1, 2, \ldots M$, and $W = [w_1, w_2, \ldots, w_K]$ are the canonical weight vectors for X_k, $\{Y_k^m\}_{m=1}^M$ and Z_k, respectively, λ_1, λ_2, $\{\lambda_3^m\}_{m=1}^M$, τ_1, and $\{\tau_2^m\}_{m=1}^M$ are the regularization parameters. The relevance of each term in the proposed formulation is explained next.

The first term of the objective function in (7.3) consists of two parts. The first part takes into account the correlation between the SNP and neuroimaging data, whereas the second part is considered to correlate the cognitive measures with the neuroimaging data. Since the cognitive outcome of patients is directly affected by the abnormal brain regions due to the degeneration of the neurons, the correlation between cognitive assessment and neuroimaging data is only considered. Moreover, the cognitive measures with the label information can influence relevant and noise-free feature selection for each diagnosis group for the multiple imaging modalities, which further helps in the association of imaging and genetic data. The second term of the objective function is taken as the $\ell_{1,1}$ (ℓ_1-norm for matrix) penalty of W, which is used to find the diagnose-specific sparse cognitive measures that implicitly highlight the diagnose-specific features in imaging modalities. The parameter λ_1 controls the sparsity of the canonical weight W.

In the third and fourth terms, $\ell_{1,1}$ penalty and the generalized fused lasso penalty are applied to the canonical weight U, to sparsify the weight vectors u_k, for the class-specific and class-consistent feature selection for the SNP data. The $\ell_{1,1}$ penalty on U is controlled by the parameter λ_2. It ensures the sparsity on each u_k. The $\ell_{1,1}$ penalty, applied to the difference of every two canonical weight vectors of U, encourages multiple features of the canonical weight vectors to be fused with each other. It helps the canonical vectors to share a similar structure over different diagnostic groups of the SNP data. The fusion level between two canonical vectors u_k and $u_{k'}$ is controlled by τ_1. The fifth and sixth terms of the objective function are for the class-specific and class-consistence feature selection of multiple neuroimaging modalities. The $\ell_{1,1}$ penalty on each V^m encourages the sparsity in each canonical vector v_k^m for the k-th class in the m-th imaging modality, which is controlled by the parameter λ_3^m. The fused lasso penalty on v_k^m and $v_{k'}^m$ selects variables that share the same information across different diagnosis groups for each imaging modality. The fusion level between two imaging canonical vectors v_k^m and $v_{k'}^m$ is controlled by the parameter τ_2^m for the m-th imaging modality.

The constraint $\|W\|_F^2 = 1$ explains the joint estimation of the canonical weights w_k across different classes, while $\|U\|_F^2 = 1$ signifies the joint estimation of different canonical vectors of the genetic data. Instead of taking $\|V^m\|_F^2 = 1$, for $m = 1, 2, \ldots, M$ as the constraint, $\sum_{m=1}^M \|V^m\|_F^2 = 1$ is taken for estimating the imaging canonical vectors jointly

across different classes and cooperatively across all the imaging modalities.

7.3.1 Optimization of MvMt-SCCA

This section presents an alternating optimization technique for the minimization problem of (7.3), in terms of the variables W, U, and $\{V^m\}_{m=1}^M$, since the objective function cannot be optimized directly. In this scenario, the objective function of (7.3) is alternatively convex with respect to each of the three variables, while the remaining variables are treated as fixed. In each case, the block coordinate descent algorithm is applied to solve the problem.

7.3.1.1 Canonical Weight of Cognitive Measure

In order to update W, while U and $\{V^m\}_{m=1}^M$ are fixed, the optimization problem in (7.3) can be expressed as follows:

$$\min_{\|W\|_F^2=1} \; -\sum_{m=1}^{M}\sum_{k=1}^{K}\frac{1}{n_k} w_k^T Z_k^T Y_k^m v_k^m + \lambda_1 \sum_{k=1}^{K}\|w_k\|_1. \tag{7.4}$$

The solution to the problem of (7.4) can be obtained by solving the K sub-problems defined as follows:

$$\min_{\|w_k\|\leqslant 1} \; -a_k^T w_k + \lambda_1 \|w_k\|_1, \tag{7.5}$$

$$\text{where} \;\; a_k = \sum_{m=1}^{M}\frac{1}{n_k} Z_k^T Y_k^m v_k^m; \;\; k = 1, 2, \ldots, K. \tag{7.6}$$

It contains the information of the sum of cross-correlation among the cognitive measures and multiple imaging modalities corresponding to each diagnosis group. The sparse solution w_k for the k-th group is in the direction where the sum of cross-correlation for the group is maximum. The solution can be obtained by the algorithm described in [273] and is given by the equation

$$\hat{w}_k = \frac{S(a_k, \lambda_1)}{\|S(a_k, \lambda_1)\|}. \tag{7.7}$$

The operator $S(\cdot)$ is soft-thresholding operator, which is defined as follows:

$$S(x, \lambda) = \text{sgn}(x)\max(|x| - \lambda, 0).$$

The solution of the problem of (7.4) is given by $W = \frac{\hat{W}}{\|\hat{W}\|}$, where $\hat{W} = [\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_k]$.

7.3.1.2 Canonical Weight of SNP Data

To update the canonical weight matrix U, while considering the other variables W and $\{V^m\}_{m=1}^M$ as constant, the optimization problem of (7.3) can be written as

$$\min_{\|U\|_F^2=1} -\sum_{m=1}^{M}\sum_{k=1}^{K}\frac{1}{n_k}u_k^T X_k^T Y_k^m v_k^m + \lambda_2\sum_{k=1}^{K}\|u_k\|_1 + \tau_1\sum_{k<k'}\|u_k - u_{k'}\|_1. \tag{7.8}$$

The problem of (7.8) can be expressed as follows:

$$\min_{\|U\|_F^2=1} -\sum_{k=1}^{K} b_k^T u_k + \lambda_2\sum_{k=1}^{K}\|u_k\|_1 + \tau_1\sum_{k<k'}\|u_k - u_{k'}\|_1,$$

which is same as solving the following problem:

$$\min_{\|U\|_F^2=1}\sum_{k=1}^{K}\|b_k - u_k\|_2^2 + \lambda_2\sum_{k=1}^{K}\|u_k\|_1 + \tau_1\sum_{k<k'}\|u_k - u_{k'}\|_1,$$
$$\text{where}\quad b_k = \sum_{m=1}^{M}\frac{1}{n_k}X_k^T Y_k^m v_k^m. \tag{7.9}$$

It includes the sum of cross-correlation among the SNP data and multiple imaging modalities corresponding to each diagnosis group. The solution u_k for the k-th group is in the direction of maximum sum of cross-correlation for that group. The problem of (7.9) can be viewed as a special case of the fused lasso signal approximator (FLSA) problem [102]. A very efficient solution to this problem is found in [52, 101]. Following three steps are required to solve the problem of (7.9):

- **Fusion:** In this step, a solution $\bar{U}$ is obtained by setting $\lambda_2 = 0$, which is necessary to fuse the features in u_k and $u_{k'}$ that do not have significant differences measured by the parameter τ_1. The optimization problem for this step is given by:

$$\bar{U} = \arg\min_{U}\sum_{k=1}^{K}\|b_k - u_k\|_2^2 + \tau_1\sum_{k<k'}\|u_k - u_{k'}\|_1. \tag{7.10}$$

  The solution to this problem can efficiently be obtained by the algorithm of FLSA [101].

- **Sparsification:** In this step, a soft-thresholding operator is applied to $\bar{U}$ to obtain the optimal solution, $\hat{U} = [\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_K]$, from the relation $\hat{u}_k = S(\bar{u}_k, \lambda_2)$.

- **Normalization:** In the final step, $\hat{U}$ is normalized by $\frac{\hat{U}}{\|\hat{U}\|}$, which gives the optimal solution for the variable U.

7.3.1.3 Canonical Weight of Neuroimaging Data

For updating the variables $\{V^m\}_{m=1}^M$, considering W and U as fixed, the objective function of (7.3) becomes

$$\min_{\{V^m\}_{m=1}^M} -\sum_{m=1}^{M}\sum_{k=1}^{K}\frac{1}{n_k}(u_k^T X_k^T + w_k^T Z_k^T)Y_k^m v_k^m + \sum_{m=1}^{M}\lambda_3^m\sum_{k=1}^{K}\|v_k^m\|_1 + \sum_{m=1}^{M}\tau_2^m\sum_{k<k'}\|v_k^m - v_{k'}^m\|_1$$
$$\text{subject to } \sum_{m=1}^{M}\|V^m\|_F^2 = 1. \tag{7.11}$$

The minimization problem in (7.11) can be broken down to M sub-problems given by

$$\min_{\|V^m\|\leqslant 1} -\sum_{k=1}^{K}\frac{1}{n_k}(u_k^T X_k^T + w_k^T Z_k^T)Y_k^m v_k^m + \lambda_3^m\sum_{k=1}^{K}\|v_k^m\|_1 + \tau_2^m\sum_{k<k'}\|v_k^m - v_{k'}^m\|_1. \tag{7.12}$$

The problem of (7.12) can also be written in the form of (7.9) as follows:

$$\min_{\|V^m\|\leqslant 1}\sum_{k=1}^{K}\|c_k^m - v_k^m\|_2^2 + \lambda_3^m\sum_{k=1}^{K}\|v_k^m\|_1 + \tau_2^m\sum_{k<k'}\|v_k^m - v_{k'}^m\|_1,$$
$$\text{where } \; c_k^m = \frac{1}{n_k}Y_k^{mT}(X_k u_k + Z_k w_k). \tag{7.13}$$

The above problem can be solved in the same three-step procedure mentioned above. Here, the optimization problem in the fusion step is

$$\bar{V}^m = \arg\min_{V^m}\sum_{k=1}^{K}\|c_k^m - v_k^m\|_2^2 + \tau_2^m\sum_{k<k'}\|v_k^m - v_{k'}^m\|_1, \tag{7.14}$$

where $\bar{V}^m$ is the solution of (7.12) obtained by setting $\lambda_3^m = 0$. All the successive steps are similar as before and the final solution of the problem of (7.11) can be obtained by normalizing each canonical loading, $\hat{V}^m$, by the sum $\sum_{m=1}^{M}\|\hat{V}^m\|_F^2$, which scales the solutions across all the imaging modalities. The main steps to solve the proposed optimization problem are summarized in Algorithm 7.1.

7.3.2 Computational Complexity Analysis

This section briefly analyzes the computational complexity of the proposed MvMt-SCCA algorithm. For convenience, assume that $q = \max_m\{q_m\}$, where q_m is the dimension of the m-th imaging modality Y^m. The computational cost for executing each step of Algorithm 7.1 is outlined below.

1. The computational cost for the initialization step is $\mathcal{O}(K(p+Mq+r)) \approx \mathcal{O}(p+Mq)$.

2. The computational cost to compute a_k in step 4 is $\mathcal{O}(rq)$ and $\hat{w}_k$ is $\mathcal{O}(r)$. Since the steps 4 and 5 are executed K times in a loop, the total complexity of step 3 to step

Algorithm 7.1 MvMt-SCCA

Input: Normalized data X_k, Y_k and Z_k. The parameters λ_1, λ_2, λ_3^m, τ_1 and τ_2^m; $m = 1, 2, \ldots, M$; $k = 1, 2, \ldots, K$. **Output:** The canonical loadings, $W = [w_1, w_2, \ldots, w_K]$, $U = [u_1, u_2, \ldots, u_K]$ and $V^m = [v_1^m, v_2^m, \ldots, v_K^m]$ for $m = 1, 2, \ldots, M$.

1: Initialize the loading matrix W, U and $\{V^m\}_{m=1}^M$.
2: **while** not converge **do**
3: **for** $k = 1$ to K **do**
4: Compute a_k according to (7.6).
5: Compute $\hat{w}_k$ according to (7.7).
6: **end for**
7: Normalize: $W = \frac{\hat{W}}{\|\hat{W}\|}$.
8: Find $\bar{U}$ according to (7.10).
9: Obtain $\hat{u}_k = S(\bar{u}_k, \lambda_2)$; $k = 1, 2, \ldots, K$.
10: Normalize: $U = \frac{\hat{U}}{\|\hat{U}\|}$.
11: **for** $m = 1$ to M **do**
12: Find $\bar{V}^m$ according to (7.14).
13: Obtain $\hat{v}_k^m = S(\bar{v}_k^m, \lambda_2)$; $k = 1, 2, \ldots, K$.
14: Normalize: $V^m = \frac{\hat{V}^m}{\|\hat{V}^m\|}$.
15: **end for**
16: Normalize: $V^m = \frac{\hat{V}^m}{\sum_{m=1}^M \|\hat{V}^m\|_F^2}$.
17: **end while**

6 is $\mathcal{O}(K(rq + r)) \approx \mathcal{O}(rq)$.

3. The normalization in step 7 requires $\mathcal{O}(Kr)$ time.

4. In step 9, the cost for solving the FLSA for the canonical vector U is at most $\mathcal{O}(p^2 \log p)$ [101]. The steps of soft-thresholding and normalization require $\mathcal{O}(Kp)$ computational cost each. Therefore, total time cost for the computation of U is $\mathcal{O}(Kp + p^2 \log p) \approx \mathcal{O}(p^2 \log p)$.

5. Similarly, the total computational cost for the steps 11 to 15 is $\mathcal{O}(M(Kq + q^2 \log q))$ $\approx \mathcal{O}(q^2 \log q)$.

6. The computational cost for the step 16 is $\mathcal{O}(MqK)$.

Therefore, the total computational complexity of the proposed algorithm is $\mathcal{O}(p + Mq + \tau(rq + Kr + p^2 \log p + q^2 \log q + MqK))$, where τ is the number of iterations required to converge the iterative procedure. Since the number of classes K, cognitive measures r, and number of imaging modalities M are very very less than p or q, the complexity can be reduced to $\mathcal{O}(\tau(p^2 \log p + q^2 \log q))$. So, the complexity is lesser than the existing MTL based SCCA methods, which require $\mathcal{O}(\tau(p^3 + q^3))$ time to solve the optimization problem. This makes the proposed algorithm efficient for large-scale imaging genetics problem.

7.4 Simulation Study

In this section, the effectiveness of the proposed method in identifying imaging genetics correlation is evaluated through a set of experiments on simulated data.

7.4.1 SIG: A Simulated Data

A simulated imaging genetics data, termed as SIG, is generated to conduct a set of experiments. Let $X \in \mathbb{R}^{n\times p}$ be the SNP data, and $Y^1 \in \mathbb{R}^{n\times q}$ and $Y^2 \in \mathbb{R}^{n\times q}$ denote two neuroimaging modalities, namely, IM1 and IM2, respectively. The SIG data contains $n = 300$ samples belonging to three diagnostic groups, where each of the three groups or classes consists of 100 samples each. The data set is generated in order to reflect the association between SNPs-QTs through both the class-consistent and class-specific features.

To reflect similar effect between the genetic and multiple imaging modalities, a latent variable model is used similar to [272]. The latent variable, $l_k \in \mathbb{R}^{n_k\times 1}$ $(k = 1, 2, 3)$, is generated by a normal distribution $\mathcal{N}(0, 1)$. Next, one sparse canonical weight matrix $U = [u_1, u_2, u_3] \in \mathbb{R}^{p\times 3}$ for the SNP data and two sparse canonical weight matrices $V^m = [v_1^m, v_2^m, v_3^m] \in \mathbb{R}^{q\times 3}$ $(m = 1, 2)$ for the two imaging modalities are generated. Using the latent variable l_k and the canonical weight matrices U, V^1, and V^2, $X_k^i \sim \mathcal{N}(l_k^i u_k, \sigma_x I_{p\times p})$, $Y_k^{1,i} \sim \mathcal{N}(l_k^i v_k, \sigma_{y_1} I_{q\times q})$, and $Y_k^{2,i} \sim \mathcal{N}(l_k^i u_k, \sigma_{y_2} I_{q\times q})$ are constructed for each i-th sample belonging to the k-th class $(k = 1, 2, 3)$. The noise strength is considered to be $\sigma_x = \sigma_{y_1} = \sigma_{y_2} = 0.5$. Finally, the SNP data X and two imaging modalities IM1 (Y^1) and IM2 (Y^2) are constructed from X_k^i, $Y_k^{1,i}$ and $Y_k^{2,i}$, respectively.

Following [132], the SNP data X is further recoded into $\{0, 1, 2\}$ in order to reflect the intrinsic property of the SNP as its minor allele count. A binomial distribution $B(2, \text{logit}^{-1}(x + \text{logit}(\kappa)))$ is used to convert each element of the SNP data X into $\{0, 1, 2\}$, where $\text{logit}(\kappa) = \log(\frac{\kappa}{1-\kappa})$ and κ is the minor allele frequency (MAF) of the SNP, generated from the uniform distribution $\mathcal{U}(0.01, 0.1)$ to include rare variant of SNPs. The cognitive assessment data Z is generated by the relation $Z = lw^T + e$, where $l = [l_1, l_2, l_3]^T$, $w \sim \mathcal{N}(0, 1)$ consisting of ten cognitive outcomes corresponding to each sample, and e is the Gaussian noise [128]. Using the above procedure, the SIG data is generated, which consists of SNP (X), two imaging modalities IM1 (Y_1) and IM2 (Y_2), and cognitive assessment score Z, with $p = 10,000$ and $q = 5000$.

7.4.2 Selection of Parameters

The proposed method has $(2M + 3)$ number of parameters, namely, $\lambda_1, \lambda_2, \{\lambda_3^m\}_{m=1}^M, \tau_1$, and $\{\tau_2^m\}_{m=1}^M$, to be tuned. A five-fold cross-validation (CV) strategy has been used to find the optimal set of parameters. The parameters for which the mean canonical correlation coefficient (CCC) is highest in the training set are considered to be optimal parameters. In the current experimental setup, the mean CCC $(\bar{\rho})$ over the five-fold CV is computed as follows:

$$\bar{\rho} = \frac{1}{5M} \sum_{f=1}^{5} \sum_{m=1}^{M} \sum_{k=1}^{K} \rho(X u_k, Y^m v_k^m), \tag{7.15}$$

where ρ is the correlation between genetic and imaging canonical vectors in the projected space. The sparsity parameters λ_1, λ_2, and λ_3 are chosen from the set $\{0.001, 0.01, 0.1, 1.0\}$, while the fusion parameters τ_1 and τ_2 are chosen from the set {0.001, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}.

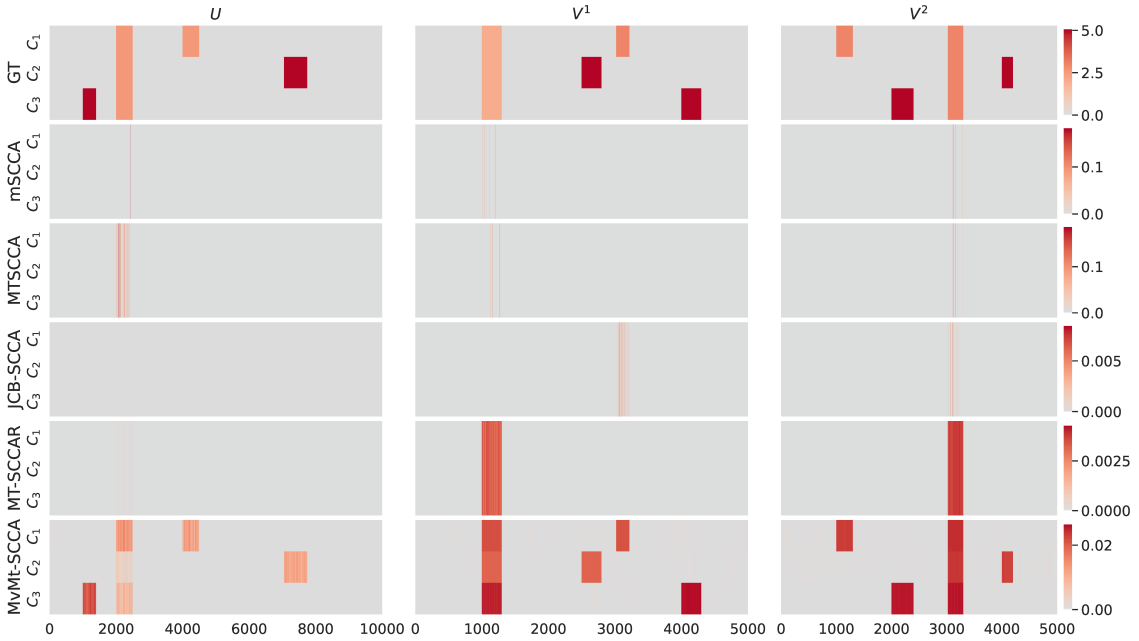


Figure 7.2: Heatmaps showing canonical weights U, V^1 and V^2 for SIG data: Rows 1 to 6 depict the results corresponding to the ground truth (GT), mSCCA, MTSCCA, JCB-SCCA, MT-SCCAR, and the proposed method, while the columns 1, 2 and 3 represent the canonical weights U, V^1 and V^2, respectively, corresponding to SNP, IM1 and IM2. Each row contains three task (C_1, C_2, C_3) specific features. The features are shown by taking the mean computed over five-fold cross-validation.

7.4.3 Experimental Results and Discussion

This section presents the experimental results on SIG data set. The detailed description of generating the canonical weight matrices, including the class-consistent and class-specific features, is given in the supplementary material. The performance of the proposed method is evaluated with respect to both diagnose-specific and modality-specific feature selection ability in identifying the SNPs-QTs association.

7.4.3.1 Modality Specific Feature Selection

The performance of the proposed MvMt-SCCA is compared with that of several multi-view modality-specific feature selection methods, namely, mSCCA [273], MTSCCA [66], JCB-SCCA [132], and MT-SCCAR [128], and the corresponding results on SIG data are reported in Fig. 7.2. In Fig. 7.2, the canonical weights U, V^1 and V^2 are represented by using the heatmaps to highlight the important features. Since the methods mSCCA,

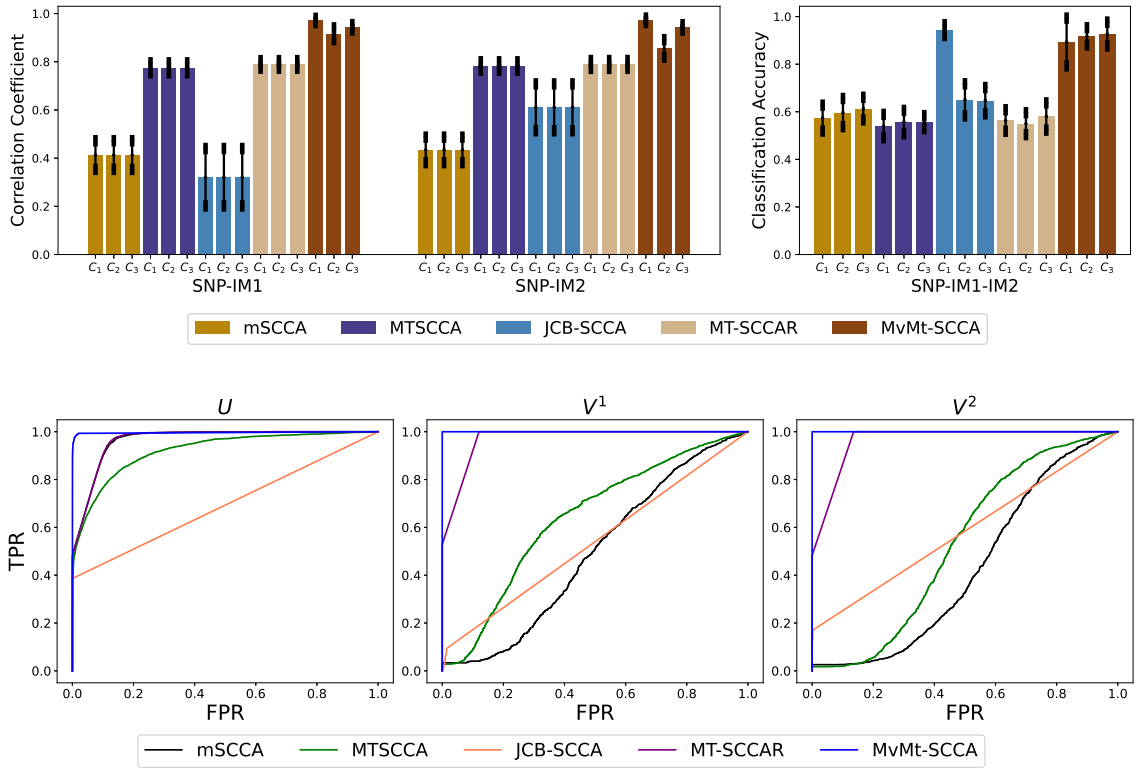


Figure 7.3: Comparison of different methods: Top row presents canonical correlation coefficient and classification accuracy, while columns 1, 2 and 3 of bottom row depict the ROC curve on detection of canonical weights U, V^1 and V^2, respectively.

MTSCCA, JCB-SCCA, and MT-SCCAR generate one canonical vector for each modality, the canonical vectors are stacked $K = 3$ times to get a heatmap comparable to the proposed method. From Fig. 7.2, it can be concluded that each of the extracted canonical vectors by the proposed method has the highest degree of similarity with the ground truth (GT) for the SIG data. All the other methods have also identified important features, more or less consistent with the GT vectors, but they are not class-specific.

The first row of Fig. 7.3 presents the mean correlation coefficient and classification accuracy on test set as indicators of the relationship between imaging QTs and SNPs. The support vector machine (SVM) is used to compute the classification accuracy of each method. By concatenating the top ten (10) selected features of the genetic and two imaging modalities, the classification accuracy is achieved. The bars of Fig. 7.3 corresponding to the correlation coefficient and classification accuracy signify that the proposed method attains the highest correlation and comparatively better classification accuracy among all the methods for the SIG data set. The proposed method indeed has the highest classification accuracy since all the existing multi-view methods are not able to obtain the class-specific features that help to classify the samples. Finally, the performance of the proposed method is compared with that of the existing algorithms in bottom row of Fig. 7.3 with respect to the receiver operating characteristic (ROC) curve by varying the cutoff of the canonical

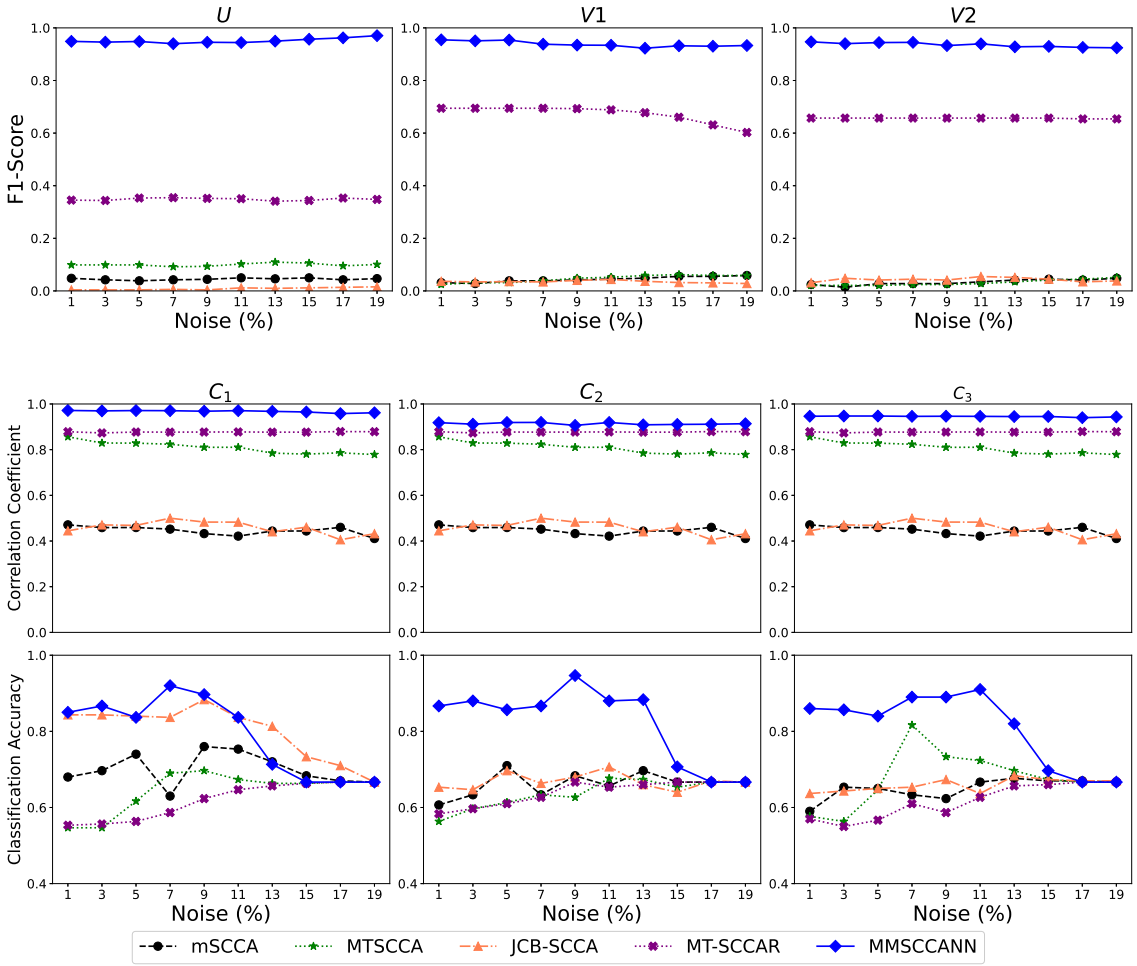


Figure 7.4: Comparative performance analysis of different methods under different noise levels, with respect to F1 score, correlation coefficient and classification accuracy.

loadings. The figure shows that the proposed method has an almost perfect curve of true positive rate (TPR) versus false positive rate (FPR) for all the canonical weights on SIG data. Both MT-SCCA and MT-SCCAR have poor performance in terms of TPR and FPR values in case of large data set like SIG. All the findings show that the proposed method outperforms all other methods, in terms of selected features that are specific to a given class, correlation value, and classification accuracy. The results on SIG data also indicate that the proposed method is capable of handling large-scale imaging genetics data.

7.4.3.2 Noise Sensitivity Analysis

To analyze the noise sensitivity of the proposed approach in comparison to the existing methods, random Gaussian noise is added to the standardized SIG data set as follows:

$$D' = D + \sigma_e E, \tag{7.16}$$

Table 7.1: Comparison of Execution Time and Space

Algorithms	Time (Sec.)	Space (MB)
JSCCA [74]	0.03	150
MT-SCCALR [63]	653.6	1600
MTL-SDCCA	8.42	500
MvMt-SCCA	0.30	**150**

where D' is the noisy data set, D is the original noise free SIG data set, E is the Gaussian noise generated from the normal distribution $\mathcal{N}(0_{d\times 1}, I_{d\times d})$ and σ_e is the noise strength. By varying σ_e from 1% to 19% with the increment of 2%, the performance of various approaches is evaluated, in terms of F1-scores, mean correlation coefficient, and mean classification accuracy. The first row of Fig. 7.4 shows the variation of the F1-score with respect to different noise levels. From the results, it can be seen that the performance of the proposed method with respect to F1-score reduces very slowly with the increase in noise levels, whereas that of different methods deteriorates significantly. All the existing methods, except MT-SCCAR, attain very low F1-score, irrespective of the noise levels. On the other hand, the proposed approach outperforms all other approaches in terms of F1-score, regardless of noise levels and modalities.

The second and third rows of Fig. 7.4 illustrate the variation of the correlation coefficient and classification accuracy, respectively, with the increase in noise levels. The correlation coefficient is computed class-wise, but averaged over all the modalities. All the results show that the correlation coefficient of the proposed method decreases slightly as the noise level increases, whereas the classification accuracy changes a lot with respect to noise after 10% noise level. However, the proposed method performs better than all other existing methods with respect to both correlation coefficient and classification accuracy, irrespective of the noise levels. Other than the proposed method, both MTSCCA and JCB-SCCA provide higher correlation value, but they do not have significance for class-specific correlations. Also, the classification accuracy of the existing methods is very poor as compared to that of the proposed method over different noise levels.

7.4.3.3 Diagnose Specific Feature Selection

In this section, the performance of the proposed method is compared with that of JSCCA [74], MTL-SDCCA, proposed in Chapter 6, and MT-SCCALR [63], which are bi-multivariate diagnose-specific feature selection methods. So, the analysis is presented for both (X, Y^1) and (X, Y^2) pairs separately. As the the algorithm, MT-SCCALR, is not scalable to high-dimensional data, another data set is simulated, which has the similar structure of having class-consistent and class-specific features like SIG data, but significantly lesser number of features and samples. In this case, $n = 150$, $p = 500$ and $q = 600$ are considered.

To compare the performance of the proposed method with that of these three methods, extensive experimentation is conducted. Fig. 7.5(a) and Fig. 7.5(b) present the heatmaps for the pairs of canonical matrices (U, V^1) and (U, V^2), respectively, while the mean testing canonical correlation coefficient and mean testing classification accuracy are presented by the bar graphs in Fig. 7.5(c) and Fig. 7.5(d), respectively. It is seen from the results that

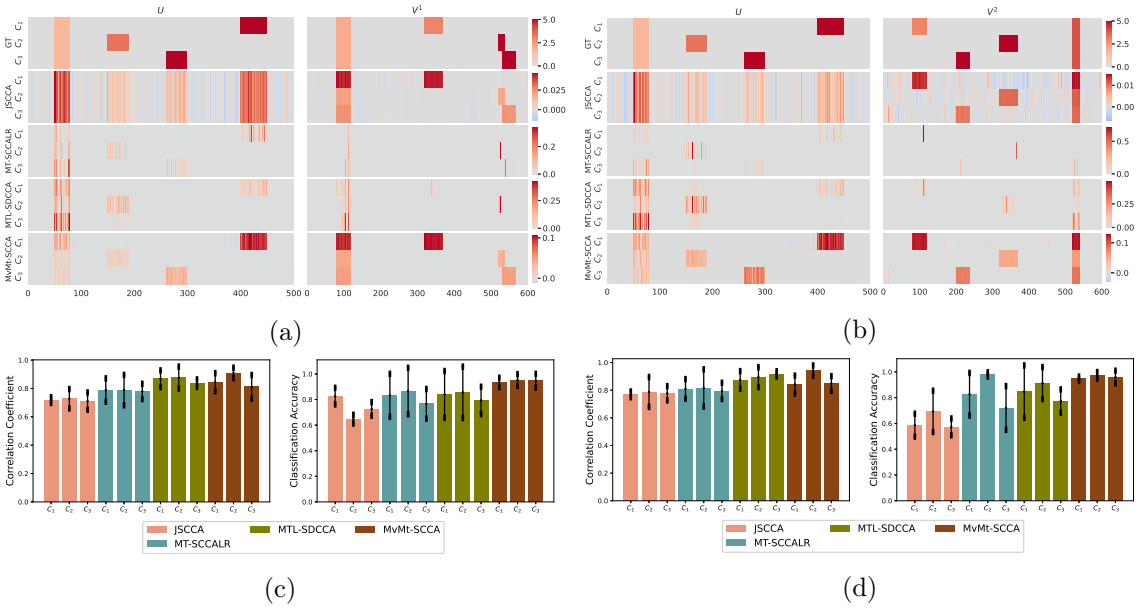


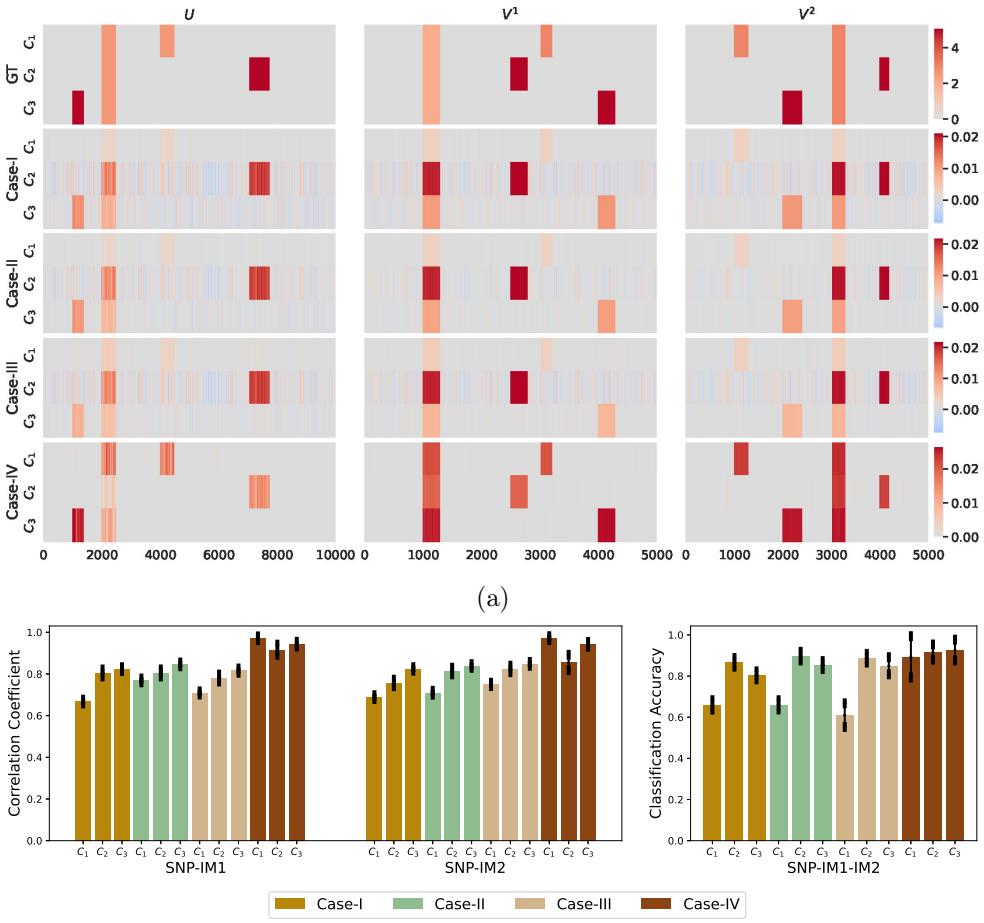
Figure 7.5: Comparative performance analysis of different methods with respect to feature selection, canonical correlation coefficient and classification accuracy. Top row: Heatmaps of canonical weights signifying feature selection corresponding to the pairs (a) (SNP, IM1) and (b) (SNP, IM2). Bottom row: Correlation coefficient and classification accuracy obtained by the selected features for the pairs (c) (SNP, IM1) and (d) (SNP, IM2).

for both pairs of canonical weights (U, V^1) and (U, V^2), the proposed MvMt-SCCA method has the best feature selection capability. It has also captured the highest correlation value and classification accuracy, irrespective of the tasks, with lower variance over the five folds, compared to MT-SCCALR and JSCCA. On the other hand, MTL-SDCCA have the second best correlation coefficient and classification accuracy on both pairs. The MT-SCCALR performs well in both cases, whereas JSCCA has only identified class-specific features for the imaging modalities. The execution time and space required for different methods are reported in Table 7.1. The results show that both JSCCA and the proposed method need the lower execution time and least memory consumption. The execution time and space required for the MTL-SDCCA is also high with respect to the methods JSCCA and MvMt-SCCA, but significantly lower than the method MT-SCCALR. Both time and space complexities of the MT-SCCALR are very high, because of the calculation of the Hessian matrix and the gradient of the graph-guided pairwise group lasso penalty.

Finally, it can be concluded from the above results that the MvMt-SCCA proposed in this chapter significantly reduces the space and time complexities compared to the MTL-SDCCA method proposed in the last chapter.

7.4.3.4 Ablation Study

This section presents the significance of incorporating various regularizer terms into the objective function of the proposed MvMt-SCCA method. The ablation study, in terms of the ability to select class-specific and class-consistent features, the mean canonical correlation



(a)

(b)

Figure 7.6: The performance of the proposed MvMt-SCCA method on SIG data set under different regularizer terms. (a) Heatmaps showing comparison of the canonical weights U and V. The rows 1 to 4 depict the results corresponding to the ground truth (GT), and different combinations of the MvMt-SCCA represented by Case-I to Case-IV, respectively. (b) Comparison with respect to mean testing canonical coefficient (left-column) and mean classification accuracy (right-column).

coefficient, and the mean classification accuracy is illustrated in Fig. 7.6 and Table 7.2. The experiments are conducted for the following four scenarios, based on two regularizer terms, namely, lasso and fused lasso.

- Case-I: None of the two regularizer terms is present in the objective function of the MvMt-SCCA, that is, λ_1, λ_2^m, τ_1, and τ_2^m all are set to zero for $m = 1, 2, \ldots, M$.
- Case-II: The lasso regularizer term, which controls the element-wise sparsity, is only present in the objective function, that is, $\lambda_1 \neq 0$ and $\lambda_2^m \neq 0$, while all the other hyperparameters for fused lasso, that is, τ_1, and τ_2^m are set to zero.
- Case-III: The fused lasso regularizer term, which controls the fusion of two similar

Table 7.2: F1-Score of Different Cases of MvMt-SCCA for U, V^1 and V^2.

MvMt-SCCA	U	V^1	V^2
Case-I	0.35	0.59	0.56
Case-II	0.46	0.74	0.70
Case-III	0.32	0.68	0.64
Case-IV	**0.96**	**0.94**	**0.94**

features, is only present in the objective function, that is, $\tau_1 \neq 0$ and $\tau_2^m \neq 0$, while all the other hyperparameters for lasso, that is, λ_1 and λ_2^m are set to zero.

- Case-IV: All the regularizer terms are present in the objective function of MvMt-SCCA, that is, all of the hyperparameters λ_1, λ_2^m, τ_1, and τ_2^m are all nonzero for $m = 1, 2, \ldots, M$.

The heatmap in Fig. 7.6a shows that, in Case-I, when neither of the two regularizer terms is part of the objective function of MvMt-SCCA, the proposed method picks up almost all the noisy features along with the class-specific and class-consistent features. Similarly, in Cases II and III, where the objective function includes either the lasso or fused lasso, the proposed method not only selects the class-specific and class-consistent features, but also captures the other noisy features. On the other hand, for Case IV, where both lasso and fused lasso are included in the objective function, the proposed method has identified the class-specific and class-consistent features more accurately. The F1-score shown in Table 7.2 gives a better understanding of the selected features for these four cases. It is clear that both the lasso and fused lasso regularizer terms are important for choosing features more accurately. The accurate feature selection is also responsible for the increase in canonical correlation coefficient and classification accuracy, which is shown in Fig. 7.6b. In Case IV, the MvMt-SCCA obtains the highest mean canonical correlation coefficient and classification accuracy. All the results justify the inclusion of two regularizer terms in the objective function of the MvMt-SCCA.

7.5 Real Neuroimaging Genetics Data Study

The real neuroimaging and genetic data sets used in this study are acquired from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD[1]. The main objective of ADNI has been to determine whether the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD) can be tracked using a combination of clinical and neuropsychological assessment, PET, serial MRI, and other biological markers. The current study includes 800 subjects considered from three diagnostic groups, namely, control normal (CN), MCI, and AD. Their basic characteristics are given in Table 7.3.

[1]The investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data, but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Table 7.3: Characteristics of the Participants

	CN	MCI	AD
Number	222	342	236
Gender (M/F)	121/101	179/163	134/102
Handedness (R/L)	202/20	305/37	212/24
Age (mean ± std)	75.66±5.87	71.69±6.98	73.54±8.03
Education (mean ± std)	16.44±2.66	16.16 ±2.64	16.10 ±2.60

7.5.1 Neuroimaging Data Acquisition and Preprocessing

The neuroimaging modalities of the participants include two types of PET scans, namely, 18-F florbetapir (AV-45) scans and fluorodeoxyglucose (FDG) scans. In AD patients, FDG and AV45 PET imaging are useful methods for measuring the amyloid-β ($A\beta$) burden and brain glucose metabolism, respectively [122]. The PET scans downloaded from ADNI go through some primary preprocessing. The steps consist of averaging across voxels, coregistration, standardization, intensity normalization and variable smoothing to obtain the standardized uptake value ratio (SUVR) images. The preprocessed images are further corrected for anterior commissure and posterior commissure, and registered to the standard Montreal Neurological Institute (MNI) space as $2 \times 2 \times 2$ mm^3 voxels. After that, the whole brain is subdivided, and 116 region of interest (ROI) level measurements are extracted according to the MarsBaR automated anatomical labelling (AAL) atlas [254]. These include glucose utilization for FDG scans and amyloid levels for AV45 images. These imaging measurements are pre-adjusted to exclude the effects of baseline age, gender, education, and handedness using the regression weights obtained from the healthy control subjects.

7.5.2 SNP Data Acquisition and Preprocessing

The SNP data of the same population is downloaded from the ADNI LONI website. The Human 610-Quad or OmniExpress Array (Illumina, Inc., San Diego, CA, USA) was used to genotype them. The SNP data is preprocessed to follow the standard quality control (QC) criteria using the PLINK tool (http://pngu.mgh.harvard.edu/) [201]. There are $1,008,823$ SNP markers to start with. The QC criteria for the SNP data include (1) call rate check per subject thresholded at 95%, (2) call rate check per SNP marker thresholded at 99.9%, (3) marker removal by the MAF ($\leqslant$ 5%), and (4) the Hardy-Weinberg equilibrium test thresholded at 1E$-$6. After the QC check, the marker count reduces to $307,281$. The majority of genes in the human genome can be found in the region of chromosome 19 sequence, where the average gene density is more than double that of the entire genome [85]. Thus, the current study includes the SNPs that are extracted from the chromosome 19 sequence. There are a total of 6032 SNPs left for further experiments.

The neuropsychological or cognitive assessment data is also collected for the 800 samples from the ADNI site (study data). The data contains four composite measures for executive functioning (ADNI-EF), memory (ADNI-MEM), language (ADNI-Lan), and visuospatial functioning (ADNI-VS). Thus, there are four modalities, namely, SNP (800×6032), PET-FDG (800×116), PET-AV45 (800×116) and cognitive measures (800×4), included in the current study. The objective is to find out a small set of SNPs that are correlated to a

Table 7.4: Comparison of Canonical Correlation Coefficient and Classification Accuracy on ADNI Data

Different Algorithms	SNP-FDG Correlation			SNP-AV45 Correlation		
	CN	MCI	AD	CN	MCI	AD
mSCCA [273]	0.01 ± 0.07	0.01 ± 0.07	0.01 ± 0.07	0.01 ± 0.05	0.01 ± 0.05	0.01 ± 0.05
MTSCCA [66]	0.06 ± 0.06	0.06 ± 0.06	0.06 ± 0.06	0.06 ± 0.05	0.06 ± 0.05	0.06 ± 0.05
JCB-SCCA [132]	0.07 ± 0.03	0.07 ± 0.03	0.07 ± 0.03	0.15 ± 0.10	0.15 ± 0.10	0.15 ± 0.10
MT-SCCAR [128]	0.14 ± 0.06	0.14 ± 0.06	0.14 ± 0.06	0.20 ± 0.06	0.20 ± 0.06	0.20 ± 0.06
MvMt-SCCA	**0.17 ± 0.05**	**0.16 ± 0.05**	**0.15 ± 0.07**	**0.28 ± 0.06**	**0.28 ± 0.06**	**0.28 ± 0.06**

Different Algorithms	Classification Accuracy		
	CN	MCI	AD
mSCCA [273]	0.56 ± 0.06	0.51 ± 0.04	0.55 ± 0.03
MTSCCA [66]	0.51 ± 0.08	0.51 ± 0.03	0.55 ± 0.02
JCB-SCCA [132]	**0.60 ± 0.06**	0.47 ± 0.04	0.53 ± 0.03
MT-SCCAR [128]	0.53 ± 0.07	0.51 ± 0.03	0.50 ± 0.02
MvMt-SCCA	0.56 ± 0.03	**0.52 ± 0.03**	**0.56 ± 0.02**

small set of imaging biomarkers with respect to AD subtypes, which may help to diagnose and prognosis of AD in the early stages.

7.5.3 Experimental Results and Discussion

In this section, the performance of the proposed method is compared with that of several multi-view methods, namely, mSCCA, MTSCCA, JCB-SCCA, and MT-SCCAR. It should be noted that the proposed method is the only one that identifies a correlation between SNP data and multiple imaging QTs with respect to disease-subtypes.

7.5.3.1 SNPs-QTs Association and Classification

The association between SNPs and multimodal imaging QTs has been assessed by reporting the canonical correlation coefficients and classification accuracy in Table 7.4 with respect to three disease-subtypes, namely, CN, MCI, and AD. A higher correlation value denotes a greater genetic connection with multimodal imaging QTs. The results show that the mSCCA has a very low correlation value for both the pairs SNP-FDG and SNP-AV45 with respect to the three disease courses since it tries to look into the strong correlation between every pair of imaging and genetic data, which may not be the case. Both MTSCCA and JCB-SCCA provide similar correlation value for the SNP-FDG pair, but JCB-SCCA has found a significantly higher correlation value for the SNP-AV45 pair for the three disease courses. The MT-SCCAR has a comparatively better correlation value than other existing methods. It has obtained correlation values of 0.14 and 0.20 for the pairs SNP-FDG and SNP-AV45, respectively, with respect to each of the three disease-subtypes. On the other hand, the proposed MvMt-SCCA has the higher correlation values compared to all the existing methods, and obtained 0.17, 0.16 and 0.15 for the pair SNP-FDG with respect

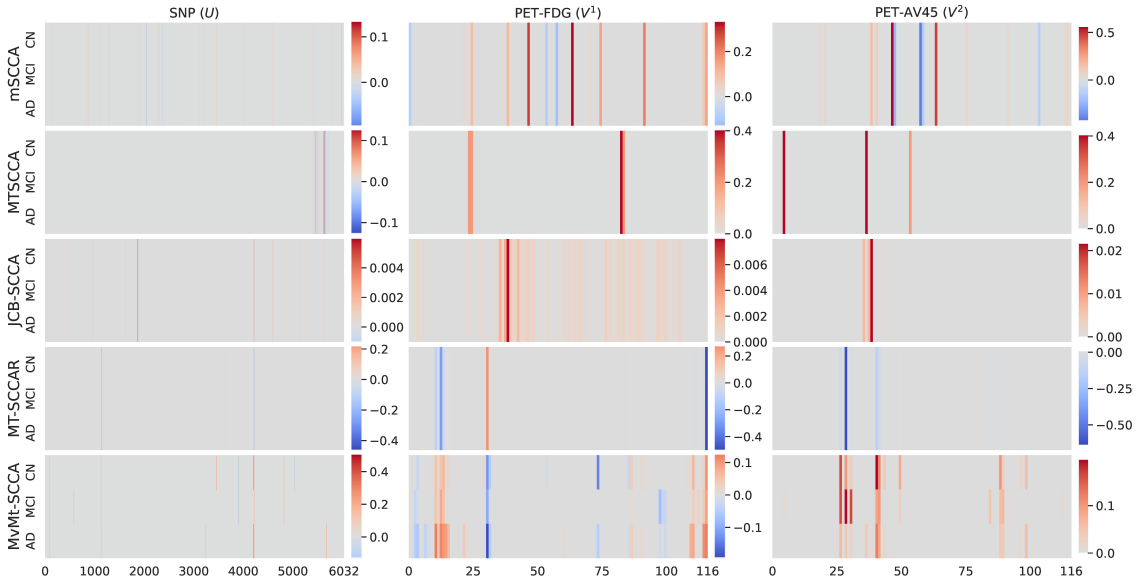


Figure 7.7: Mean canonical weights of SNPs (U: 1st column), PET-FDG (V^1: 2nd column) and PET-AV45 (V^2: 3rd column) obtained over five-fold cross-validation.

to CN, MCI, and AD, respectively. In the case of SNP-AV45, it has obtained 0.28 for all three diagnostic groups. Both the proposed method and MT-SCCAR have included the cognitive data into the model, which helps in associating the relevant and noise-free imaging measures to the genetic data, and results in a better correlation value. The selected SNPs and ROIs, which are responsible for correlating the genetic data with multimodal imaging QTs, are shown through the heatmaps of Fig. 7.7. The results show that only the proposed MvMt-SCCA method is able to select diagnose-consistent and diagnose-specific SNPs as well as imaging QTs, which is also justified by the higher correlation values, reported in Table 7.4.

The classification accuracy (one versus all) reported in Table 7.4 is obtained using the SVM, based on 30 features formed by concatenating 10 features for each of the three modalities, namely, SNP, PET-FDG, and PET-AV45. The accuracy establishes how well the selected features can separate one diagnostic group from the others. The classification accuracy reported in this table indicates that the proposed method has identified MCI and AD, better than any other methods. However, it provides second highest accuracy for the third diagnostic group by the selected features. This is mainly due to the fact that the selected features contain more shared correlation information compared to class-discriminative information.

7.5.3.2 Interpretation of Selected SNPs and Corresponding Genes

According to the canonical weight matrix U, there are 16 SNPs selected by the proposed method. The diagnose-consistent and diagnose-specific SNPs and their corresponding genes are reported by a Venn diagram in Fig. 7.8. There are 7 SNPs, which are same across all three diagnostic groups and are shown in middle of the first Venn diagram of Fig. 7.8. This 7 SNPs correspond to 5 genes, namely, TOMM40, NECTIN2, MUC16, PRTN3 and PRR12.

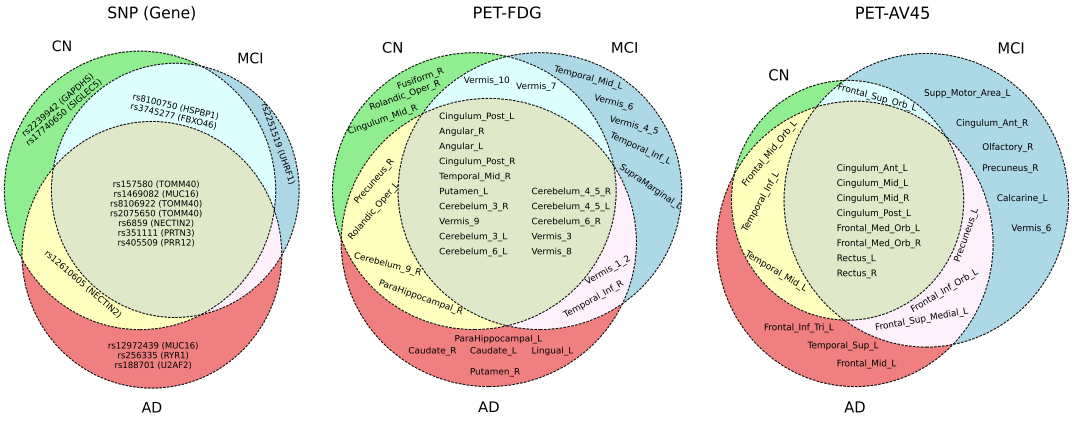


Figure 7.8: Venn diagrams for SNPs and ROIs of PET-FDG and PET-AV45, selected by the proposed method using ADNI data, corresponding to three diagnostic groups, namely, CN, MCI and AD.

Out of 7 SNPs, three SNPs, namely, rs2075650, rs157580, and rs8106922, come from the gene TOMM40. The genetic variants of translocase of outer mitochondrial membrane 40 (TOMM40) cause neuroinflammation and are believed to increase the risk of AD in different populations [46, 87]. Three genetic markers, namely, rs2075650, rs157580, and rs8106922 of TOMM40, selected by the proposed method, are proven to be associated with the progression of late-onset AD [48, 233].

Three SNPs, namely, rs12972439, rs256335 and rs188701, which are selected only for the diagnostic group AD, belong to the genes MUC16, RYR1 and U2AF2, respectively. The expression and function alteration of ryanodine receptor (RYR1) gene, which code for ryanodine receptors, and involve in βAPP processing and Aβ peptide production, may play an important role in the progression of AD [56]. In recent studies, the genes MUC16 [299] and U2AF2 [216] were discovered as the genetic biomarkers for AD [299]. Two SNPs, namely, rs6859 and rs12610605, selected for the AD group, come from the NECTIN2 gene. The SNP rs6859 is shared by all the three diagnostic groups, whereas rs12610605 is shared by the AD and CN. According to [172], NECTIN2 is expressed in particular brain regions and is crucial for processes including synaptic development and astrocyte and neuron homeostasis. The rs6859 SNP of NECTIN2 can be associated with cognitive declination in AD patients [208]. The other SNP, rs12610605 of the NECTIN2 gene, is also considered to be associated with AD [138]. Although the selected gene PRR12 is thought to have a role in brain development, its function and possible significance in human illness remain unknown [144]. All other selected genes, such as PRTN3 [214], HSPBP1 [61], and FBXP046 [142], selected for MCI or AD, are directly or indirectly related to the progression of AD. All the results of Fig. 7.7 and Fig. 7.8 suggest that the SNPs selected by the proposed method are either diagnose-consistent or diagnose-specific, and their corresponding genes are all significant towards the progression of AD.

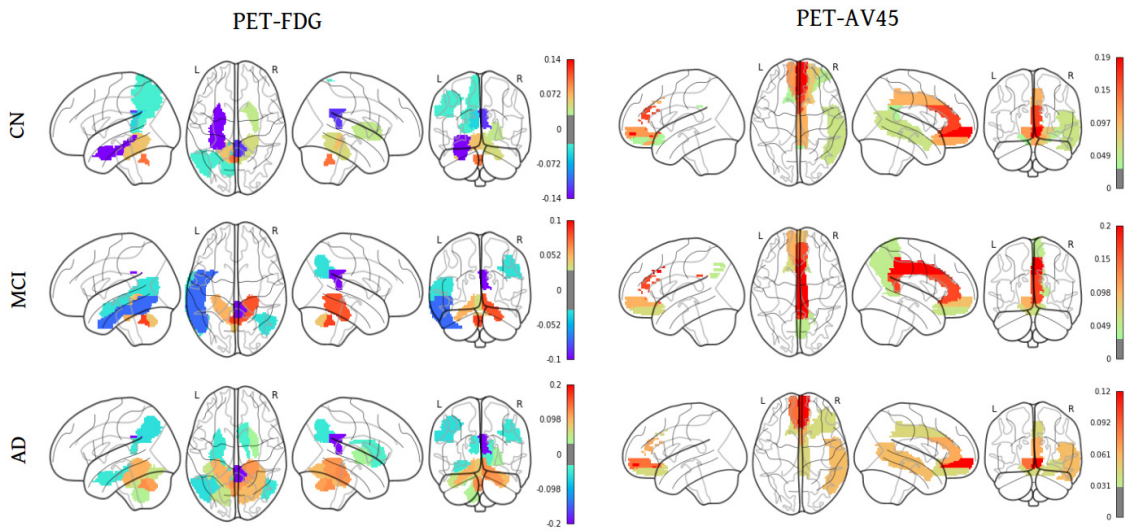


Figure 7.9: Visual representation of region of interests (ROIs) corresponding to PET-FDG (1st column) and PET-AV45 (2nd column) obtained by the proposed method. Rows 1, 2 and 3 correspond to ROIs of three diagnostic groups, namely, CN, MCI and AD.

7.5.3.3 Interpretation of Selected ROIs

The selected brain ROIs for the two imaging QTs, namely, PET-FDG and PET-AV45, are shown by a glass brain plot in Fig. 7.9, while the names of ROIs corresponding to diagnose-specific and diagnose-consistent are shown in Fig. 7.8 using Venn diagrams. It has been demonstrated that there is a strong correlation between the selected imaging QTs and SNPs in developing AD. Fig. 7.9 shows that there exist several ROIs for both modalities, which are mostly diagnose-consistent. Moreover, the selected ROIs for FDG and AV45 are almost similar with respect to the diagnostic status of CN and AD, which suggests that the degradation of those ROIs in normal people over time may increase the risk of AD. It can also be concluded from Fig. 7.8 and Fig. 7.9 that both imaging modalities have different ROIs selected by the proposed method. In case of the FDG modality, most of them are from the parieto-temporal lobe and cerebelum, whereas the selected ROIs in case of AV45 modality are mostly from the temporal lobe and medial frontal lobe.

The FDG-PET measures the cerebral metabolic rates of glucose (CMRglc), a stand-in for neuronal activity in AD patients. Some of the important ROIs, specific to AD, selected by the proposed method from the FDG modality are the posterior singulate cortex (PCC), angular gyrus, parahippocampal gyrus, caudate, putamen, lingual gyrus, and precuneus. The PCC is the only brain region that is selected for all the modalities and classes, and it is proven to be the most affected brain region in early AD. With strong anatomical and functional links to several other brain areas, the PCC is a region of the brain with increased metabolic activity. According to [146], AD patients show PCC hypometabolism, volume atrophy, and connection corruption. In a longitudinal study with MCI patients, it has been seen that changes in PCC connectivity over time are associated with cognitive decline [270]. Therefore, the selection of PCC is significant for MCI and AD. Along with PCC, in the parieto-temporal regions such as the angular gyrus [225] and medial temporal

Table 7.5: Comparison of Execution Time and Space

Different Algorithms	SIG		ADNI	
	Time (Sec.)	Space (GB)	Time (Sec.)	Space (GB)
mSCCA [273]	217.4	3.5	82.20	1.0
MTSCCA [66]	378.41	3.5	170.46	1.1
JCB-SCCA [132]	45.66	2.1	7.05	0.7
MT-SCCAR [128]	194.79	5.1	47.84	1.3
MvMt-SCCA	**1.93**	**0.2**	**0.14**	**0.2**

lobe such as the hippocampus and parahippocampus [181], AD patients consistently have CMRglc abnormalities. According to [115], people at risk for AD had far worse glucose metabolism in the right angular gyrus, compared to controls. In [126], it has been shown that the glucose metabolism in the precuneus gets reduced in patients developing AD. All the other selected ROI markers, such as parahippocampal gyrus [256], caudate [197], putamen [54], and lingual gyrus [159], have been shown to be associated with AD.

The PET-AV45 detects Aβ load in patients with AD and can provide information on both neurodegeneration and amyloid deposition. Some of the important brain ROIs selected by the proposed method from the AV45 modality are different parts of the frontal cortex (Frontal_Med_Orb, Frontal_Inf_Orb, Frontal_Mid_Orb), anterior cingulate cortex, different parts of the temporal lobe (Temporal Mid L, Temporal Inf L, Temporal Sup L), gyrus rectus, and precuneus. In [88], a systematic review has shown that the inferior frontal gyrus, anterior cingulate cortex, and medial temporal lobe are the areas most commonly related to ROIs associated with AD. The insula, orbitofrontal cortex, posterior cingulate cortex, medial frontal gyrus, and superior frontal gyrus are further important areas. All the ROIs, except insula, have been selected by the proposed method with respect to AD. The gyrus rectus is also considered to be associated with MCI and AD [38], which justifies the selection of this ROI.

7.5.4 Execution Time and Memory Usage

Finally, the performance of various approaches is compared in Table 7.5, in terms of memory usage and execution time. The results show that the proposed method performs significantly better, in terms of memory usage and execution time on both SIG and ADNI data sets. It takes only 1.93 seconds on SIG data having $10,000$ features, while 0.14 seconds on ADNI data set having 6032 features. All other methods, except JCB-SCCA, take a huge amount of time and space on both the data sets. This is due to the fact that the mSCCA, MTSCCA, and MT-SCCALR require computation of the inverse of a high-dimensional $(p \times p)$ covariance matrix, which has a complexity of $\mathcal{O}(p^3)$. On the other hand, the proposed method uses a lasso or a fused lasso solver, and optimizes the algorithm using the block coordinate descent technique, which has a complexity of $\mathcal{O}(p^2 \log p)$ at most. The coordinate-wise optimization technique helps the proposed method to occupy less memory while executing it on large data sets. The JCB-SCCA also uses the block coordinate descent algorithm to compute the canonical vectors, but it incorporates a GraphNet regularizer, for which there are more computations involving graph Laplacian in the intermediate steps

of the algorithm.

7.6 Conclusion

A multi-view multi-task SCCA model (MvMt-SCCA) has been proposed in this chapter, which integrates multimodal neuroimaging data, genetic data, and cognitive assessment data. It also uses the diagnostic status to find out complex connections between genetic and multimodal neuroimaging data with respect to disease-subtypes. It uses the lasso and fused lasso penalty together to identify the diagnose-specific and diagnose-consistent features, exploring the natural ordering in the SNPs and brain ROIs. The iterative algorithm that solves the proposed optimization problem uses the block coordinate descent technique, which is both time and space efficient. The proposed model has a time complexity of $\mathcal{O}(p^2 \log p)$, which is significantly lesser than that of the existing MTL based SCCA methods, which have a complexity of $\mathcal{O}(p^3)$ or $\mathcal{O}(p^4)$. The results on simulated data, SIG, establish that the proposed method can capture a higher correlation value, select relevant and noise-free features, with lower computational cost. The real neuroimaging study over ADNI data has identified some of the important genetic and imaging biomarkers for AD.

Chapter 8

Conclusion and Future Directions

This chapter provides an overview of the major contributions of MVL algorithms discussed in the various chapters of this thesis. This chapter also covers the potential applications and future directions of the proposed research work.

8.1 Major Contributions

A few novel MVL methods for multi-view data integration are presented in the thesis. There are five key challenges involved in multi-view data analysis, which include (i) integrating heterogeneous views while constructing discriminant subspaces, (ii) finding the intrinsic non-linear class-geometry of the data across all the views, (iii) handling "high-dimension low-sample size" nature of different views, (iv) selecting relevant and informative views while discarding the noisy and redundant views, and (v) applying MVL under a multi-task learning framework, for learning multiple related tasks simultaneously to improve the performance of single-task MVL. This thesis has addressed each of the aforementioned concerns. The key characteristics of the proposed methods in the thesis are briefly summarized and discussed next.

Chapter 3 introduces a novel supervised MVL algorithm, termed as CSP-MvCDA, by judiciously combining the merits of both MCCA and LDA. It jointly optimizes the inter-set correlation across all the views and intra-set discrimination in each view to obtain a common discriminative latent space, where the global class-structure of the data can be preserved. A label guided regularizer term, based on the locality preserving projection, is incorporated into the proposed framework to preserve the local class-structure of the data in each view. In effect, both local and global class-structure of the data can be preserved, which makes the model compatible for non-linear data. The proposed formulation of CSP-MvCDA can be solved easily by solving a generalized eigenvalue problem and used as an alternative to MCCA, in case of supervised multi-view subspace learning. Although the model CSP-MvCDA considers both intra-view and inter-view information to integrate multiple views, it may suffer from the presence of noisy and irrelevant views. For example, some of the views may not contain the desired information, and integrating all of them may lead to poor performance with respect to a particular task. Moreover, the "high-dimension low-sample size" nature of different views makes the class-scatter matrices ill conditioned. In

this regard, a new supervised graph regularized MVL algorithm, termed as SGR-MCCDA, based on the unsupervised subspace learning algorithm MCCA-MAXVAR, is introduced in Chapter 4. It uses a low rank representation of the class-scatter matrices to form a sample-by-sample non-singular normalized variance matrix, which makes it applicable for high-dimension low-sample size case. The method utilizes the prior knowledge of the data in terms of the class label and the structural geometry of the source vector. The class labels are used to encode the structural geometry of the source vectors by constructing the within-class and between-class graphs, which have been invoked as the regularizer term into the model. Due to the supervised graph information, the common subspace learned by the proposed method has also the ability to preserve the class structure of the data. Moreover, it learns a proper weight according to the relevance of each view, which helps selecting informative views while pay less attention to noisy and redundant views with respect to a desired task. Finally, to prove the diverse potency of both CSP-MvCDA and SGR-MCCDA methods, several cancer and benchmark data sets with different characteristics have been used.

In Chapter 5, both CSP-MvCDA and SGR-MCCDA have been applied successfully to extract features for imaging genetics data. In imaging genetics research, the main objective is to investigate the complex genotype-phenotype association for the disease under study. To understand the impact of genetic variations over the brain functions and structure, the genotypic data such as SNP can be integrated with the phenotypic data such as imaging quantitative traits. Both CSP-MvCDA and SGR-MCCDA extract features for imaging genetics correlation and select important biomarkers for the disease subtype classification, which highlights the key characteristics of the imaging genetics study. A comparative performance analysis with the three primary MVL algorithms of imaging genetics domain, namely, CCA, CoRe and CoopLe, is presented in this chapter. A real imaging genetics data set, obtained from ADNI cohort, is used for the analysis by using the AD related genetic variants (SNPs) and brain imaging (fMRI) modalities. The results establish that although, both the methods perform well in terms of classifying AD disease subtypes, they are not capable of handling the noise of such SNP or fMRI data and is unable to identify disease-specific imaging genetics association.

In the field of imaging genetics, identification of the diagnose-specific imaging measures and genetic markers are of great importance in personalized treatments. Most of the models in this domain are based on sparse CCA, and thus unsupervised in nature. Though few supervised models have been developed in the recent years, they cannot identify the diagnose-specific biomarkers. This necessitates the sparse multi-task model in this field. In this regard, a supervised multi-task model, termed as MTL-SDCCA, has been proposed in Chapter 6, by judiciously integrating the merits of both CCA and LDA under MTL framework. It uses lasso and group lasso penalties to select the diagnosis-specific and diagnosis-consistent features from the large number of features to identify group-wise imaging genetic associations. The performance of the proposed method is compared with that of several existing supervised models, namely, CoRe, CoopLe, JSCCA, JCB-SCCA, MT-SCCALR and MT-SCCAR, considering two synthetic data and one real imaging genetic data obtained from ADNI cohort. The results show that the proposed method has better feature selection capability than several existing methods and can successfully identify the class-specific and class-consistent features, irrespective of the data sets used. However, like most of the existing approaches, the proposed MTL-SDCCA require computing the inverse

of a very high-dimensional covariance matrix, which has large time and space complexity. It makes the models inefficient for large-scale imaging genetics studies. Moreover, the proposed model is unable to utilize the multiple imaging and genetic modalities. In this regard, a multi-view multi-task SCCA model, termed as MvMt-SCCA, has been proposed in Chapter 7, which integrates multimodal neuroimaging data, genetic data, and cognitive assessment data for identifying disease-specific imaging genetics association. It uses the lasso and fused lasso penalty together to identify modality-wise diagnose-specific and diagnose-consistent features, exploring the natural ordering in the SNPs and brain ROIs. An iterative algorithm based on the block coordinate descent technique is introduced to solve the optimization problem, which is both time and space efficient. The proposed model has a time complexity of $\mathcal{O}(p^2 \log p)$, which is significantly lesser than that of the existing MTL based SCCA methods, which have a complexity of $\mathcal{O}(p^3)$ or $\mathcal{O}(p^4)$, where p is the maximum feature size of the imaging or genetic data. The results on simulation and real ADNI data supports the claim.

8.2 Future Directions

The work presented in this thesis covers a wide range of significant MVL algorithms that can be extended further to advance the multi-view data analysis under both single and multi-task learning framework. The research can proceed with the following improvements and future directions:

- **Hierarchical Fusion**: All the MVL algorithms proposed in this thesis fuses multiple views, which are all in the same level of abstraction. However, complex tasks in the fields of computer vision and biomedical imaging require integration of features at various stages of processing. By combining features at different levels, the model can produce richer and significant representations. While high-level features can offer a wider context and improve the model's ability to generalize, low-level features might capture crucial details. Thus, all the MVL algorithms presented in this thesis can be extended to incorporate the hierarchical feature extraction technique for model generalization.

- **Incomplete Views**: The thesis makes the assumption that every view is complete, that is, every view has the same set of common samples. However, in real-world applications, it may occur that measurement mistakes and pre-processing leave the data set with incomplete views. Sometimes, the information about a sample may not be present in some of the views or the information about the sample may partially be observed in some of the views. In both the scenarios, the MVL algorithms should work. Therefore, the multi-view data integration techniques discussed in the thesis may be extended to deal with the view missing or variable missing problem and recover missing samples by creating connections between the views, without discarding the missing sample from all views, as described by Xu (2015) [280].

- **Updation in Database:** The MVL models described in the thesis are constructed under the premise that all the views are accessible a priori for the analysis. However, a substantial volume of data is consistently being added to the current databases.

Either more observations are incorporated or an entirely new perspective is developed to enhance the interpretation of the current data. The databases in TCGA have undergone more than 20 updates in the past 5 years, including additions to observations and views. Therefore, it is necessary to alter the learning process of the proposed MVL models in such a manner so that they can effectively adjust to any change in the databases.

- **Optimization on Manifold**: Several real-world data sets exhibit significant structures that are located on a low-dimensional manifold within a higher-dimensional Euclidean space [221]. Since each view might have its own manifold structure, the multi-view data set can be seen as a combination of manifolds. According to this hypothesis, all the techniques for integrating multi-view data provided in the thesis may be improved to include the construction of discriminant subspaces from a low-dimensional manifold.

- **Deep Network Based Optimization**: Shallow optimization strategies and analysis based on eigendecomposition are the main interests of the thesis. On the other hand, non-linear transformations may be learned from various modalities via deep learning architectures like the multimodal deep Boltzmann machines [234]. An attempt will be made to integrate multi-view data sets utilizing a deep learning framework as extensions of the proposed methods in order to learn the latent subspaces. The deep optimization techniques can be used to solve the optimization problems of different chapters.

- **Active Learning**: All the MVL algorithms proposed in this thesis are supervised in nature. In supervised MVL, data from multiple viewpoints must be labeled in order to train the models. However, it will cost more to label the data for multiple viewpoints. Consequently, it is important to decrease the amount of labeled data without compromising the multi-view learning efficiency. Active learning is an useful approach which selects fewer and more valuable data points for labeling. It constructs the classifier with data acquisition by rating the unlabeled data to make recommendations for the following query with the highest training utility. As a result, the training set can be kept as small as feasible while the learner's full potential is explored on both labeled and unlabeled data. This lowers the cost of data labeling dramatically and exploit the most informative data samples. Thus integrating MVL and active learning jointly can improve the efficiency of MVL algorithms by selecting more informative data points across multiple views.

- **Views Observed in Heterogeneous Measurement Spaces**: All the MVL algorithms proposed in the thesis assume that every view is seen in a real-valued Euclidean space. However, some of the views may not be embedded into the Euclidean space in realistic applications. For example, textual information and categorical data may represent different views along with the Euclidean data. Thus, certain modifications can be made to the proposed MVL algorithms, which will enable the integration of heterogeneous multi-view data consisting of several views in various measurement spaces.

- **Sparse Data Integration**: The sparsity of the data sets is not directly addressed

by the techniques in this thesis. However, while training model, data sparsity can lead to issues like overfitting and sub-optimal solutions. Therefore, it is necessary to enhance the data integration methods so that the sparse multi-view data sets can be merged effectively.

- **Weak Supervision**: All the algorithms proposed in this thesis needs labeled data in order to train the models. However, labeling data manually one by one take huge effort and time. In a weak supervision model, artificial labeling, using high-level and frequently noisy sources of supervision, can generate significantly larger training sets much faster than the manual supervision. By enabling noisy labeled samples to supervise the multi-view classification algorithms, new methods can be developed to enhance the learning performance.

- **Task-heterogeneity in Multi-task Learning**: The thesis proposes a multi-task multi-view model in Chapter 7 for imaging genetics studies. The tasks here are homogeneous and uses same set of labels for different views. Task heterogeneity is major challenge in MTL which can be dealt with developing new methodologies for multi-task multi-view learning.

- **Federated Multi-Task Learning**: The classical machine learning techniques, proposed in the thesis, use centralized data training, in which the training is done at the central server once the data are collected. This training presents multiple privacy risks to participant data when shared with the central cloud server, despite of significant convergence. Through federated learning, users can work together to train local models on local data without disclosing private information to a central cloud server. However, while training machine learning models across distributed networks of devices, federated learning presents significant statistical and technical issues. Multi-task learning is well-suited to manage this setting's statistical problems. In this regard, new federated multi-task learning models can be developed to enhance the generalization performance of both multi-task learning and federated learning.

Appendix A

Description of Data Sets

A quick summary of the multi-view benchmark and multi-omics cancer data sets has been presented in this appendix. These data sets have been used to compare the performance of the proposed method with the existing multi-view learning algorithms, in the thesis. There are five multi-omics cancer data sets, namely, breast invasive carcinoma (BRCA), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) and lower grade glioma (LGG), colorectal carcinoma (CRC) and glioblastoma multiforme (GBM); and eight benchmark data sets, namely, CiteSeer, Cora, Reuters, Nus-Wide-Object (NW-Object), Caltech-101, Caltech-20, ALOI and 100Leaves, included in the thesis.

A.1 Benchmark Data Sets

This section presents a brief description of the eight benchmark data sets.

1. **CiteSeer**: The CiteSeer dataset have been downloaded from http://lig-membres.imag.fr/grimal/data.html. The set is generated by sampling scientific documents from CiteSeer digital library. 3312 documents from the six labels—Agents, IR, DB, AI, HCI, and ML—are included in the archive. It consists of four perspectives on the same documents: cites, content, inbound, and outbound. In the content view, the papers are described in 3703 words. In the inbound, outbound, and citations views, there are 4732 links that connect the documents. Every word with a document frequency lower than ten is eliminated. A word vector with values of 0 or 1, which denotes the existence or absence of the matching word in the document, describes every publication in the database.

2. **Cora**: The Cora dataset is obtained from http://lig-membres.imag.fr/grimal/data.html. Over the course of the seven labels—Neural_Networks, Rule_Learning, Reinforcement_Learning, Probabilistic_Methods, Theory, Genetic_Algorithms, and Case_Based—the repository includes 2708 documents. It consists of four perspectives on the same documents: content, inbound, outbound, and citations. The materials are explained in 1433 words in the content view and 5429 links in the cites, inbound, and outbound views. The documents-words matrix, which has 0/1 values, represents a word's presence or absence in a document.

3. **Reuters**: This multilingual data has been downloaded from http://archive.ics.uci.edu/ml/machine-learning-databases/00259/. This collection includes the common set of six categories' feature characteristics from English-language texts along with their matching translations into French, German, Spanish, and Italian. Multilingual classification and multi-view learning research can benefit from the use of this collection. After being translated and preprocessed, the documents are provided as feature characteristics in a "bag of words" style. The 18758 documents are described using the following five feature sets and divided into six categories: CCAT, C15, ECAT, E21, GCAT, and M11.

4. **Nus-Wide-Object (NW-Object)**: This data has been downloaded from https://lms.comp.nus.edu.sg/wp-content/uploads/2019/research/nuswide/NUS-WIDE.html. This database was developed for object recognition tasks by the National University of Singapore's Lab for Media Search. The 30000 photos are organised into 31 classes. The five feature sets correspond to the 30000 photos in the database.

5. **Caltech**: This data has been downloaded from http://www.vision.caltech.edu/Image_Datasets/Caltech101/. Images of items in 101 categories make up Caltech-101. For every category, there are 40–800 photos. The majority of the categories contain roughly fifty photographs that were gathered by Fei-Fei Li, Marco Andreetto, and Marc Aurelio Ranzato in September 2003. Each image is about 300×200 pixels in size. The Caltech-20 is a subset of Caltech-101, with a total of 20 classes. The performance of the algorithms proposed in chapter 3 and 4 are examined using both the Caltech-101 and Caltech-20 data sets in the current research.

6. **ALOI**: This is the Amsterdam Library of Object Image data set, available at https://elki-project.github.io/datasets/multiview. The data set consists of 11,0250 images of 1000 small objects. Each image is represented with four types of features, that is, RGB color histogram, HSV color histogram, color similiarity and Haralick features. In the current research work, a subset of 10800 samples with 100 classes having 108 samples in each, has been used to conduct the experiments.

7. **100Leaves**: It is a one-hundred plant species leaves data set https://archive.ics.uci.edu/ml/datasets/One-hundred+plant+species+leaves+data+set. The data set includes 1,600 samples, where each type of leaf consists of sixteen samples. The texture histogram, fine scale margin, and form descriptors are the three sets of image attributes that each sample is represented by. A 64-element vector is provided for every feature for every leaf sample. 64-element feature vectors are contained in a single file. The class label starts each row. The feature vector is comprised of the remaining 64 elements.

A.2 Omics Data Sets

This section describes the five multi-omics cancer data sets, obtained from The Cancer Genome Atlas (TCGA) [2] and downloaded through the Genomic Data Commons (GDC) Data Portal [1].

1. **breast invasive carcinoma (BRCA):** Breast cancer is one of the most common cancers with greater than 1,300,000 cases and 450,000 deaths each year worldwide [214]. During the last 15 years, four intrinsic molecular subtypes of breast cancer, namely, Luminal A, Luminal B, HER2-enriched, and Basal-like, have been identified and intensively studied [111, 135, 232]. The BRCA data set consists of 134 samples and is classified into four groups, namely, LuminalA, LuminalB, HER2-enriched, and Basal-like subtype, respectively.

2. **Cervical Carcinoma (CESC)** More than any other gynaecological tumour, this disease causes 266,000 deaths and 528,000 new cases every year, globally [75]. Three subtypes of CESC have been identified by the TCGA research network through extensive integrated analysis [188]. There are 104 samples in the CESC data set: 26 samples belong to the adenocarcinoma-rich subgroup, 32 samples are in the keratin-low squamous subgroup, and 46 samples are in the keratin-high squamous subgroup.

3. **Lower Grade Glioma (LGG):** According to World Health Organization, lower grade gliomas are benign tumours classified as grades I and II. They are made up of diffuse low-grade and intermediate-grade gliomas. It is exceedingly difficult to predict LGG based on histologic class because of its highly varied clinical behavior [189]. Some develop into glioblastoma slowly, while others do so rapidly. The performance of each method is analyzed using 374 LGG samples in the current study effort. With 180 samples, the first subtype shows IDH mutation without 1p/19q co-deletion. There are 129 instances in the second subtype that have both 1p/19q co-deletion and an IDH mutation. The third subtype, with 65 samples, is the wild-type IDH subtype.

4. **Colorectal Carcinoma (CRC):** It accounts for 9% of all cancer-related fatalities and is the third most frequent cancer in both men and women [166]. As components of the digestive system, the colon and the rectum can develop cancer. The CRC data set has 261 samples. The CRC samples are classified into two subtypes, colon carcinoma and rectum carcinoma, with 192 and 69 samples, respectively, based on the place of origin.

5. **Glioblastoma multiforme (GBM):** It is the most common primary brain tumour in adults and the most aggressive and widespread subtype of glioma. It has four subtypes identified in the study by Veerhak *et al.* [70]. The subtypes consist of proneural, neural, classical, and mesenchymal. The data set consists of 213 samples from four genomic modalities, namely, miRNA, RNA, DNA, and CNV. The data set contains 39, 52, 21, 64, and 37 samples of proneural, classical, G-CIMP, mesenchymal, and neural subtypes, respectively.

These subtypes offer a framework for patient classification and focused therapy trials, and they are clinically significant. Each and every omics data collection has made use of the information regarding DNA methylation (mDNA). On the other side, the LGG data set uses reverse phase protein array expression (RPPA). MicroRNA (miRNA) is used in sequence form in the previously described data sets, while GBM uses the information of miRNA in expression form. Gene expression offers information associated to genes in the GBM data set, whereas gene details have been extracted from RNA sequences in the

LGG data. The GBM and LGG data sets make use of copy number segmentation (CNS) information.

Data Platforms and Preprocessing

The reverse-phase protein array data from the MDA_RPPA_Core platform is used to obtain the protein modality. The number of proteins is different for each sample. Only a set of common proteins which is present in all the samples is considered to construct the protein expression data set. The H-miRNA_8x15Kv2 and H-miRNA_8x15K platforms are used to extract the information of miRNA for OV and GBM, respectively. On the other hand, the sequence-based miRNA expression data from the Illumina HiSeq platform is used for other data sets, which contain RPM (reads per million miRNA mapped) values for 1046 miRNAs. The miRNA sequence data is also log-transformed. The expression values of this modality are not available for most of the samples in these data sets. As there are too many missing values, the feature having more than 5% missing values is discarded. The missing values are replaced by 0 for the feature which has less than or equal to 5% missing values.

For the DNA methylation modality, methylation β-values from Illumina Human Methylation 450 platform are used on LUNG, KIDNEY, and LGG data sets. On the other hand, methylation β-values of GBM and OV data sets consist of Illumina Human Methylation 27 platform. The Human Methylation 450 gives methylation β-values of approximately 450,000 CpG sites, while Human Methylation 27 covers 27,000 CpG sites. The CpG locations having missing gene information are excluded. In the current research work, the top 2,000 CpG sites having the most variance, are used. In all omics data sets, CNV is generated from Affymetrix SNP Array 6.0 platform. To reduce the redundant copy number regions, the CNregions function of iCluster+ R-package [173] has been used in raw copy number segmented data. There is an epsilon parameter in the CNregions function, which has been used to compute the maximum Euclidean distance between adjacent probes tolerated for defining a non-redundant region. The value of this epsilon parameter gives the number of non-redundant copy number regions. According to [173], the value of this epsilon parameter has to be selected in this manner so that the reduced dimension becomes less than 10,000. In all the data sets, the default value that is 0.005 has been considered for the epsilon parameter of the CNregions function. For the RNA modality of LUNG, KIDNEY, and LGG data sets, RNA-sequence data from the Illumina HiSeq platform is used which contains normalized RPKM (reads per kilobase of exon per million) counts for 20,531 genes. The data is then log transformed and 2,000 most variable genes based on their expression profile across the samples are considered. The RNA modality of the GBM and OV data sets are prepared using the platform HT_HG-U133A and AgilentG4502A_07_3, respectively, and consists of log-ratio based expression data for 12,042 genes amongst which 2,000 genes having the most variance are considered.

A.3 Simulated Data Sets

This section describes the canonical weight matrices to construct the simulated imaging genetics data sets used in Chapter 6 and Chapter 7. The concept of setting canonical

weights has been adopted from [66].

- **DS-A:** Set the canonical weight matrices U and V as follows: $u_1 = (\underbrace{1,\ldots,1}_{20},\underbrace{0,\ldots,0}_{60},\underbrace{1,\ldots,1}_{10},\underbrace{0,\ldots,0}_{50},\underbrace{1,\ldots,1}_{40},\underbrace{0,\ldots,0}_{80},\underbrace{1,\ldots,1}_{20},\underbrace{0,\ldots,0}_{20})$, $u_2 = (\underbrace{1,\ldots,1}_{20},\underbrace{0,\ldots,0}_{60},\underbrace{1,\ldots,1}_{10},\underbrace{0,\ldots,0}_{50},\underbrace{1,\ldots,1}_{40},\underbrace{0,\ldots,0}_{80},\underbrace{1,\ldots,1}_{20},\underbrace{0,\ldots,0}_{20})$, $u_3 = (\underbrace{1,\ldots,1}_{20},\underbrace{0,\ldots,0}_{60},\underbrace{1,\ldots,1}_{10},\underbrace{0,\ldots,0}_{50},\underbrace{1,\ldots,1}_{40},\underbrace{0,\ldots,0}_{80},\underbrace{1,\ldots,1}_{20},\underbrace{0,\ldots,0}_{20})$, $v_1 = (\underbrace{0,\ldots,0}_{60},\underbrace{1,\ldots,1}_{30},\underbrace{0,\ldots,0}_{90},\underbrace{1,\ldots,1}_{20},\underbrace{0,\ldots,0}_{130},\underbrace{1,\ldots,1}_{30},\underbrace{0,\ldots,0}_{90})$, $v_2 = \underbrace{0,\ldots,0}_{60},\underbrace{1,\ldots,1}_{30},\underbrace{0,\ldots,0}_{90},\underbrace{1,\ldots,1}_{20},\underbrace{0,\ldots,0}_{130},\underbrace{1,\ldots,1}_{30},\underbrace{0,\ldots,0}_{90}$, and $v_3 = \underbrace{0,\ldots,0}_{60},\underbrace{1,\ldots,1}_{30},\underbrace{0,\ldots,0}_{90},\underbrace{1,\ldots,1}_{20},\underbrace{0,\ldots,0}_{130},\underbrace{1,\ldots,1}_{30},\underbrace{0,\ldots,0}_{90}$.
- **DS-B:** Set the canonical weight matrices U and V as follows: $u_1 = (\underbrace{0,\ldots,0}_{800},\underbrace{1,\ldots,1}_{200},\underbrace{0,\ldots,0}_{13025},\underbrace{2,\ldots,2}_{150},\underbrace{0,\ldots,0}_{160},\underbrace{0,\ldots,0}_{80},\underbrace{1,\ldots,1}_{20})$, $u_2 = (\underbrace{0,\ldots,0}_{800},\underbrace{1,\ldots,1}_{200},\underbrace{0,\ldots,0}_{12050},\underbrace{2,\ldots,2}_{300},\underbrace{0,\ldots,0}_{1950},\underbrace{5,\ldots,5}_{200},\underbrace{0,\ldots,0}_{4500})$, $u_3 = (\underbrace{0,\ldots,0}_{800},\underbrace{1,\ldots,1}_{200},\underbrace{0,\ldots,0}_{9500},\underbrace{3,\ldots,3}_{300},\underbrace{0,\ldots,0}_{9300})$, $v_1 = (\underbrace{0,\ldots,0}_{1000},\underbrace{1,\ldots,1}_{250},\underbrace{0,\ldots,0}_{150},\underbrace{5,\ldots,5}_{70},\underbrace{0,\ldots,0}_{1030},\underbrace{2,\ldots,2}_{80},\underbrace{0,\ldots,0}_{420})$, $v_2 = (\underbrace{0,\ldots,0}_{1000},\underbrace{1,\ldots,1}_{250},\underbrace{0,\ldots,0}_{750},\underbrace{3,\ldots,3}_{60},\underbrace{0,\ldots,0}_{640},\underbrace{5,\ldots,5}_{70},\underbrace{0,\ldots,0}_{230})$, and $v_3 = (\underbrace{0,\ldots,0}_{1000},\underbrace{1,\ldots,1}_{250},\underbrace{0,\ldots,0}_{770},\underbrace{2,\ldots,2}_{80},\underbrace{0,\ldots,0}_{900})$.
- **SIG:** Set the canonical weight matrices U, V^1 and V^2 as follows: $u_1 = (\underbrace{0,\ldots,0}_{2000},\underbrace{1,\ldots,1}_{500},\underbrace{0,\ldots,0}_{1500},\underbrace{1,\ldots,1}_{500},\underbrace{0,\ldots,0}_{5500})$, $u_2 = (\underbrace{0,\ldots,0}_{2000},\underbrace{1,\ldots,1}_{500},\underbrace{0,\ldots,0}_{4550},\underbrace{2,\ldots,2}_{700},\underbrace{0,\ldots,0}_{2250})$, $u_3 = (\underbrace{0,\ldots,0}_{1000},\underbrace{2,\ldots,2}_{400},\underbrace{0,\ldots,0}_{600},\underbrace{1,\ldots,1}_{500},\underbrace{1,\ldots,1}_{7500})$; $v_1^1 = (\underbrace{0,\ldots,0}_{1000},\underbrace{2,\ldots,2}_{300},\underbrace{0,\ldots,0}_{1720},\underbrace{3,\ldots,3}_{200},\underbrace{0,\ldots,0}_{1780})$, $v_2^1 = (\underbrace{0,\ldots,0}_{1000},\underbrace{2,\ldots,2}_{300},\underbrace{0,\ldots,0}_{1200},\underbrace{5,\ldots,5}_{300},\underbrace{0,\ldots,0}_{2200})$, $v_3^1 = (\underbrace{0,\ldots,0}_{1000},\underbrace{2,\ldots,2}_{300},\underbrace{0,\ldots,0}_{2700},\underbrace{5,\ldots,5}_{300},\underbrace{0,\ldots,0}_{700})$, $v_1^2 = (\underbrace{0,\ldots,0}_{1000},\underbrace{3,\ldots,3}_{300},\underbrace{0,\ldots,0}_{1720},\underbrace{3,\ldots,3}_{280},\underbrace{0,\ldots,0}_{1700})$, $v_2^2 = (\underbrace{0,\ldots,0}_{3020},\underbrace{3,\ldots,3}_{280},\underbrace{0,\ldots,0}_{700},\underbrace{5,\ldots,5}_{200},\underbrace{0,\ldots,0}_{800})$, $v_3^2 = (\underbrace{0,\ldots,0}_{2000},\underbrace{5,\ldots,5}_{400},\underbrace{0,\ldots,0}_{620},\underbrace{3,\ldots,3}_{280},\underbrace{0,\ldots,0}_{1700})$.
- **Small SIG:** Set the canonical weight matrices U, V^1 and V^2 as follows: $u_1 = (\underbrace{0,\ldots,0}_{50},\underbrace{1,\ldots,1}_{30},\underbrace{0,\ldots,0}_{320},\underbrace{3,\ldots,3}_{50},\underbrace{0,\ldots,0}_{50})$, $u_2 = (\underbrace{0,\ldots,0}_{50},\underbrace{1,\ldots,1}_{30},\underbrace{0,\ldots,0}_{70},\underbrace{2,\ldots,2}_{40},\underbrace{0,\ldots,0}_{310})$, $u_3 = (\underbrace{0,\ldots,0}_{50},\underbrace{1,\ldots,1}_{30},\underbrace{0,\ldots,0}_{180},\underbrace{3,\ldots,3}_{40},\underbrace{1,\ldots,1}_{300})$; $v_1^1 = (\underbrace{0,\ldots,0}_{80},\underbrace{2,\ldots,2}_{40},$

$\underbrace{0,\ldots,0}_{200},\underbrace{3,\ldots,3}_{50},\underbrace{0,\ldots,0}_{230}),\ v_2^1 = (\underbrace{0,\ldots,0}_{80},\underbrace{2,\ldots,2}_{40},\underbrace{0,\ldots,0}_{500},\underbrace{5,\ldots,5}_{40},\underbrace{0,\ldots,0}_{40}),\ v_3^1 = (\underbrace{0,\ldots,0}_{80},\underbrace{2,\ldots,2}_{40},\underbrace{0,\ldots,0}_{510},\underbrace{5,\ldots,5}_{40},\underbrace{0,\ldots,0}_{30}),\ v_1^2 = (\underbrace{0,\ldots,0}_{80},\underbrace{3,\ldots,3}_{40},\underbrace{0,\ldots,0}_{400},\underbrace{4,\ldots,4}_{20},\underbrace{0,\ldots,0}_{60}),\ v_2^2 = (\underbrace{0,\ldots,0}_{320},\underbrace{5,\ldots,5}_{50},\underbrace{0,\ldots,0}_{150},\underbrace{4,\ldots,4}_{20},\underbrace{0,\ldots,0}_{60}),\ v_3^2 = (\underbrace{0,\ldots,0}_{200},\underbrace{5,\ldots,5}_{40},\underbrace{0,\ldots,0}_{260},\underbrace{4,\ldots,4}_{20},\underbrace{0,\ldots,0}_{60}).$

List of Related Publications

International Journal Papers

- **Sankar Mondal** and Pradipta Maji. Multi-Task Learning and Sparse Discriminant Canonical Correlation Analysis for Identification of Diagnosis-Specific Genotype-Phenotype Association. ***IEEE/ACM Transactions on Computational Biology and Bioinformatics***, vol. 21, no. 5, pp. 1390-1402, Sept.-Oct. 2024, DOI:10.1109/TCBB.2024.3386406.

- **Sankar Mondal** and Pradipta Maji. Class-Structure Preserving Multi-View Correlated Discriminant Analysis for Multiblock Data. ***International Journal of Machine Learning and Cybernetics. Springer***, pages 1-22, 2024. DOI:10.1007/s13042-024-02270-9

- **Sankar Mondal** and Pradipta Maji. Correlated Joint Subspace Learning: A Framework for Class-Structure Preservation. ***IEEE Transactions on Knowledge and Data Engineering,*** pages 1-14. (**Submitted**, Manuscript ID: TKDE-2024-12-3185)

- **Sankar Mondal** and Pradipta Maji. Multi-View Multi-Task Sparse Canonical Correlation Analysis for Imaging Genetics Study. ***IEEE Transactions on Emerging Topics in Computational Intelligence***, pages 1-10. (**Submitted**, Manuscript ID: TETCI-2024-1533).

References

[1] GDC Data Portal. https://gdc-portal.nci.nih.gov/.

[2] TCGA Research Network. http://cancergenome.nih.gov/.

[3] A. H. Abdulnabi, B. Shuai, Z. Zuo, L.-P. Chau, and G. Wang. Multimodal Recurrent Neural Networks With Information Transfer Layers for Indoor Scene Labeling. *IEEE Transactions on Multimedia*, 20(7):1656–1671, 2018.

[4] S. Abney. Bootstrapping. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 360–367, 2002.

[5] Ayan Acharya, Raymond J Mooney, and Joydeep Ghosh. Active Multitask Learning Using Both Latent and Supervised Shared Topics. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 190–198. SIAM, 2014.

[6] F. Aiolli and M. Donini. EasyMKL: A Scalable Multiple Kernel Learning Algorithm. *Neurocomputing*, 169:215–224, 2015.

[7] S. Akaho. A Kernel Method for Canonical Correlation Analysis. In *Proceedings of the International Meeting of Psychometric Society*, 2001.

[8] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy. Deep Variational Information Bottleneck. In *Proceedings of the International Conference on Learning Representations*, 2017.

[9] Muaath Ebrahim AlMansoori, Sherlyn Jemimah, Ferial Abuhantash, and Aamna AlShehhi. Predicting early Alzheimer's with blood biomarkers and clinical features. *Scientific Reports*, 14(1):6039, 2024.

[10] M. R. Amer, T. Shields, B. Siddiquie, A. Tamrakar, A. Divakaran, and S. Chai. Deep Multimodal Fusion: A Hybrid Approach. *International Journal of Computer Vision*, 126(2-4):440–456, 2018.

[11] Rie Kubota Ando, Tong Zhang, and Peter Bartlett. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of machine learning research*, 6(11), 2005.

[12] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR, 2013.

[13] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. *Advances in neural information processing systems*, 19, 2006.

[14] F. R. Bach and M. I. Jordan. Kernel Independent Component Analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

[15] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple Kernel Learning, Conic Duality, and the SMO Algorithm. In *Proceedings of the Twenty-First International Conference on Machine Learning*, page 6, 2004.

[16] Yuntong Bai, Zille Pascal, Wenxing Hu, Vince D. Calhoun, and Yu-Ping Wang. Biomarker identification through integrating fmri and epigenetics. *IEEE Transactions on Biomedical Engineering*, 67(4):1186–1196, 2020.

[17] BJ Bakker and TM Heskes. Task Clustering and Gating for Bayesian Multitask Learning. *Journal of Machine Learning Research*, 2003.

[18] M.F. Balcan, A. Blum, and Y. Ke. Co-training and Expansion: Towards Bridging Theory and Practice. In *Proceedings of the 17th International Conference on Neural Information Processing SystemsDecember*, pages 89–96, 2004.

[19] Xiangze Bao, Yun-Hao Yuan, Yun Li, Jipeng Qiang, and Yi Zhu. Learning Supervised Covariation Projection Through General Covariance. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[20] Gaston Baudat and Fatiha Anouar. Generalized Discriminant Analysis Using a Kernel Approach. *Neural Computation*, 12(10):2385–2404, 2000.

[21] A. Benton, H. Khayrallah, B. Gujral, D. A. Reisinger, S. Zhang, and R. Arora. Deep Generalized Canonical Correlation Analysis. *arXiv:1702.02519*, 2017.

[22] Xia-an Bi, Xi Hu, et al. Multimodal Data Analysis of Alzheimer's Disease Based on Clustering Evolutionary Random Forest. *IEEE Journal of Biomedical and Health Informatics*, 24(10):2973–2983, 2020.

[23] Xia-an Bi, Yingchao Liu, Yiming Xie, Xi Hu, and Qinghua Jiang. Morbigenous brain region and gene detection with a genetically evolved random neural network cluster approach in late mild cognitive impairment. *Bioinformatics*, 36(8):2561–2568, 2020.

[24] S. Bickel and T. Scheffer. Multi-view Clustering. In *Proceedings of the Fourth IEEE International Conference on Data Mining*, pages 19–26, 2004.

[25] F. Biessmann, F. C. Meinecke, A. Gretton, A. Rauch, G. Rainer, N. K. Logothetis, and K. R. Muller. Temporal Kernel CCA and its Application in Multimodal Neuronal Data Analysis. *Machine Learning*, 79(1-2):5–27, 2010.

[26] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, New York, 2006. ISBN: 978-0-387-31073-2.

[27] Matthew B Blaschko and Christoph H Lampert. Correlational Spectral Clustering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[28] Mia E-L Blomqvist, Niels Andreasen, Nenad Bogdanovic, Kaj Blennow, Anthony J Brookes, and Jonathan A Prince. Genetic variation in CTNNA3 encoding alpha-3 catenin and Alzheimer's disease. *Neuroscience letters*, 358(3):220–222, 2004.

[29] A. Blum and T. Mitchell. Combining Labeled and Unlabeled Data with Co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 92–100, 1998.

[30] Edwin V Bonilla, Kian Chai, and Christopher Williams. Multi-task Gaussian process prediction. *Advances in neural information processing systems*, 20, 2007.

[31] David Boutte and Jingyu Liu. Sparse Canonical Correlation Analysis Applied to fMRI and Genetic Data Fusion. In *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine*, pages 422–426, 2010.

[32] Ulf Brefeld and Tobias Scheffer. Co-EM Support Vector Learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 16, 2004.

[33] Matthew Brett, Jean-Luc Anton, et al. Region of Interest Analysis Using An SPM Toolbox. In *Proceedings of 8th International Conference on Functional Mapping of the Human Brain*, volume 16, page 497. Sendai, 2002.

[34] Randy L Buckner, Jessica R Andrews-Hanna, et al. The Brain's Default Network: Anatomy, Function, and Relevance to Disease. *Annals of the New York Academy of Sciences*, 2008.

[35] Randy L Buckner, Abraham Z Snyder, et al. Molecular, Structural, and Functional Characterization of Alzheimer's Disease: Evidence for a Relationship Between Default Activity, Amyloid, and Memory. *Journal of Neuroscience*, 25(34):7709–7717, 2005.

[36] Rita Cacace, Bavo Heeman, et al. Loss of DPP6 in Neurodegenerative Dementia: A Genetic Player in the Dysfunction of Neuronal Excitability. *Acta Neuropathologica*, 137(6):901–918, 2019.

[37] Xiao Cai, Feiping Nie, Weidong Cai, and Heng Huang. Heterogeneous image features integration via multi-modal semi-supervised learning model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1737–1744, 2013.

[38] Antti Cajanus, Eino Solje, Koikkalainen, et al. The Association Between Distinct Frontal Brain Volumes and Behavioral Symptoms in Mild Cognitive Impairment, Alzheimer's disease, and Frontotemporal Dementia. *Frontiers in Neurology*, 10:1059, 2019.

[39] Rich Caruana. Multitask Learning. *Machine Learning*, 28:41–75, 1997.

[40] Donato Cascio, Vincenzo Taormina, and Giuseppe Raso. Deep CNN for IIF Images Classification in Autoimmune Diagnostics. *Applied Sciences*, 9(8):1618, 2019.

[41] Nicoletta Cera, Roberto Esposito, et al. Altered Cingulate Cortex Functional Connectivity in Normal Aging and Mild Cognitive Impairment. *Frontiers in Neuroscience*, 13:857, 2019.

[42] Kamalika Chaudhuri, Sham M Kakade, Karen Livescu, and Karthik Sridharan. Multi-view Clustering via Canonical Correlation Analysis. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 129–136, 2009.

[43] Jia Chen, Gang Wang, and Georgios B Giannakis. Graph multiview canonical correlation analysis. *IEEE Transactions on Signal Processing*, 67(11):2826–2838, 2019.

[44] Jia Chen, Gang Wang, Yanning Shen, and Georgios B Giannakis. Canonical correlation analysis of datasets with a common source graph. *IEEE Transactions on Signal Processing*, 66(16):4398–4408, 2018.

[45] Jianhui Chen, Lei Tang, Jun Liu, and Jieping Ye. A Convex Formulation for Learning Shared Structures from Multiple Tasks. In *Proceedings of the 26th annual international conference on machine learning*, pages 137–144, 2009.

[46] Jingchun Chen, Natalia G Samoylova, and Melika G Cummings. TOMM40, A Risk Gene for Alzheimer's Disease, is Upregulated During Proinflammatory Response. *Alzheimer's & Dementia*, 17:e058711, 2021.

[47] Y. Chen, S. Wang, C. Peng, Z. Hua, and Y. Zhou. Generalized Nonconvex Low-Rank Tensor Approximation for Multi-View Subspace Clustering. *IEEE Transactions on Image Processing*, 30:4022–4035, 2021.

[48] Yi-Chun Chen, Shih-Cheng Chang, Yun-Shien Lee, Wei-Min Ho, Yu-Hua Huang, Yah-Yuan Wu, Yi-Chuan Chu, Kuan-Hsuan Wu, Li-Shan Wei, Hung-Li Wang, et al. TOMM40 Genetic Variants Cause Neuroinflammation in Alzheimer's Disease. *International Journal of Molecular Sciences*, 24(4):4085, 2023.

[49] Shing Wan Choi, Timothy Shin-Heng Mak, and Paul F O'Reilly. Tutorial: A Guide to Performing Polygenic Risk Score Analyses. *Nature Protocols*, 15(9):2759–2772, 2020.

[50] Hyonho Chun and Sunduz Keles. Sparse Partial Least Squares Regression for Simultaneous Dimension Reduction and Variable Selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25, 2010.

[51] A. Damianou, C. Ek, M. Titsias, and N. Lawrence. Manifold Relevance Determination. In *Proceedings of International Conference on Machine Learning*, pages 145–152, 2012.

[52] Patrick Danaher, Pei Wang, and Daniela M Witten. The Joint Graphical Lasso for Inverse Covariance Estimation Across Multiple Classes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(2):373–397, 2014.

[53] J. P. Van de Geer. Linear Relations Among k Sets of Variables. *Psychometrika*, 49(1):79–94, 1984.

[54] Laura W de Jong, Karin van der Hiele, Ilya M Veer, JJ Houwing, RGJ Westendorp, ELEM Bollen, Paul W de Bruin, HAM Middelkoop, Mark A van Buchem, and Jeroen van der Grond. Strongly Reduced Volumes of Putamen and Thalamus in Alzheimer's Disease: An MRI Study. *Brain*, 131(12):3277–3285, 2008.

[55] Ofer Dekel, Philip M Long, and Yoram Singer. Online Learning of Multiple Tasks with a Shared Loss. *Journal of Machine Learning Research*, 8(10), 2007.

[56] Dolores Del Prete, Frederic Checler, and Mounia Chami. Ryanodine Receptors: Physiological Function and Deregulation in Alzheimer Disease. *Molecular Neurodegeneration*, 9:1–15, 2014.

[57] F. Deleus and M. M. V. Hulle. A Connectivity-Based Method for Defining Regions-of-Interest in fMRI Data. *IEEE Transactions on Image Processing*, 18(8):1760–1771, 2009.

[58] Tom Diethe, David R Hardoon, and John Shawe-Taylor. Multiview Fisher Discriminant Analysis. In *Proceedings of NIPS Workshop on Learning from Multiple Sources*, 2008.

[59] Daisy Yi Ding, Shuangning Li, Balasubramanian Narasimhan, and Robert Tibshirani. Cooperative learning for multiview analysis. *Proceedings of the National Academy of Sciences*, 119(38):e2202113119, 2022.

[60] J. Donahue, P. Krahenbuhl, and T. Darrell. Adversarial Feature Learning. In *Proceedings of the 5th International Conference on Learning Representations*, 2017.

[61] Yeqing Dong, Tongxin Li, Zhonghui Ma, Chi Zhou, Xinxu Wang, Jie Li, et al. HSPA1A, HSPA2, and HSPA8 are Potential Molecular Biomarkers for Prognosis among HSP70 Family in Alzheimer's Disease. *Disease Markers*, 2022, 2022.

[62] C. Du, C. Du, G. Long, X. Jin, and Y. Li. Efficient Bayesian Maximum Margin Multiple Kernel Learning. In *Machine Learning and Knowledge Discovery in Databases*, pages 165–181, 2016.

[63] Lei Du, Fang Liu, et al. Identifying Diagnosis-Specific Genotype-Phenotype Associations via Joint Multitask Sparse Canonical Correlation Analysis and Classification. *Bioinformatics*, 36(Supplement_1):i371–i379, 2020.

[64] Lei Du, Fang Liu, Kefei Liu, Xiaohui Yao, Shannon L Risacher, Junwei Han, Andrew J Saykin, and Li Shen. Associating Multi-Modal Brain Imaging Phenotypes and Genetic Risk Factors via a Dirty Multi-Task Learning Method. *IEEE Transactions on Medical Imaging*, 39(11):3416–3428, 2020.

[65] Lei Du, Kefei Liu, et al. Detecting Genetic Associations with Brain Imaging Phenotypes in Alzheimer's Disease via a Novel Structured SCCA Approach. *Medical Image Analysis*, 61:101656, 2020.

[66] Lei Du, Kefei Liu, et al. Multi-Task Sparse Canonical Correlation Analysis with Application to Multi-Modal Brain Imaging Genetics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(1):227–239, 2021.

[67] Lei Du, Kefei Liu, Tuo Zhang, Xiaohui Yao, Jingwen Yan, Shannon L Risacher, Junwei Han, Lei Guo, Andrew J Saykin, Li Shen, et al. A Novel SCCA Approach via Truncated l1-norm and Truncated Group Lasso for Brain Imaging Genetics. *Bioinformatics*, 34(2):278–285, 2018.

[68] Lloyd T Elliott, Kevin Sharp, et al. Genome-wide association studies of brain imaging phenotypes in uk biobank. *Nature*, 562(7726):210–216, 2018.

[69] Sam Ereira, Sheena Waters, Adeel Razi, and Charles R Marshall. Early detection of dementia with default-mode network effective connectivity. *Nature Mental Health*, pages 1–14, 2024.

[70] R. G. W. Verhaak et al. Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, 17(1):98–110, 2010.

[71] Theodoros Evgeniou and Massimiliano Pontil. Regularized multi-task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117, 2004.

[72] Q. Fan, Z. Wang, H. Zha, and D. Gao. MREKLM: A Fast Multiple Empirical Kernel Learning Machine. *Pattern Recognition*, 61:197–209, 2017.

[73] Wenqi Fan, Yao Ma, Han Xu, Xiaorui Liu, Jianping Wang, Qing Li, and Jiliang Tang. Deep adversarial canonical correlation analysis. In *Proceedings of the SIAM International Conference on Data Mining*, pages 352–360. SIAM, 2020.

[74] Jian Fang, Dongdong Lin, et al. Joint Sparse Canonical Correlation Analysis for Detecting Differential Imaging Genetics Modules. *Bioinformatics*, 32(22):3480–3488, 2016.

[75] Jacques Ferlay, Isabelle Soerjomataram, et al. Cancer Incidence and Mortality Worldwide: Sources, Methods and Major Patterns in GLOBOCAN 2012. *International Journal of Cancer*, 136(5):E359–E386, 2015.

[76] Ronald A Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[77] Hui Gao, Yu Tao, et al. Functional Enrichment Analysis of Three Alzheimer's Disease Genome-Wide Association Studies Identities DAB1 as a Novel Candidate Liability/Protective Gene. *Biochemical and Biophysical Research Communications*, 463(4):490–495, 2015.

[78] Lei Gao, Lin Qi, Enqing Chen, and Ling Guan. Discriminative Multiple Canonical Correlation Analysis for Multi-feature Information Fusion. In *Proceedings of IEEE International Symposium on Multimedia*, pages 36–43. IEEE, 2012.

[79] Quanxue Gao, Huanhuan Lian, Qianqian Wang, and Gan Sun. Cross-modal subspace clustering via deep canonical correlation analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3938–3945, 2020.

[80] A. Golugula, G. Lee, S. R. Master, M. D. Feldman, J. E. Tomaszewski, D. W. Speicher, and A. Madabhushi. Supervised Regularized Canonical Correlation Analysis: Integrating Histologic and Proteomic Measurements for Predicting Biochemical Recurrence Following Prostate Surgery. *BMC Bioinformatics*, 12(483), 2011.

[81] Abhishek Golugula, George Lee, et al. Supervised Regularized Canonical Correlation Analysis: Integrating Histologic and Proteomic Measurements for Predicting Biochemical Recurrence Following Prostate Surgery. *BMC Bioinformatics*, 12(1):1–13, 2011.

[82] Shaoqing Gong, Brenda Bin Su, Hugo Tovar, ChunXiang Mao, Valeria Gonzalez, Ying Liu, Yongke Lu, Ke-Sheng Wang, and Chun Xu. Polymorphisms within RYR3 gene are associated with risk and age at onset of hypertension, diabetes, and Alzheimer's disease. *American Journal of Hypertension*.

[83] Claudia Grellmann, Sebastian Bitzer, Jane Neumann, Lars T. Westlye, Ole A. Andreassen, Arno Villringer, and Annette Horstmann. Comparison of variants of canonical correlation analysis and partial least squares for combined analysis of MRI and genetic data. *NeuroImage*, 107:289–310, 2015.

[84] Matthias Grieder, Danny JJ Wang, Thomas Dierks, Lars-Olof Wahlund, and Kay Jann. Default mode network complexity and cognitive decline in mild Alzheimer's disease. *Frontiers in neuroscience*, 12:770, 2018.

[85] Jane Grimwood, Laurie A Gordon, Anne Olsen, Astrid Terry, Jeremy Schmutz, Jane Lamerdin, Uffe Hellsten, David Goodstein, Olivier Couronne, Mary Tran-Gyamfi, et al. The DNA Sequence and Biology of Human Chromosome 19. *Nature*, 428(6982):529–535, 2004.

[86] Samuel M Gross and Robert Tibshirani. Collaborative regression. *Biostatistics*, 16(2):326–338, 2015.

[87] Rita J Guerreiro and John Hardy. TOMM40 Association with Alzheimer Disease: Tales of APOE and Linkage Disequilibrium. *Archives of Neurology*, 69(10):1243–1244, 2012.

[88] Brendan Hallam, Justin Chan, Sergi Gonzalez Costafreda, Rohan Bhome, and Jonathan Huntley. What are the neural correlates of meta-cognition and anosognosia in alzheimer's disease? a systematic review. *Neurobiology of Aging*, 94:250–264, 2020.

[89] Lu Han, Xiao-Yuan Jing, and Fei Wu. Multi-view Local Discrimination and Canonical Correlation Analysis for Image Classification. *Neurocomputing*, 275:1087–1098, 2018.

[90] Y. Han, K. Yang, Y. Ma, and G. Liu. Localized Multiple Kernel Learning Via Sample-Wise Alternating Optimization. *IEEE Transactions on Cybernetics*, 44(1):137–148, 2014.

[91] M. Hanafi. PLS Path Modelling: Computation of Latent Variables with the Estimation Mode B. *Computational Statistics*, 22(2):275–292, 2007.

[92] M. Hanafi and H. A. L. Kiers. Analysis of K sets of Data, with Differential Emphasis on Agreement Between and Within Sets. *Computational Statistics and Data Analysis*, 51(3):1491–1508, 2006.

[93] Xiaoke Hao, Chanxiu Li, Lei Du, Xiaohui Yao, Jingwen Yan, Shannon L Risacher, Andrew J Saykin, Li Shen, and Daoqiang Zhang. Mining Outcome-Relevant Brain Imaging Genetic Associations via Three-Way Sparse Canonical Correlation Analysis in Alzheimer's Disease. *Scientific Reports*, 7(1):44272, 2017.

[94] Xiaoke Hao, Qihao Tan, Yingchun Guo, Yunjia Xiao, Ming Yu, Meiling Wang, Jing Qin, Daoqiang Zhang, and Alzheimer's Disease Neuroimaging Initiative. Identifying Modality-Consistent and Modality-Specific Features via Label-Guided Multi-Task Sparse Canonical Correlation Analysis for Neuroimaging Genetics. *IEEE Transactions on Biomedical Engineering*, 70(3):831–840, 2023.

[95] Xiaoke Hao, Xiaohui Yao, Shannon L Risacher, Andrew J Saykin, Jintai Yu, Huifu Wang, Lan Tan, Li Shen, and Daoqiang Zhang. Identifying Candidate Genetic Associations with MRI-derived AD-related ROI via Tree-Guided Sparse Learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(6):1986–1996, 2018.

[96] Ahmad R Hariri, Emily M Drabant, and Daniel R Weinberger. Imaging Genetics: Perspectives from Studies of Genetically Driven Variation in Serotonin Function and Corticolimbic Affective Processing. *Biological Psychiatry*, 59(10):888–897, 2006.

[97] K. Hassani and A. H. Khasahmadi. Contrastive Multi-View Representation Learning on Graphs. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[98] Jingrui He and Rick Lawrence. A graphbased framework for multi-task multi-view learning. In *ICML*, pages 25–32, 2011.

[99] Derrek P Hibar, Jason L Stein, et al. Voxelwise Gene-Wide Association Study (vGeneWAS): Multivariate Gene-Based Association Testing in 731 Elderly Subjects. *Neuroimage*, 56(4):1875–1891, 2011.

[100] Akitoyo Hishimoto, Olga Pletnikova, Doyle Lu Lang, Juan C Troncoso, Josephine M Egan, and Qing-Rong Liu. Neurexin 3 transmembrane and soluble isoform expression and splicing haplotype are associated with neuron inflammasome and Alzheimer's disease. *Alzheimer's research & therapy*, 11:1–15, 2019.

[101] Toby Dylan Hocking, Armand Joulin, et al. Clusterpath an algorithm for clustering using convex fusion penalties. In *28th international conference on machine learning*, page 1, 2011.

[102] Holger Hoefling. A Path Algorithm for the Fused Lasso Signal Approximator. *Journal of Computational and Graphical Statistics*, 19(4):984–1006, 2010.

[103] C. Hong, J. Yu, J. Wan, D. Tao, and M. Wang. Multimodal Deep Autoencoder for Human Pose Recovery. *IEEE Transactions on Image Processing*, 24(12):5659–5670, 2015.

[104] Zhibin Hong, Xue Mei, Danil Prokhorov, and Dacheng Tao. Tracking via Robust Multi-Task Multi-View Joint Sparse Representation. In *Proceedings of the IEEE international conference on computer vision*, pages 649–656, 2013.

[105] Paul Horst. Generalized Canonical Correlations and Their Applications to Experimental Data. *Journal of Clinical Psychology*, 17(4):331–347, 1961.

[106] Paul Horst. Relations Among m Sets of Measures. *Psychometrika*, 26(2):129–149, 1961.

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

[108] Che-Yu Hsu, Kuan-Ting Lee, Tzu-Yu Sun, Chun-I Sze, Shenq-Shyang Huang, Li-Jin Hsu, and Nan-Shan Chang. WWOX and its binding proteins in neurodegeneration. *Cells*, 10(7):1781, 2021.

[109] W. Hu, D. Lin, S. Cao, J. Liu, J. Chen, V. D. Calhoun, and Y. Wang. Adaptive Sparse Multiple Canonical Correlation Analysis With Application to Imaging (Epi)Genomics Study of Schizophrenia. *IEEE Transactions on Biomedical Engineering*, 65(2):390–399, 2018.

[110] Wenxing Hu, Dongdong Lin, Shaolong Cao, Jingyu Liu, Jiayu Chen, Vince D Calhoun, and Yu-Ping Wang. Adaptive Sparse Multiple Canonical Correlation Analysis with Application to Imaging (epi) Genomics Study of Schizophrenia. *IEEE Transactions on Biomedical Engineering*, 65(2):390–399, 2017.

[111] Z. Hu et al. The Molecular Portraits of Breast Tumors are Conserved Across Microarray Platforms. *BMC Genomics*, 7:96, April 2006.

[112] Meiyan Huang, Thomas Nichols, et al. FVGWAS: Fast Voxelwise Genome Wide Association Analysis of Large-Scale Imaging Genetic Data. *Neuroimage*, 118:613–627, 2015.

[113] Shudong Huang, Ivor Tsang, Zenglin Xu, Jiancheng Lv, and Quan-Hui Liu. Multi-view clustering on topological manifold. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6944–6951, 2022.

[114] Z. Huang, J. T. Zhou, H. Zhu, C. Zhang, J. Lv, and X. Peng. Deep Spectral Representation Learning From Multi-View Data. *IEEE Transactions on Image Processing*, 30:5352–5362, 2021.

[115] Aoife Hunt, Peter Schonknecht, Marcus Henze, Ulrich Seidl, Uwe Haberkorn, and Johannes Schroder. Reduced Cerebral Glucose Metabolism in Patients at Risk for Alzheimer's Disease. *Psychiatry Research: Neuroimaging*, 155(2):147–154, 2007.

[116] Xiaodong Jia, Xiao-Yuan Jing, et al. Semi-Supervised Multi-View Deep Discriminant Representation Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7):2496–2509, 2020.

[117] Yuanbing Jiang, Xiaopu Zhou, Fanny C Ip, Philip Chan, Yu Chen, Nicole CH Lai, Kit Cheung, Ronnie MN Lo, Estella PS Tong, Bonnie WY Wong, et al. Large-scale plasma proteomic profiling identifies a high-performance biomarker panel for Alzheimer's disease screening and staging. *Alzheimer's & Dementia*, 18(1):88–102, 2022.

[118] Zhibin Jiang, Zekang Bian, and Shitong Wang. Multi-view Local Linear KNN Classification: Theoretical and Experimental Studies on Image Classification. *International Journal of Machine Learning and Cybernetics*, 11:525–543, 2020.

[119] Xin Jin, Fuzhen Zhuang, Shuhui Wang, Qing He, and Zhongzhi Shi. Shared Structure Learning for Multiple Tasks with Multiple Views. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part II 13*, pages 353–368. Springer, 2013.

[120] Xin Jin, Fuzhen Zhuang, Hui Xiong, Changying Du, Ping Luo, and Qing He. Multi-Task Multi-View Learning for Heterogeneous Tasks. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pages 441–450, 2014.

[121] Zhong Jin, Jing-Yu Yang, Zhong-Shan Hu, and Zhen Lou. Face Recognition Based on the Uncorrelated Discriminant Transformation. *Pattern Recognition*, 34(7):1405–1416, 2001.

[122] Jiaojiao Jing, Feng Zhang, Li Zhao, Jinghui Xie, Jianwen Chen, Rujia Zhong, Yanjun Zhang, and Chunbo Dong. Correlation Between Brain 18F-AV45 and 18F-FDG PET Distribution Characteristics and Cognitive Function in Patients with Mild and Moderate Alzheimer's Disease. *Journal of Alzheimer's Disease*, 79(3):1317–1325, 2021.

[123] Sham M Kakade and Dean P Foster. Multi-view Regression via Canonical Correlation Analysis. In *Proceedings of International Conference on Computational Learning Theory*, pages 82–96. Springer, 2007.

[124] Meina Kan, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. Multi-view Discriminant Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):188–194, 2015.

[125] M. Kanai, R. Togo, T. Ogawa, and M. Haseyama. Aesthetic Quality Assessment of Images Via Supervised Locality Preserving CCA. In *Proceedings of the IEEE 6th Global Conference on Consumer Electronics*, pages 1–2, 2017.

[126] Takashi Kato, Yoshitaka Inui, Akinori Nakamura, and Kengo Ito. Brain Fluorodeoxyglucose (FDG) PET in Dementia. *Ageing Research Reviews*, 30:73–84, 2016.

[127] Katzorke, Andrea and Zeller, Julia BM and Muller, Laura D and Lauer, Martin and Polak, Thomas and Reif, Andreas and Deckert, Jurgen and Herrmann, Martin J. Reduced activity in the right inferior frontal gyrus in elderly APOE-E4 carriers during a verbal fluency task. *Frontiers in human neuroscience*, 11:46, 2017.

[128] Fengchun Ke, Wei Kong, et al. Identifying Imaging Genetics Biomarkers of Alzheimer's Disease by Multi-Task Sparse Canonical Correlation Analysis and Regression. *Frontiers in Genetics*, 12, 2021.

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

[130] M. R. Khan and J. E. Blumenstock. Multi-GCN: Graph Convolutional Networks for Multi-View Networks, with Applications to Global Poverty. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[131] M. Kim, J. H. Won, J. Youn, and H. Park. Joint-Connectivity-Based Sparse Canonical Correlation Analysis of Imaging Genetics for Detecting Biomarkers of Parkinson's Disease. *IEEE Transactions on Medical Imaging*, 39(1):23–34, 2020.

[132] Mansu Kim, Ji Hye Won, et al. Joint-Connectivity-Based Sparse Canonical Correlation Analysis of Imaging Genetics for Detecting Biomarkers of Parkinson's Disease. *IEEE Transactions on Medical Imaging*, 39(1):23–34, 2019.

[133] A. Klami, S. Virtanen, E. Leppaaho, and S. Kaski. Group Factor Analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 26(9):2136–2147, 2015.

[134] M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien. Lp-Norm Multiple Kernel Learning. *Journal of Machine Learning Research*, 12:953–997, 2011.

[135] DCFR Koboldt, Robert Fulton, Michael McLellan, Heather Schmidt, Joelle Kalicki-Veizer, Joshua McMichael, Lucinda Fulton, David Dooling, Li Ding, Elaine Mardis, et al. Comprehensive Molecular Portraits of Human Breast Tumours. *Nature*, 490(7418):61–70, 2012.

[136] Omid Kohannim, Derrek P Hibar, et al. Discovery and Replication of Gene Influences on Brain Structure Using LASSO Regression. *Frontiers in Neuroscience*, 6:115, 2012.

[137] A. Krishnan, L. J. Williams, A. R. McIntosh, and H. Abdi. Partial Least Squares (PLS) Methods for Neuroimaging: A Tutorial and Review. *NeuroImage*, 56:455–475, 2010.

[138] Alexander M Kulminski, Jian Huang, Jiayi Wang, Liang He, Yury Loika, and Irina Culminskaya. Apolipoprotein E Region Molecular Signatures of Alzheimer's Disease. *Aging Cell*, 17(4):e12779, 2018.

[139] A. Kumar, A. Niculescu-mizil, K. Kavukcoglu, and H. Daume. A Binary Classification Framework for Two-Stage Multiple Kernel Learning. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

[140] A. Kumar, P. Rai, and H. Daume III. Co-regularized Spectral Clustering with Multiple Kernels. 2010.

[141] A. Kumar, P. Rai, and H. Daume III. Co-regularized Multi-view Spectral Clustering. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

[142] Man Ki Kwok, Shi Lin Lin, and C Mary Schooling. Re-Thinking Alzheimer's Disease Therapeutic Targets Using Gene-Based Tests. *EBioMedicine*, 37:461–470, 2018.

[143] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the Kernel Matrix with Semidefinite Programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

[144] Magalie S Leduc, Marianne Mcguire, Suneeta Madan-Khetarpal, Damara Ortiz, Susan Hayflick, Kory Keller, Christine M Eng, Yaping Yang, and Weimin Bi. De novo apparent loss-of-function mutations in PRR12 in three patients with intellectual disability and iris abnormalities. *Human Genetics*, 137:257–264, 2018.

[145] George Lee, Asha Singanamalli, et al. Supervised multi-view canonical correlation analysis (smvcca): Integrating histologic and proteomic features for predicting recurrent prostate cancer. *IEEE transactions on medical imaging*, 34(1):284–297, 2014.

[146] Robert Leech and David J Sharp. The role of the posterior cingulate cortex in cognition and disease. *Brain*, 137(1):12–32, 2014.

[147] Hui Li, Xuejun Liao, and Lawrence Carin. Multi-task Reinforcement Learning in Partially Observable Stochastic Environments.s. *Journal of Machine Learning Research*, 10(5), 2009.

[148] M. Li, Y. Liu, G. Feng, Z. Zhou, and D. Hu. OI and fMRI Signal Separation Using Both Temporal and Spatial Autocorrelations. *IEEE Transactions on Biomedical Engineering*, 57(8):1917–1926, 2010.

[149] Rui Li, Xia Wu, et al. Attention-Related Networks in Alzheimer's Disease: A Resting Functional MRI Study. *Human Brain Mapping*, 33(5):1076–1088, 2012.

[150] Xuelong Li, Han Zhang, Rong Wang, and Feiping Nie. Multiview Clustering: A Scalable and Parameter-Free Bipartite Graph Fusion Method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):330–344, 2020.

[151] Y. Li, M. Yang, and Z. Zhang. A Survey of Multi-View Representation Learning. *IEEE Transactions on Knowledge and Data Engineering*, 31(10):1863–1883, 2019.

[152] Naiyao Liang, Zuyuan Yang, Zhenni Li, and Shengli Xie. Co-Consensus Semi-Supervised Multi-View Learning with Orthogonal Non-Negative Matrix Factorization. *Information Processing & Management*, 59(5):103054, 2022.

[153] Naiyao Liang, Zuyuan Yang, Zhenni Li, Shengli Xie, and Chun-Yi Su. Semi-Supervised Multi-View Clustering with Graph-Regularized Partially Shared Non-Negative Matrix Factorization. *Knowledge-Based Systems*, 190:105185, 2020.

[154] Zhizheng Liang and Pengfei Shi. Uncorrelated Discriminant Vectors Using a Kernel Method. *Pattern Recognition*, 38(2):307–310, 2005.

[155] Dongdong Lin, Vince D Calhoun, and Yu-Ping Wang. Correspondence between fMRI and SNP Data by Group Sparse Canonical Correlation Analysis. *Medical Image Analysis*, 18(6):891–902, 2014.

[156] F-F Liu, Z Zhang, W Chen, H-Y Gu, and Q-J Yan. Regulatory mechanism of microRNA-377 on CDH13 expression in the cell model of Alzheimer's disease. *European Review for Medical & Pharmacological Sciences*, 22(9), 2018.

[157] Qiuhua Liu, Xuejun Liao, and Lawrence Carin. Semi-Supervised Multitask Learning. *Advances in Neural Information Processing Systems*, 20, 2007.

[158] X. Liu, L. Wang, J. Zhang, and J. Yin. Sample-Adaptive Multiple Kernel Learning. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 1975–1981, 2014.

[159] Xiaozheng Liu, Wei Chen, Hongtao Hou, Xingli Chen, Jiangtao Zhang, Jian Liu, Zhongwei Guo, and Guanghui Bai. Decreased Functional Connectivity Between the Dorsal Anterior Cingulate Cortex and Lingual Gyrus in Alzheimer's Disease Patients with Depression. *Behavioural Brain Research*, 326:132–138, 2017.

[160] JCL Looi, Leif Svensson, et al. Putaminal Volume in Frontotemporal Lobar Degeneration and Alzheimer Disease: Differential Volumes in Dementia Subtypes and Controls. *American Journal of Neuroradiology*, 30(8):1552–1560, 2009.

[161] Y. Luo, D. Tao, K. Ramamohanarao, C. Xu, and Y. Wen. Tensor Canonical Correlation Analysis for Multi-View Dimension Reduction. *IEEE Transactions on Knowledge and Data Engineering*, 27(11):3111–3124, 2015.

[162] P. Maji and A. Mandal. Multimodal Omics Data Integration Using Max Relevance-Max Significance Criterion. *IEEE Transactions on Biomedical Engineering*, 64(8):1841–1851, 2017.

[163] A. Mandal and P. Maji. FaRoC: Fast and Robust Supervised Canonical Correlation Analysis for Multimodal Omics Data. *IEEE Transactions on Cybernetics*, 48(4):1229–1241, 2018.

[164] Ankita Mandal and Pradipta Maji. Multiview Regularized Discriminant Canonical Correlation Analysis: Sequential Extraction of Relevant Features From Multiblock Data. *IEEE Transactions on Cybernetics*, 53(9):5497–5509, 2023.

[165] R. Mane, N. Robinson, A. P. Vinod, S.-W. Lee, and C. Guan. A Multi-View CNN with Novel Variance Layer for Motor Imagery Brain Computer Interface. In *Proceedings of the 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2950–2953, 2020.

[166] I. Mármol, C. Sánchez de Diego, A. P. Dieste, E. Cerrada, and M. J. R. Yoldi. Colorectal Carcinoma: A General Overview and Future Perspectives in Colorectal Cancer. *International Journal of Molecular Sciences*, 18(1):197, 2017.

[167] Almudena Medina-Samame, Eva Paller, Mateo R Bril, Ana Archvadze, Madalena BC Simoes-Abade, Patricia Estanol-Cayuela, and Chloe LeMaoult. Role of neurexins in Alzheimer's disease. *Journal of Neuroscience*, 43(23):4194–4196, 2023.

[168] Yanying Mei, Zhenwen Ren, Bin Wu, Yanhua Shao, and Tao Yang. Robust Graph-Based Multi-View Clustering in Latent Embedding Space. *International Journal of Machine Learning and Cybernetics*, pages 1–12, 2022.

[169] Charles Micchelli and Massimiliano Pontil. Kernels for Multi-task Learning. *Advances in neural information processing systems*, 17, 2004.

[170] Jeremy A Miller, Randall L Woltjer, et al. Genes and Pathways Underlying Regional and Cell Type Changes in Alzheimer's Disease. *Genome Medicine*, 5(5):1–17, 2013.

[171] Akinori Miyashita, Hiroyuki Arai, et al. Genetic Association of CTNNA3 with Late-Onset Alzheimer's Disease in Females. *Human Molecular Genetics*, 16(23):2854–2869, 2007.

[172] Kiyohito Mizutani, Muneaki Miyata, Hajime Shiotani, Takeshi Kameyama, and Yoshimi Takai. NECTIN-2 in general and in the brain. *Molecular and Cellular Biochemistry*, 477(1):167–180, 2022.

[173] Q. Mo and R. Shen. iClusterPlus: Integrative Clustering of Multiple Genomic Data Sets. *R package version 1.19.0*, 2018.

[174] A. Mohammadi-Nejad, G. Hossein-Zadeh, and H. Soltanian-Zadeh. Structured and Sparse Canonical Correlation Analysis as a Brain-Wide Multi-Modal Data Fusion Approach. *IEEE Transactions on Medical Imaging*, 36(7):1438–1448, 2017.

[175] A. R. Mohammadi-Nejad, G. A. Hossein-Zadeh, and H. Soltanian-Zadeh. Structured and Sparse Canonical Correlation Analysis as a Brain-Wide Multi-Modal Data Fusion Approach. *IEEE Transactions on Medical Imaging*, 36(7):1438–1448, 2017.

[176] Sankar Mondal and Pradipta Maji. Class-Structure Preserving Multi-View Correlated Discriminant Analysis for Multiblock Data. *International Journal of Machine Learning and Cybernetics*, pages 1–22, 2024, DOI: 10.1007/s13042-024-02270-9.

[177] Sankar Mondal and Pradipta Maji. Multi-task learning and sparse discriminant canonical correlation analysis for identification of diagnosis-specific genotype-phenotype association. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 1–13, 2024, DOI: 10.1109/TCBB.2024.3386406.

[178] Sankar Mondal and Pradipta Maji. Multi-view multi-task sparse canonical correlation analysis for imaging genetics study. *IEEE Transactions on Emerging Topics in Computational Intelligence*, pages 1–10, 2024 (Manuscript ID: TETCI-2024-1533, Submitted).

[179] Sankar Mondal and Pradipta Maji. Correlated Joint Subspace Learning: A Framework for Class-Structure Preservation. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–14, 2024 (Manuscript ID: TKDE-2024-12-3185, Submitted).

[180] Janusz Morys, Matthew Bobinski, et al. Alzheimer's Disease Severely Affects Areas of the Claustrum Connected with the Entorhinal Cortex. *Journal fur Hirnforschung*, 37(2):173–180, 1996.

[181] Lisa Mosconi. Brain Glucose Metabolism in the Early and Specific Diagnosis of Alzheimer's Disease: FDG-PET Studies in MCI and AD. *European Journal of Nuclear Medicine and Molecular Imaging*, 32:486–510, 2005.

[182] T Muller, C Loosse, et al. The AICD Interacting Protein DAB1 is Up-regulated in Alzheimer Frontal Cortex Brain Samples and Causes Deregulation of Proteins involved in Gene Expression Changes. *Current Alzheimer Research*, 8(5):573–582, 2011.

[183] I. Muslea, S. Minton, and C. A. Knoblock. Active + Semi-supervised Learning = Robust Multi-View Learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 435–442, 2002.

[184] I. Muslea, S. Minton, and C. A. Knoblock. Active Learning with Strong and Weak Views: A Case Study on Wrapper Induction. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, volume 18, pages 415–420, 2003.

[185] I. Muslea, S. Minton, and C. A. Knoblock. Active Learning with Multiple Views. *Journal of Artificial Intelligence Research*, 27(1):203–233, 2006.

[186] Farouk S Nathoo, Linglong Kong, Hongtu Zhu, and Alzheimer's Disease Neuroimaging Initiative. A Review of Statistical Methods in Imaging Genetics. *Canadian Journal of Statistics*, 47(1):108–131, 2019.

[187] A. Nazarpour and P. Adibi. Two-Stage Multiple Kernel Learning for Supervised Dimensionality Reduction. *Pattern Recognition*, 48(5):1854–1862, 2015.

[188] Cancer Genome Atlas Research Network et al. Integrated genomic and molecular characterization of cervical cancer. *Nature*, 543(7645):378, 2017.

[189] TCGA Research Network. Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *The New England Journal of Medicine*, 372(26):2481–2498, 2015.

[190] Feiping Nie, Xiaowei Zhao, Rong Wang, and Xuelong Li. Fast locality discriminant analysis with adaptive manifold embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9315–9330, 2022.

[191] Allan Aasbjerg Nielsen. Multiset Canonical Correlations Analysis and Multispectral, Truly Multitemporal Remote Sensing Data. *IEEE Transactions on image Processing*, 11(3):293–305, 2002.

[192] K. Nigam and R. Ghani. Analyzing the Effectiveness and Applicability of Co-training. In *Proceedings of the Ninth International Conference on Information and Knowledge Management*, pages 86–93, 2000.

[193] Yue Pan, Ke-Sheng Wang, and Nagesh Aragam. NTM and NR3C2 Polymorphisms Influencing Intelligence: Family-Based Association Studies. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 35(1):154–160, 2011.

[194] Elena Parkhomenko, David Tritchler, et al. Sparse Canonical Correlation Analysis with Application to Genomic Data Integration. *Statistical Applications in Genetics and Molecular Biology*, 8(1), 2009.

[195] Lucas C Parra. Multi-set Canonical Correlation Analysis Simply Explained. *ArXiv Preprint arXiv:1802.03759*, 2018.

[196] Jinye Peng, Peng Luo, Ziyu Guan, and Jianping Fan. Graph-Regularized Multi-View Semantic Subspace Learning. *International Journal of Machine Learning and Cybernetics*, 10:879–895, 2019.

[197] K Persson, VD Bohbot, Nenad Bogdanovic, Geir Selbaek, Anne Braekhus, and Knut Engedal. Finding of Increased Caudate Nucleus in Patients with Alzheimer's Disease. *Acta Neurologica Scandinavica*, 137(2):224–232, 2018.

[198] Barbara B Pineda-Bautista et al. General Framework for Class-Specific Feature Selection. *Expert Systems with Applications*, 38(8):10018–10024, 2011.

[199] Florian Prive, Hugues Aschard, et al. Portability of 245 polygenic scores when derived from the uk biobank and applied to 9 ancestry groups from the same cohort. *The American Journal of Human Genetics*, 109(1):12–23, 2022.

[200] Dmitry Prokopenko et al. Region-Based Analysis of Rare Genomic Variants in Whole-Genome Sequencing Datasets Reveal Two Novel Alzheimer's Disease-Associated Genes: DTNB and DLG2. *MedRxiv*, 2021.

[201] Shaun Purcell, Benjamin Neale, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[202] M. A. Qadar and A. Seghouane. A Projection CCA Method for Effective fMRI Data Analysis. *IEEE Transactions on Biomedical Engineering*, 66(11):3247–3256, 2019.

[203] Huihui Qi, Yang Hu, Yingru Lv, and Peijun Wang. Primarily disrupted default subsystems cause impairments in inter-system interactions and a higher regulatory burden in Alzheimer's disease. *Frontiers in Aging Neuroscience*, 12:593648, 2020.

[204] Junyang Qian, Yosuke Tanigawa, et al. A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the uk biobank. *PLoS genetics*, 16(10):e1009141, 2020.

[205] Junyang Qian, Yosuke Tanigawa, Ruilin Li, Robert Tibshirani, Manuel A Rivas, and Trevor Hastie. Large-scale multivariate sparse regression with applications to uk biobank. *The annals of applied statistics*, 16(3):1891, 2022.

[206] Neha S Raghavan, Logan Dumitrescu, Elizabeth Mormino, Emily R Mahoney, Annie J Lee, Yizhe Gao, Murat Bilgel, David Goldstein, Theresa Harrison, Corinne D Engelman, et al. Association between common variants in RBFOX1, an RNA-binding protein, and brain amyloidosis in early and preclinical Alzheimer disease. *JAMA neurology*, 77(10):1288–1298, 2020.

[207] Marcus E Raichle, Ann Mary MacLeod, et al. A Default Mode of Brain Function. *Proceedings of the National Academy of Sciences*, 98(2):676–682, 2001.

[208] Aravind Lathika Rajendrakumar, Konstantin G Arbeev, Olivia Bagley, Anatoliy I Yashin, Svetlana Ukraintseva, and Alzheimer's Disease Neuroimaging Initiative. The SNP rs6859 in NECTIN2 gene is associated with underlying heterogeneous trajectories of cognitive changes in older adults. *BMC Neurology*, 24(1):78, 2024.

[209] N. Rappoport and R. Shamir. Multi-Omic and Multi-View Clustering Algorithms: Review and Cancer Benchmark. *Nucleic Acids Research*, 46(20):10546–10562, 2018.

[210] Roi Reichart, Katrin Tomanek, Udo Hahn, and Ari Rappoport. Multi-Task Active Learning for Linguistic Annotations. In *Proceedings of ACL-08: HLT*, pages 861–869, 2008.

[211] Peggy Reiner, Eric Jouvent, et al. Sulcal Span in Azheimer's Disease, Amnestic Mild Cognitive Impairment, and Healthy Controls. *Journal of Alzheimer's Disease*, 29(3):605–613, 2012.

[212] D. Reinsel, J. Gantz, and J. Rydning. The Digitization of the World From Edge to Core. Technical report, Data Age 2025, An IDC White Paper - #US44413318, Sponsored by Seagate, November 2018.

[213] T. Rohlfing, A. Pfefferbaum, E. V. Sullivan, and C. R. Maurer. Information Fusion in Biomedical Image Analysis: Combination of Data vs. Combination of Interpretations. In *Proceedings of the 19th International Conference on Information Processing in Medical Imaging*, pages 150–161, 2005.

[214] Rita R Romito-DiGiacomo, Harry Menegay, Samantha A Cicero, and Karl Herrup. Effects of Alzheimer's Disease on Different Cortical Layers: The Role of Intrinsic Differences in Aβ Susceptibility. *Journal of Neuroscience*, 27(32):8496–8504, 2007.

[215] J. Rupnik and J. Shawe-Taylor. Multi-View Canonical Correlation Analysis. 2010.

[216] Agnieszka Rybak-Wolf and Mireya Plass. RNA Dynamics in Alzheimer's disease. *Molecules*, 26(17):5113, 2021.

[217] Ruslan Salakhutdinov and Geoffrey Hinton. Deep Boltzmann Machines. In *Artificial Intelligence and Statistics*, pages 448–455. PMLR, 2009.

[218] A. Sano, W. Chen, D. Lopez-Martinez, S. Taylor, and R. W. Picard. Multimodal Ambulatory Sleep Detection Using LSTM Recurrent Neural Networks. *IEEE Journal of Biomedical and Health Informatics*, 23(4):1607–1617, 2019.

[219] Stephen W Scheff, Douglas A Price, et al. Synaptic Loss in the Inferior Temporal Gyrus in Mild Cognitive Impairment and Alzheimer's Disease. *Journal of Alzheimer's Disease*, 24(3):547–557, 2011.

[220] Stephen W Scheff, Douglas A Price, et al. Synaptic Change in the Posterior Cingulate Gyrus in the Progression of Alzheimer's Disease. *Journal of Alzheimer's disease*, 43(3):1073–1090, 2015.

[221] H. S. Seung and D. D. Lee. Cognition. The Manifold Ways of Perception. *Science*, 290(5500):2268–2269, 2000.

[222] Abhishek Sharma, Abhishek Kumar, Hal Daume, and David W Jacobs. Generalized Multiview Analysis: A discriminative Latent Space. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2160–2167. IEEE, 2012.

[223] S. S. Shiju and S. Sumitra. Multiple Kernel Learning using Single Stage Function Approximation for Binary Classification Problems. *International Journal of Systems Science*, 48(16):3569–3580, 2017.

[224] AKAHO Shotaro. A kernel method for canonical correlation analysis. In *International Meeting of Psychometric Society, 2001*, 2001.

[225] Daniel HS Silverman, Gary W Small, Carol Y Chang, Carolyn S Lu, Michelle A Kung de Aburto, Wei Chen, Johannes Czernin, Stanley I Rapoport, Pietro Pietrini, Gene E Alexander, et al. Positron Emission Tomography in Evaluation of Dementia: Regional Brain Metabolism and Long-Term Outcome. *Jama*, 286(17):2120–2127, 2001.

[226] V. Sindhwani, P. Niyogi, and M. Belkin. A Co-regularization Approach to Semi-supervised Learning with Multiple Views. In *Proceedings of the Workshop on Learning with Multiple Views at 22nd International Conference on Machine Learning*, pages 1135–1142, 2005.

[227] Andy Sivils, Felix Yang, John Q Wang, and Xiang-Ping Chu. Acid-sensing ion channel 2: Function and modulation. *Membranes*, 12(2):113, 2022.

[228] Xiaomu Song, Lawrence P Panych, et al. Data-Driven and Predefined ROI-Based Quantification of Long-Term Resting-State fMRI Reproducibility. *Brain Connectivity*, 6(2):136–151, 2016.

[229] Xueli Song, Rongpeng Li, Kaiming Wang, Yuntong Bai, Yuzhu Xiao, and Yu-Ping Wang. Joint Sparse Collaborative Regression on Imaging Genetics Study of Schizophrenia. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(2):1137–1146, 2023.

[230] S. Sonnenburg, G. Ratsch, C. Schafe, and B. Scholkopf. Large Scale Multiple Kernel Learning. *Journal of Machine Learning Research*, 7(57):1531–1565, 2006.

[231] S. Sonnenburg, G. Ratsch, and C. Schafer. A General and Efficient Multiple Kernel Learning Algorithm. In *Advances in Neural Information Processing Systems*, volume 18, 2005.

[232] T. Sorlie, C. M. Perou, R. Tibshirani, T. Aas, et al. Gene Expression Patterns of Breast Carcinomas Distinguish Tumor Subclasses with Clinical Implications. *Proceedings of the National Academy of Sciences U.S.A.*, 98(19):10869–10874, September 2001.

[233] Selma M Soyal, Markus Kwik, Ognian Kalev, Stefan Lenz, Greta Zara, Peter Strasser, Wolfgang Patsch, and Serge Weis. A TOMM40/APOE allele encoding APOE-E3 predicts high likelihood of late-onset Alzheimer's disease in autopsy cases. *Molecular Genetics & Genomic Medicine*, 8(8):e1317, 2020.

[234] N. Srivastava and R. Salakhutdinov. Multimodal Learning with Deep Boltzmann Machines. *Journal of Machine Learninig Research*, 15(1):2949–2980, 2014.

[235] Jason L Stein, Xue Hua, et al. Voxelwise Genome-Wide Association Study (vGWAS). *Neuroimage*, 53(3):1160–1174, 2010.

[236] Julia M Stephen, Rebecca Montano, et al. Somatosensory Responses in Normal Aging, Mild Cognitive Impairment, and Alzheimer's Disease. *Journal of Neural Transmission*, 117(2):217–225, 2010.

[237] S. Sun, L. Mao, Z. Dong, and L. Wu. *Multiview Machine Learning*. Springer Singapore, Singapore, 2019. ISBN: 978-981-13-3029-2.

[238] Shiliang Sun. A Survey of Multi-view Machine Learning. *Neural Computing and Applications*, 23(7):2031–2038, 2013.

[239] Shiliang Sun, Xijiong Xie, and Mo Yang. Multiview Uncorrelated Discriminant Analysis. *IEEE Transactions on Cybernetics*, 46(12):3272–3284, 2015.

[240] T. Sun and S. Chen. Locality Preserving CCA With Applications to Data Visualization and Pose Estimation. *Image and Vision Computing*, 25(5):531–543, 2007.

[241] Ting-Kai Sun, Song-Can Chen, Zhong Jin, and Jing-Yu Yang. Kernelized Discriminative Canonical Correlation Analysis. In *Proceedings of International Conference on Wavelet Analysis and Pattern Recognition*, volume 3, pages 1283–1287. IEEE, 2007.

[242] Tingkai Sun and Songcan Chen. Locality Preserving CCA with Applications to Data Visualization and Pose Estimation. *Image and Vision Computing*, 25(5):531–543, 2007.

[243] Tingkai Sun, Songcan Chen, et al. A Novel Method of Combined Feature Extraction For Recognition. In *Proceedings of IEEE International Conference on Data Mining*, pages 1043–1048. IEEE, 2008.

[244] Domizio Suva, Isabelle Favre, et al. Primary Motor Cortex Involvement in Alzheimer Disease. *Journal of Neuropathology and Experimental Neurology*, 58(11):1125–1134, 1999.

[245] Yoshio Takane, Heungsun Hwang, and Hervé Abdi. Regularized multiple-set canonical correlation analysis. *Psychometrika*, 73(4):753–775, 2008.

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

[247] A. Tenenhaus and M. Tenenhaus. Regularized Generalized Canonical Correlation Analysis for Multiblock or Multigroup Data Analysis. *European Journal of Operational Research*, 238(2):391–403, 2014.

[248] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, Inc., USA, 4th edition, 2008. ISBN: 9781597492720.

[249] Chris Thornton. Separability is a Learner's Best Friend. In *Proceedings of 4th Neural Computation and Psychology Workshop, London, 9–11 April 1997*, pages 40–46, London, 1998. Springer London.

[250] Qing Tian, Chuang Ma, Meng Cao, Songcan Chen, and Hujun Yin. A Convex Discriminant Semantic Correlation Analysis for Cross-View Recognition. *IEEE Transactions on Cybernetics*, 52(2):849–861, 2022.

[251] Y. Tian, X. Peng, L. Zhao, S. Zhang, and D. N. Metaxas. CR-GAN: Learning Complete Representations for Multi-view Generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 942–948, 2018.

[252] Y. Tian, L. Sigal, F. De la Torre, and Y. Jia. Canonical Locality Preserving Latent Variable Model for Discriminative Pose Inference. *Image and Vision Computing*, 31(3):223–230, 2013.

[253] Amy J Tibbo and Gonzalo Sand others Tejeda. Understanding PDE4's Function in Alzheimer's Disease: A Target for Novel Therapeutic Approaches. *Biochemical Society Transactions*, 47(5):1557–1565, 2019.

[254] Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Octave Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot. Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain. *Neuroimage*, 15(1):273–289, 2002.

[255] Uddin, Lucina Q and Yeo, BT and Spreng, R Nathan. Towards a universal taxonomy of macro-scale functional human brain networks. *Brain topography*, 32(6):926–942, 2019.

[256] Gary W Van Hoesen, Jean C Augustinack, Jason Dierking, Sarah J Redman, and Ramasamy Thangavel. The Parahippocampal Gyrus in Alzheimer's Disease: Clinical and Preclinical Neuroanatomical Correlates. *Annals of the New York Academy of Sciences*, 911(1):254–274, 2000.

[257] M. Varma and B. R. Babu. More Generality in Efficient Multiple Kernel Learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1065–1072, 2009.

[258] Annalena Venneri and Michael Shanks. The Claustrum and Alzheimer's Disease. *The Claustrum*, pages 263–275, 2014.

[259] Hrishikesh D Vinod. Canonical Ridge and Econometrics of Joint Production. *Journal of Econometrics*, 4(2):147–166, 1976.

[260] S. Virtanen, A. Klami, S. Khan, and S. Kaski. Bayesian Group Factor Analysis. In *Proceedings of International Conference on Artificial Intelligence and Statistics*, pages 1269–1277, 2012.

[261] Peter M Visscher, Naomi R Wray, et al. 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

[262] F. Wang and D. Zhang. A New Locality-Preserving Canonical Correlation Analysis Algorithm for Multi-View Dimensionality Reduction. *Neural Processing Letters*, 37:135–146, 2012.

[263] Hua Wang, Feiping Nie, et al. Identifying Quantitative Trait Loci via Group-Sparse Multitask Regression and Feature Selection: An Imaging Genetics Study of the ADNI Cohort. *Bioinformatics*, 28(2):229–237, 2012.

[264] Ke-Sheng Wang, Ying Liu, et al. Family-Based Association Analysis of NAV2 Gene with the Risk and Age at Onset of Alzheimer's Disease. *Journal of Neuroimmunology*, 310:60–65, 2017.

[265] Meiling Wang, Wei Shao, Xiaoke Hao, Li Shen, and Daoqiang Zhang. Identify Consistent Cross-Modality Imaging Genetic Patterns via Discriminant Sparse Canonical Correlation Analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(4):1549–1561, 2019.

[266] T. Wang, J. Lu, and G. Zhang. Two-Stage Fuzzy Multiple Kernel Learning Based on Hilbert-Schmidt Independence Criterion. *IEEE Transactions on Fuzzy Systems*, 26(6):3703–3714, 2018.

[267] W. Wang and Z. H. Zhou. A New Analysis of Co-training. In *Proceedings of the 27th International Conference on Machine Learning*, pages 1135–1142, 2010.

[268] W. Wang and Z.H. Zhou. Analyzing co-training style algorithms. In *Machine Learning: ECML 2007*, pages 454–465, 2007.

[269] Z. Wang, S. Chen, and T. Sun. MultiK-MHKS: A Novel Multiple Kernel Learning Algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):348–353, 2008.

[270] Zhiqun Wang, Peipeng Liang, Xiuqin Jia, Guangwei Jin, Haiqing Song, Ying Han, Jie Lu, and Kuncheng Li. The Baseline and Longitudinal Changes of PCC Connectivity in Mild Cognitive Impairment: A Combined Structure and Resting-State fMRI Study. *PloS One*, 7(5):e36838, 2012.

[271] Wang, Ke-Sheng and Tonarelli, Silvina and Luo, Xingguang and Wang, Liang and Su, Brenda and Zuo, Lingjun and Mao, ChunXiang and Rubin, Lewis and Briones, David and Xu, Chun. Polymorphisms within ASTN2 gene are associated with age at onset of Alzheimer's disease. *Journal of Neural Transmission*, 122:701–708, 2015.

[272] Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A Penalized Matrix Decomposition, with Applications to Sparse Principal Components and Canonical Correlation Analysis. *Biostatistics*, 10(3):515–534, 2009.

[273] Daniela M Witten and Robert J Tibshirani. Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data. *Statistical Applications in Genetics and Molecular Biology*, 8(1), 2009.

[274] Herman Wold. Path Models with Latent Variables: The NIPALS Approach. In *Quantitative Sociology*, pages 307–357. Elsevier, 1975.

[275] J. Wu, X. Xie, L. Nie, Z. Lin, and H. Zha. Unified Graph and Low-Rank Tensor Learning for Multi-View Clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2020.

[276] Jie Xiang, Xin Wang, Yuan Gao, Ting Li, Rui Cao, Ting Yan, Yunxiao Ma, Yan Niu, Jiayue Xue, and Bin Wang. Phosphodiesterase 4D gene modifies the functional network of patients with mild cognitive impairment and Alzheimer's disease. *Frontiers in Genetics*, 11:890, 2020.

[277] Xiaolin Xiao, Yongyong Chen, et al. Prior Knowledge Regularized Multiview Self-Representation and Its Applications. *IEEE Transactions on Neural Networks and Learning Systems*, 32(3):1325–1338, 2020.

[278] Xijiong Xie, Yanfeng Li, and Shiliang Sun. Deep Multi-View Multiclass Twin Support Vector Machines. *Information Fusion*, 91:80–92, 2023.

[279] Y. Xie, D. Tao, W. Zhang, Y. Liu, L. Zhang, and Y. Qu. On Unifying Multi-View Self-Representations for Clustering by Tensor Multi-Rank Minimization. *International Journal of Computer Vision*, 126:1157–1179, 2018.

[280] C. Xu, D. Tao, and C. Xu. Multi-View Learning With Incomplete Views. *IEEE Transactions on Image Processing*, 24(12):5812–5825, 2015.

[281] Z. Xu, R. Jin, S. Zhu, M. R. Lyu, and I. King. Smooth Optimization for Effective Multiple Kernel Learning. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, pages 637–642, 2010.

[282] F. Xue, X. Wu, S. Cai, and J. Wang. Learning Multi-View Camera Relocalization With Graph Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11372–11381, 2020.

[283] Chaogan Yan and Yufeng Zang. DPARSF: A MATLAB Toolbox for Pipeline Data Analysis of Resting-State fMRI. *Frontiers in Systems Neuroscience*, 4:13, 2010.

[284] Jingwen Yan, Shannon L Risacher, Kwangsik Nho, Andrew J Saykin, and LI Shen. Identification of Discriminative Imaging Proteomics Associations in Alzheimer's Disease via a Novel Sparse Correlation Model. In *Proceedings of Pacific Symposium on Biocomputing 2017*, pages 94–104. World Scientific, 2017.

[285] J. Yang and X. Zhang. Feature-Level Fusion of Fingerprint and Finger-Vein for Personal Identification. *Pattern Recognition Letters*, 33(5):623–628, 2012.

[286] Mo Yang and Shiliang Sun. Multi-view Uncorrelated Linear Discriminant Analysis with Applications to Handwritten Digit Recognition. In *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, pages 4175–4181. IEEE, 2014.

[287] Pei Yang and Wei Gao. Information-theoretic multi-view domain adaptation: A theoretical and empirical study. *Journal of Artificial Intelligence Research*, 49:501–525, 2014.

[288] X. Yang, W. Liu, and W. Liu. Tensor Canonical Correlation Analysis Networks for Multi-View Remote Sensing Scene Recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(06):2948–2961, 2022.

[289] Z. Yang, L. Tang, K. Zhang, and P. K. Wong. Multi-View CNN Feature Aggregation with ELM Auto-Encoder for 3D Shape Recognition. *Cognitive Computation*, 10:908–921, 2018.

[290] Michael A Yassa, Guillermo Verduzco, et al. Altered fMRI Activation During Mental Rotation in those at Genetic Risk for Alzheimer Disease. *Neurology*, 70(20):1898–1904, 2008.

[291] Takamasa Yokoi, Hirohisa Watanabe, et al. Involvement of the Precuneus/Posterior Cingulate Cortex is Significant for the Development of Alzheimer's Disease: A PET (THK5351, PiB) and resting fMRI Study. *Frontiers in Aging Neuroscience*, 10:304, 2018.

[292] Xinge You, Jiamiao Xu, Wei Yuan, Xiao-Yuan Jing, Dacheng Tao, and Taiping Zhang. Multi-view Common Component Discriminant Analysis for Cross-view Classification. *Pattern Recognition*, 92:37–51, 2019.

[293] Enyan Yu, Zhengluan Liao, et al. Directed Functional Connectivity of Posterior Cingulate Cortex and Whole Brain in Alzheimer's Disease and Mild Cognitive Impairment. *Current Alzheimer Research*, 14(6):628–635, 2017.

[294] Guoxian Yu, Yuying Xing, Jun Wang, et al. Multiview Multi-Instance Multilabel Active Learning. *IEEE Transactions on Neural Networks and Learning Systems*, 33(9):4311–4321, 2021.

[295] S. Yu, B. Krishnapuram, R. Rosales, and R. B. Rao. Bayesian Co-training. *The Journal of Machine Learning Research*, pages 2649–2680, 2011.

[296] Yuyuan Yu, Guoxu Zhou, Haonan Huang, Shengli Xie, and Qibin Zhao. A Semi-Supervised Label-Driven Auto-Weighted Strategy for Multi-View Data Classification. *Knowledge-Based Systems*, 255:109694, 2022.

[297] Yun-Hao Yuan, Jin Li, Yun Li, Jipeng Qiang, Yi Zhu, Xiaobo Shen, and Jianping Gou. Learning Canonical f-Correlation Projection for Compact Multiview Representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19260–19269, 2022.

[298] Yun-Hao Yuan, Yun Li, et al. Laplacian Multiset Canonical Correlations for Multi-view Feature Extraction and Image Recognition. *Multimedia Tools and Applications*, 76(1):731–755, 2017.

[299] An Zeng, Huabin Rong, Dan Pan, Longfei Jia, Yiqun Zhang, Fengyi Zhao, Shaoliang Peng, and Alzheimer's Disease Neuroimaging Initiative (ADNI). Discovery of Genetic Biomarkers for Alzheimer's Disease Using Adaptive Convolutional Neural Networks Ensemble and Genome-wide Association Studies. *Interdisciplinary Sciences: Computational Life Sciences*, 13(4):787–800, 2021.

[300] Shanhua Zhan, Weijun Sun, Cuifeng Du, and Weifang Zhong. Diversity-Promoting Multi-View Graph Learning for Semi-Supervised Classification. *International Journal of Machine Learning and Cybernetics*, 12(10):2843–2857, 2021.

[301] Bin Zhang, Qianyao Qiang, Fei Wang, and Feiping Nie. Fast multi-view semi-supervised learning with learned graph. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):286–299, 2020.

[302] Jingjing Zhang and Rui Wang. Deregulated lncRNA MAGI2-AS3 in Alzheimer's disease attenuates amyloid-β induced neurotoxicity and neuroinflammation by sponging miR-374b-5p. *Experimental gerontology*, 144:111180, 2021.

[303] Jintao Zhang and Jun Huan. Inductive Multi-Task Learning with Multiple View Data. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 543–551, 2012.

[304] N. Zhang, S. Ding, H. Liao, and W. Jia. Multimodal Correlation Deep Belief Networks for Multi-View Classification. *Applied Intelligence*, 49(5):1925–1936, 2019.

[305] Nan Zhang and Shiliang Sun. Multiview Graph Restricted Boltzmann Machines. *IEEE Transactions on Cybernetics*, 2021.

[306] Xiao-Lei Zhang. Convex Discriminative Multitask Clustering. *IEEE transactions on pattern analysis and machine intelligence*, 37(1):28–40, 2014.

[307] Z. Zhang, Q. Zhu, G.-S. Xie, Y. Chen, Z. Li, and S. Wang. Discriminative Margin-Sensitive Autoencoder for Collective Multi-View Disease Analysis. *Neural Networks*, 123:94–107, 2020.

[308] H. Zhao, Z. Ding, and Y. Fu. Multi-View Clustering via Deep Matrix Factorization. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 2921–2927, 2017.

[309] Jing Zhao, Zengyu Qiu, and Shiliang Sun. Multi-View Multi-Label Active Learning with Conditional Bernoulli Mixtures. *International Journal of Machine Learning and Cybernetics*, pages 1–13, 2022.

[310] Jing Zhao, Xijiong Xie, Xin Xu, and Shiliang Sun. Multi-view Learning Overview: Recent Progress and New Challenges. *Information Fusion*, 38:43–54, 2017.

[311] Xuran Zhao, Nicholas Evans, and Jean-Luc Dugelay. A subspace co-training framework for multi-view clustering. *Pattern Recognition Letters*, 41:73–82, 2014.

[312] Yue Zhao, Xinge You, Shujian Yu, Chang Xu, Wei Yuan, Xiao-Yuan Jing, Taiping Zhang, and Dacheng Tao. Multi-view manifold learning with locality alignment. *Pattern Recognition*, 78:154–166, 2018.

[313] Jun-Juan Zheng, Wen-Xing Li, et al. Low Expression of Aging-Related NRXN3 is Associated with Alzheimer Disease: A Systematic Review and Meta-Analysis. *Medicine*, 97(28), 2018.

[314] W. Zheng, X. Zhou, C. Zou, and L. Zhao. Facial Expression Recognition Using Kernel Canonical Correlation Analysis (KCCA). *IEEE Transactions on Neural Networks*, 17(1):233–238, 2006.

[315] Pascal Zille, Vince D. Calhoun, and Yu-ping Wang. Enforcing co-expression within a brain-imaging genomics regression framework. *IEEE Transactions on Medical Imaging*, 37(12):2561–2571, 2018.