Multi-View Discriminant Canonical Correlation Analysis: Regularization, Scalability to Adaptability

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

by

Ankita Mandal
Senior Research Fellow

Under the supervision of
Dr. Pradipta Maji, Professor

Machine Intelligence Unit
Indian Statistical Institute, Kolkata

December 2022

To the source of my endurance:
Baba and Maa.

Acknowledgements

Apart from a thesis with several research works, a Ph.D. is a journey of experience and memories of ups and downs. Looking back to the time, I realize that I am blessed for having continuous support from several kind-hearted people. They had given me the confidence and endurance to complete this journey successfully.

A profound sense of gratitude binds me to my thesis supervisor and mentor Prof. Pradipta Maji. He has been the backbone in molding my research enhancement since my post-graduate days. I am fortunate to have him as my teacher, who has guided me not only in my research work but also taught me life lessons. His immeasurable support, training, and guidance kept me motivated throughout this journey. He taught me how to do research work in an organized way. He had always been available for me, whenever I needed any technical support. I have learned from him the importance of discipline and honesty in research. No word of thanks can sum up the gratitude that I owe to him.

I express my sense of indebtedness to the Dean of Studies and the Director of the Indian Statistical Institute for providing me the research fellowship and grants, and a peerless infrastructure and environment for research. I owe my sincere gratitude to all the faculty members of the Machine Intelligence Unit, Indian Statistical Institute, for their continued support, encouragement, and helpful suggestions during my Ph.D. tenure. I am also grateful to the authorities of the institute for providing the facilities, which have helped me to complete my research smoothly. I would also like to acknowledge all the timely supports that I have received from the office staffs of our institute throughout the tenure of my Ph.D.

I would like to express my gratitude to all my Biomedical Imaging and Bioinformatics Lab members and alumni especially, Debamita Kumar, Suman Mahapatra, Sankar Mondal, Ekta Shah, Abhirup Banerjee, Aparajita Khan, Gunjan Gautam, Pratik Dutta, Nabina Dey, Shaswati Roy, Sushmita Paul, Partha Garai, and Debanjan Chakraborty for creating such a healthy environment to carry out my thesis work. I also want to thank all my teachers from my childhood. Without their priceless blessings and teachings, it would be very difficult for me to complete this thesis.

Finally, yet importantly, I sought inspiration and I owe a great deal to my beloved parents, Mr. Ranjit Mandal and Mrs. Maya Mandal, for being the pillars of my dreams. Their unconditional love, support, and encouragement give me endurance not only during my Ph.D. journey but also throughout my life. I am blessed to have Mr. Santi Ranjan Paul and Mrs. Sabita Paul, as my father-in-law and mother-in-law. Their constant support, encouragement, and blessings have enabled me to pursue my Ph.D. degree. I would also like to mention all the well wishes and encouragements that I have received from my uncles, aunts, cousins, and other family members. Last but certainly not the least, I must express a deep sense of gratitude to a very special person, Santanu Paul, my husband, for his constant support during my Ph.D. tenure.

I thank all of them whose names have been missed unintentionally but have contributed in one way or another to my thesis work.

Ankita Mandal

Abstract

Multi-view learning is an emerging machine learning paradigm that focuses on discovering patterns in data represented by multiple distinct views. One of the important issues associated with real-life high-dimensional multi-view data is how to integrate relevant and complementary information from multiple views, while generating discriminative subspaces for analysis. Although the integration of multi-view data is expected to provide an intrinsically more powerful model than its single-view counterpart, it poses its own set of challenges. The most important problems associated with multi-view data analysis are presence of noisy, irrelevant and heterogeneous views, high-dimension low-sample size nature of individual views, and updating the databases with new views.

In this regard, the thesis addresses the problem of multi-view data integration, for both static and dynamic data sets, in the presence of high-dimensional noisy and redundant views. The main contribution of the present work is to design some novel algorithms, based on the theory of canonical correlation analysis (CCA), to extract informative subspaces for multi-view classification, and theoretically analyze the important properties of these transformed spaces and new algorithms. The "curse of dimensionality" problem due to "high-dimension low-sample size" characteristics of real-life data is addressed, by judiciously integrating the CCA and ridge regression optimization technique. The relation between CCA and its regularized counterpart is established, which enables extraction of relevant and significant features sequentially from bimodal data sets for classification and addresses the scalability issue of real-life high-dimensional data.

To integrate multi-view data using multiset CCA (MCCA), a new block matrix representation is introduced. It facilitates generation of discriminative subspaces having maximum pairwise correlation, and makes the MCCA model scalable to high-dimensional multi-view data. Integration of MCCA with multiset ridge regression model addresses the "curse of dimensionality" problem of individual views. In order to integrate dynamic multi-view data, a novel adaptive MCCA model is proposed, which incrementally updates canonical variables when new views are available for the analysis. The adaptive model ensures selection of relevant and complementary views during data integration, while discarding irrelevant and redundant ones. To make the adaptive framework scalable to high-dimensional data, a new model is introduced under common latent representation. Finally, a graph based approach is judiciously integrated with this adaptive model to utilize the underlying geometry of the data in different views.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Data is a representation of facts, concepts, or instructions in a formalized manner, which should be suitable for communication, interpretation, or processing by a human or electronic machine. It can exist in various forms: as numbers or text recorded on paper, as bits or bytes stored in electronic memory, or as facts living in a person's mind. In computing, data is the knowledge that has been translated into a form that is efficient for conditioning or processing. Relative to present-day computers and transmission media, data is information converted into binary digital form. The growth of the web and smartphones over the past decade led to a surge in digital data creation. Data now includes text, audio, and video information, as well as log and web activity records. Data streams in from every picture taken, every file saved, every search query submitted to a search engine, every social media interaction, and every experiment performed. As data is sprawling across more devices, applications and cloud platforms, and is available in more formats, it is growing at an exponential rate with 90% of the world's data being generated in the last two years alone. It is predicted that the global data volume will reach 175 zettabytes by 2025 [221]. And, it is not only the volume of the data that has grown drastically, but also the variety of it. Moreover, having an abundance of data by itself does not make any sense, it is more important to analyze the data and get the benefit from it.

A pattern gives the knowledge of data. Hence, to analyze the data, one needs to understand or recognize the pattern properly. Pattern recognition is the automated recognition of patterns and regularities in data [265]. It tries to simulate the human brain's neural network capabilities, which further advances artificial intelligence. It uses machine learning [29,79] algorithms to identify patterns. It analyzes data based on statistical information or knowledge gained from patterns and their representation. Machine learning is a branch of artificial intelligence and computer science that focuses on the use of data and algorithms to imitate the way humans learn; gradually improving its accuracy. Pattern recognition and machine learning is a versatile practice that has found their way into many different industries and social contexts.

In pattern recognition and machine learning, a feature is an individual measurable property or attribute used to characterize a data set. Features can be in the raw form of data that cannot be used in all types of real-life problems. For example, a color can be represented in RGB format or HSV format. Thus, a color can have two different representations or encodings. Both of these representations or encodings can be used to solve different kinds of problems. Some tasks that may be difficult with one representation can become easy with another. For example, the task “select all red pixels in the image” is simpler in the RGB format, whereas “make the image less saturated” is simpler in the HSV format. Machine learning algorithms can be broadly categorized into the following three groups, namely, supervised learning, unsupervised learning, and semi-supervised learning [265], depending on the learning strategy.

- **Supervised learning** is a type of machine learning algorithm where a set of labelled data is used to predict the labels of unknown objects. These algorithms use a two-stage methodology for identifying the patterns. The first stage includes the development or construction of a model, and the second stage involves the prediction of new or unseen objects using the developed model. One practical example of supervised learning problems is the text classification problem. Here, the goal is to predict the class label of a given piece of text. One particularly popular topic in text classification is to predict the sentiment of a piece of text, like a tweet or a product review. This is widely used in the e-commerce industry to help companies to determine negative comments made by the customers.

- In **unsupervised learning**, the objective is to learn patterns from a data set without using any prior information. Since the data is not labeled, the machine should learn to categorize the data based on the similarity and finds patterns in the data. Clustering is an unsupervised technique where the goal is to find natural groups or clusters by interpreting the input data. It is commonly used for determining customer segments to build marketing or other business strategies. For example, an e-commerce site uses clustering algorithms to implement a user-specific recommendation system. Another example is grouping subscribers of a YouTube channel. The channel owner has a lot of data about the subscribers. By using these data, a clustering algorithm can group the subscribers, which helps the owner to create content of a video for each group.

- In **semi-supervised learning**, both the labeled and unlabeled data are used to train the model. In several application domains, acquiring data is easy, but acquiring labeled data turns out to be expensive. Hence, the combination of a very small amount of labeled data and a very large amount of unlabeled data may help to learn a semi-supervised model. An initial model is developed by using the limited set of training labeled samples and unlabeled data is used to refine the model. An example of semi-supervised learning is speech analysis. Labeling audio files typically is a very intensive task that requires a lot of human resources. Applying semi-supervised learning techniques can help to improve traditional speech analytic models.

Figure 1.1 represents the difference between supervised, unsupervised, and semi-supervised learning.

The number of input variables or features of a data set is referred to as the dimensionality of the data set. As the dimensionality of the input data set increases, machine learning algorithms become more complex and more prone to incorrect predictions. This is known as the “curse of dimensionality” [23, 265]. A higher number of dimensions theoretically allows more information to be stored, but practically it rarely helps due to the higher possibility of noise and redundancy in the real-world data. If the machine learning model is trained

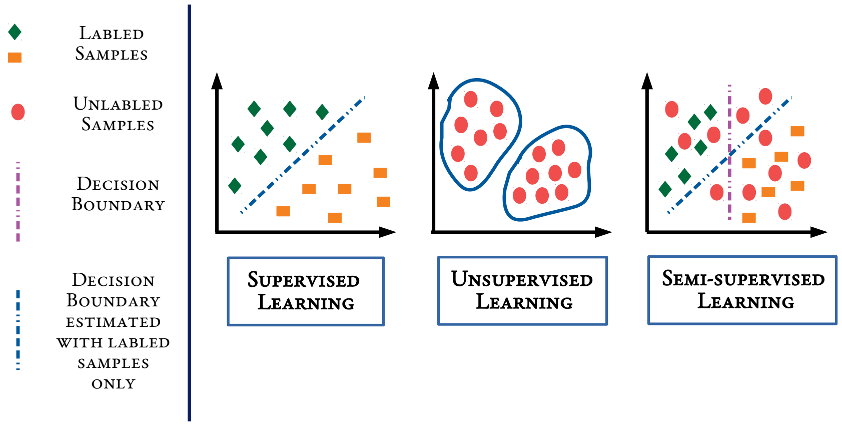


Figure 1.1: Difference between supervised, unsupervised, and semi-supervised learning.

on high-dimensional data, it becomes overfit and results in poor generalization to unseen data in many cases. **Dimensionality reduction** refers to the techniques that reduce the number of input variables or features in a data set. Besides eliminating the overfitting and redundancy problem, dimensionality reduction also leads to better human interpretations and less computational cost with the simplification of models. It helps machine learning algorithms to learn the intricate pattern of a data set with lesser cost and more accuracy. An example of dimensionality reduction is to identify an email as spam or not. This task can have several features such as the title of an email, whether it is generic or specific, the contents of the email, whether the email is based on a template, and so on. Many of these features may also overlap with each other where the dimensionality reduction can be used to separate spam from important emails. Dimensionality reduction can be performed in the following two ways, namely, feature selection and feature extraction [87].

- **Feature selection** reduces the dimensionality of the measurement space by discarding redundant or least information-carrying features.
- **Feature extraction** is a process of dimensionality reduction where all the information contained in the original measurement space are used to obtain a new transformed space, thereby mapping a higher dimensional pattern to a lower dimensional one.

Both feature selection and feature extraction are processes where each sample or observation in a high-dimensional measurement space is transformed into a low-dimensional space. The main objective of feature selection and feature extraction is to retain or generate the optimum salient characteristics necessary for the recognition process and to reduce the dimensionality of the measurement space so that effective and easily computable algorithms can be devised for efficient class labels determination. The problem of feature selection and feature extraction has two aspects, namely, formulating a suitable criterion to evaluate the goodness of a feature set and searching for the optimal set in terms of the criterion. In general, those features are considered to have optimal saliencies for which interclass (respectively, intraclass) distances are maximized (respectively, minimized). The criterion for a good feature is that it should be unchanging with any other possible variation within a class while emphasizing differences that are important in discriminating between patterns of different types.

There are generally two ways to represent the data for a machine learning algorithm, namely, feature vector-based data and relational data [180]. In feature vector-based representation, n samples are observed in p-dimensional feature space or measurement space, where each feature can be represented by numerical, textual, or categorical values. For example, a color image has three color components of a pixel, namely, red, green, and blue. On the other hand, the pairwise relationships between n samples are measured in relational data. One practical example of relational data is news article categorization. Here, two articles or related topics can be considered to be alike if there is a connection between these articles. A set of n observations or samples, represented by either p-dimensional feature vectors or by n^2 pairwise relationships, is referred to as a "modality" or "view" of a data set. There are several real-life applications, where a single type of information may fail to analyze a given problem or distinguish the patterns of a data set completely. For example, it cannot be claimed that two news articles belong to the same news category if they share a hyperlink connection. The resemblance between their content has to be assessed before making such an assertion. On the other hand, multiple views of the same set of observations can capture complementary information. In this regard, the dimensionality reduction problem associated with multi-view data sets is addressed in this thesis.

1.1 Multi-View Data Analysis

Multi-view data analysis is one of the emerging areas of machine learning. The main objective of multi-view learning is to analyze patterns in data represented by multiple views [251]. Due to the huge evolution in several data collection, measurement, and representation techniques, the multi-view data sets are almost everywhere in recent practical-world applications. During the last decade, the idea of combining knowledge from diverse sources has taken over the conventional single-view learning models. It becomes an operational area of study due to the massive success in a broad scope of real-life applications, such as biomedical imaging, integration of multi-omics data, multi-source text mining, multi-camera face and facial expression recognition, imaging genetics, multi-source image retrieval and so on [158, 220, 320]. Some of the various application areas of multi-view learning are demonstrated in Figure 1.2.

There are various reasons behind the immense success of multi-view learning over the single-view analyses. Some of these highlights are described next.

- **Comprehensive View of the System**: Different views have different characteristics. If the relevant views are combined to analyze a pattern, it provides more impact on the learning process. For example, multiple cameras capture different angles and views of a person, which helps to identify the person more conveniently than a single camera view. As facial appearance may vary due to the various lighting condition, light angles, pose, or facial expressions, multiple cameras are able to capture a significant number of images of a face in different poses and lighting cir-

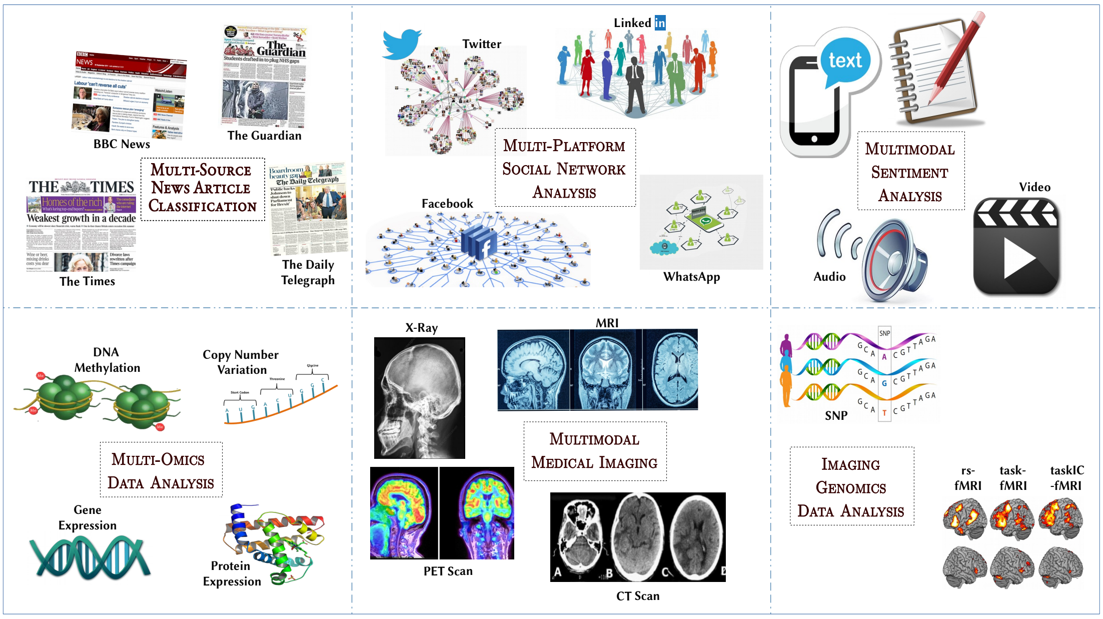


Figure 1.2: Various application areas of multi-view data analysis.

cumstances. It provides more robust and precise face recognition outcomes than a single-camera/single-view analysis.

- **Complementary Information**: The information associated with different views may have complementary nature. Integrating this complementary information may provide more insight into the problem. For example, both copy number variation and gene expression share the genetic knowledge of an individual. The copy number variation indicates how many times a particular gene sequence has been replicated within the DNA, whereas the overexpression or underexpression of a gene is represented by gene expression data. Integration of both complementary and compatible views is supposed to increase learning performance.

- **Cross-Platform Analysis**: As multiple views are available, it is feasible to draw a connection between variables observed in various views. If the information corresponding to functional magnetic resonance imaging is combined with that of single nucleotide polymorphism (SNP), then it helps to identify the brain region alterations which is triggered by corresponding SNP changes in genes.

- **Resilience to Noise**: It may be possible that a real-life data set has noise. If the information associated with different views is combined, then there is a chance that noisy observations in a particular view can be neutralized by the complementary observations of relevant views.

In spite of having an abundance of benefits in multi-view data analysis, there are several challenges and hurdles associated with it [320], which are addressed briefly in the next section.

1.2 Challenges in Multi-View Data Analysis

Conventional machine learning algorithms are developed to work on single-view data. For example, support vector machines, artificial neural networks, discriminant analysis, spectral clustering, and kernel machines are supposed to analyze single-view data. As multi-view data sets have their own set of challenges, few modifications have to be done to these algorithms to learn multi-view data sets. The challenges of multi-view data analysis, which are mostly focused on dimensionality reduction are discussed below.

- **Data Heterogeneity**: The easiest process to analyze the information of multi-view data sets using the traditional machine learning algorithms is to join all multiple views into one single view. But, this naive integration is not good as each modality has its distinct characteristics. Each view has a different scale, unit, and variance. Hence, different views may not be compatible with each other. For example, DNA methylation data is made up of β-values which lie in $[0, 1]$, while RNA sequence-based gene expression data is estimated in RPM (reads per million) and represented by real values in the order of 10^5. The naive integration of features from these two heterogeneous views is more likely to be dominated by the view, which has high variance. Hence, the integration process has to be unbiased so that the inherent properties are conserved during the learning process.

- **Curse of Dimensionality Problem**: In real-world applications, data sets consist of an enormous number of observed variables. For example, an image has nearly 10^6 pixels, DNA microarrays have almost 20K genes, thousands of words are present in a document file, and so on. On the contrary, the number of observed samples is generally very small. Due to the limited number of training samples, the learning models incline to overfit the data, thus the generalization performance decreases. The high-dimension low-sample data also have multicollinearity problems. Hence, the consistency properties of the eigenvalues and the corresponding eigenvectors of the rank deficient sample covariance matrix are degraded [126]. In high dimensions, the feature space becomes geometrically sparse, which leads to the non-invertibility of the covariance matrix.

- **Irrelevant and Redundant Views**: The observations in various views can be corrupted by noise due to measurement errors in real-life applications. The noise has to be taken care of explicitly, otherwise, it may be propagated in distinct views or even overstated during the data fusion procedure. On the other hand, most of the machine learning algorithms have an assumption that each view is knowledgeable and obtain consistent and homogeneous information about the data set. Hence, these algorithms incorporate all the available views in the learning process. But, in reality, some of the views may provide redundant, insignificant, or even worse information. Because of the presence of irrelevant, redundant, and noisy views, the integration of all available views can reduce the performance of the learning process.

- **View Disagreement**: Different views are supposed to follow a global class structure in multi-view data analysis. That means each sample should belong to the same class in all views. However, in real-world applications, the views are often corrupted by noise. As a result of that, a set of observations in some views may be corrupted, while in other views it may remain unaffected. One practical example of this situation is view disagreement in multi-sensory data sets. A sensor may have an incorrect state between two normal states by mistake, which creates confusion between various views. If there is a disagreement or corruption present in the data set, the classes recognized in different views would not correlate with each other, which makes the integration process more difficult [48].

- **Noisy and Low-Rank Geometry of Views**: The intrinsic geometrical structure hidden in a data set has the power to enhance the learning performance of dimensionality reduction, data reconstruction, clustering, and classification [22,44,104,105, 125, 227, 237, 261]. In this regard, many approaches have been proposed in recent years to identify geometrical knowledge of the data by integrating the information from multiple views. The performance of most of the existing graph-based methods relies on the predefined graph. If the data is noisy, and has the incomplete and/or heterogeneous views, then a consensus graph from the data has to be learned. However, it should be noted that the difference between the nearest and farthest neighbor points from a certain point is insignificant for high-dimensional data sets [5].

- **Updation in Database**: Every day, a huge amount of data is being added to the existing databases. Sometimes new instances may be added to the existing samples or new modalities may be considered for better analysis. For example, The Cancer Genome Atlas (TCGA) (`https://cancergenome.nih.gov/`) updates and releases the new data, both samples and modalities, twenty-two times in the last five years. Incremental learning is a machine learning paradigm where the learning process takes place whenever new data is merged with or deleted from the existing data set and the solutions already obtained are only modified. Thus, the multi-view learning algorithms should be adaptive or incremental in nature.

- **Incomplete Views**: A common assumption of multi-view data analysis algorithms is all the views have a unique set of samples. But, in practical application, there may be various failures or faults in collecting and pre-processing the data on different views. Because of that, a sample may not be observed in one view, which makes the view incomplete. In effect, this missing sample has to be discarded from all other views, which reduces the sample size. A small set of training samples may lead to the overfitting of the data. Hence, by establishing a connection between the views the missing sample can be restored with the help of the complete views [299] without discarding the missing sample from all views.

Few of these aforementioned challenges, for example, data heterogeneity, presence of irrelevant, redundant, and incomplete views, disagreement among different views, and updation in the database, are applicable for multi-view data sets only, while other challenges like the presence of noise and high-dimension low-sample size characteristic are valid for both the single-view and multi-view data sets. Thus, some advanced machine learning algorithms have to be developed that can address these challenges efficiently and extract latent information from multi-view data sets.

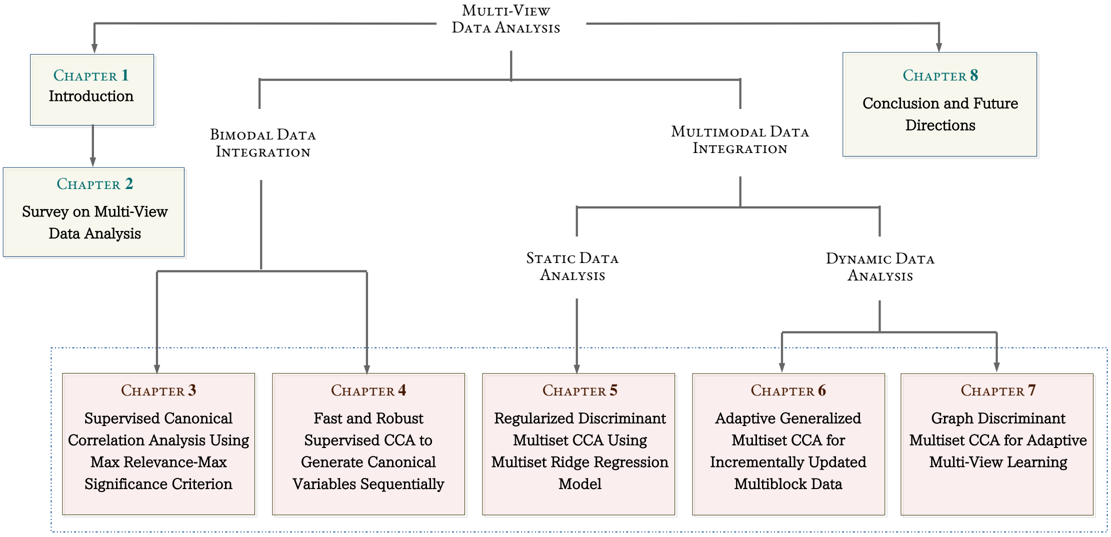

Multi-View Data Analysis
Chapter 1
Introduction
Chapter 2
Survey on Multi-View Data Analysis
Bimodal Data Integration
Multimodal Data Integration
Chapter 8
Conclusion and Future Directions
Static Data Analysis
Dynamic Data Analysis
Chapter 3
Supervised Canonical Correlation Analysis Using Max Relevance-Max Significance Criterion
Chapter 4
Fast and Robust Supervised CCA to Generate Canonical Variables Sequentially
Chapter 5
Regularized Discriminant Multiset CCA Using Multiset Ridge Regression Model
Chapter 6
Adaptive Generalized Multiset CCA for Incrementally Updated Multiblock Data
Chapter 7
Graph Discriminant Multiset CCA for Adaptive Multi-View Learning

Figure 1.3: Outline of the thesis.

1.3 Scope and Organization of Thesis

In this context, the thesis presents a set of learning algorithms to address some of the problems associated with multi-view data analysis. The high-dimension low-sample size characteristic of individual views is one of the crucial challenges related to multi-view learning. This leads to ill-conditioning of the sample covariance matrix of the high-dimensional view. Furthermore, a small subset, among the huge amount of extracted features, is effective to perform a certain task. Hence, the goal of multimodal data analysis is to extract a reduced set of most relevant features. Instead of generating all possible features, if each feature is generated sequentially, the quality of each extracted feature can be evaluated, and finally, the required number of features can be extracted from the multimodal data sets. On the other hand, real-life data sets are often plagued with noise. It may also happen that some of the views provide disparate, redundant, or even worse information. Moreover, the views may be added with time. So, it is necessary to develop a model that can generate the new feature from that of the existing modalities and the new modality without repeating the same procedure with the original data augmented by the new modality. The key contribution of this thesis is to design some novel algorithms to extract relevant and significant features from multi-view data sets and theoretically analyze the salient characteristics of these transformed feature spaces.

Figure 1.3 represents the outline of the thesis. The thesis comprises eight chapters. The importance of multi-view data analysis is described in Chapter 1. Some challenges associated with multi-view learning are also discussed in this chapter. A brief study on existing multi-view data integration algorithms is presented in Chapter 2.

Chapter 3 presents a novel supervised regularized canonical correlation analysis (CCA), termed as CuRSaR, to extract relevant and significant features from bimodal multidimensional data sets. The proposed algorithm extracts a new set of features from two multidimensional data sets by maximizing the relevance of extracted features with respect to sample categories and significance among them. It integrates judiciously the merits of regularized CCA and rough hypercuboid approach. An analytical formulation, based on spectral decomposition, is introduced to establish the relationship between the covariance matrices of different regularization parameters. It makes the computational complexity of the proposed algorithm significantly lower than that of the existing methods. The concept of hypercuboid equivalence partition matrix of rough hypercuboid is used to compute both relevance and significance of a feature. The equivalence partition matrix offers an efficient way to find optimum regularization parameters. The superiority of the proposed algorithm over other existing methods, in terms of computational complexity and classification accuracy, is established extensively on several real-life cancer data sets.

One of the main problems associated with real-life multi-view data sets is how to extract relevant and significant features sequentially. The algorithm presented in Chapter 3 extracts relevant and significant features simultaneously from two multidimensional data sets. In general, a huge number of irrelevant and insignificant features may be present in the extracted feature set, which may degrade the classification accuracy by reducing the useful information. Thus, if the features are extracted sequentially, then the required number of features can be generated by evaluating the quality of each feature. In this regard, a fast and robust feature extraction algorithm, termed as FaRoC, is presented in Chapter 4, integrating judiciously the merits of CCA and rough sets. The proposed algorithm extracts new features sequentially from two multidimensional data sets by maximizing their relevance with respect to class labels and significance with respect to already-extracted features. To generate canonical variables sequentially, an analytical formulation is introduced to establish the relation between regularization parameters and CCA. The formulation enables the proposed algorithm to extract required number of correlated features sequentially with lesser computational cost as compared to existing methods. To compute both significance and relevance measures of a feature, the concept of hypercuboid equivalence partition matrix of a rough hypercuboid approach is used. It also provides an efficient way to find optimum regularization parameters employed in CCA. The efficacy of the proposed FaRoC algorithm, along with a comparison with other existing methods, is extensively established on several real-life cancer data sets.

Both CuRSaR and FaRoC, presented in Chapter 3 and Chapter 4, respectively, can only account for two sets of variables. The multiset CCA (MCCA) is a well-known statistical method for multi-view data integration. It finds a linear subspace that maximizes the correlations among different views. However, the existing methods to find the multiset canonical variables are computationally very expensive, which restricts the application of the MCCA in real-life big data analysis. The covariance matrix of each high-dimensional view may also suffer from the singularity problem due to the limited number of samples. Moreover, the MCCA based existing feature extraction algorithms are, in general, unsupervised in nature. In this regard, a new supervised feature extraction algorithm, termed as ReDMiCA, is presented in Chapter 5, which integrates multimodal multidimensional data sets by solving the maximal correlation problem of the MCCA. A new block matrix representation is introduced to reduce the computational complexity of computing the canonical variables of the MCCA. The analytical formulation enables efficient computation of the multiset canonical variables under the supervised ridge regression optimization technique. It deals with the "curse of dimensionality" problem associated with high-dimensional data and facilitates the sequential generation of relevant features with significantly lower computational cost. The effectiveness of the proposed multiblock data integration algorithm, along with a comparison with other existing methods, is demonstrated on several benchmark and real-life cancer data sets.

One of the major problems in real-life multiblock dynamic data analysis is that all the modalities may not be available initially. The databases are generally updated incrementally. New modalities may be added to the existing modalities. So, it is necessary to develop a model that can generate the new features from that of the existing modalities and the new modality without repeating the same procedure with the original data augmented by the new modality. Moreover, it may also happen that some of the views have noisy or even inconsistent information with respect to other views. So, it is necessary to evaluate the quality of a new modality before considering it for feature extraction. In this regard, a new MCCA, termed as incremental MCCA (IMCCA), is presented in Chapter 6. When a new modality is available for the analysis, the IMCCA generates the new canonical variables from that of the earlier modalities, without repeating the same procedure with the original data augmented by the new modality. The proposed IMCCA deals with the "curse of dimensionality" problem associated with multidimensional data sets, by using the ridge regression optimization technique. Using the proposed IMCCA model, a new feature extraction algorithm, termed as SeFGeIM is introduced, which considers a new modality for the analysis if it has relevant and significant information with respect to existing modalities. The proposed algorithm starts with the two most relevant modalities, and the remaining modalities are added sequentially according to their relevance. The optimum regularization parameters for the proposed algorithm are estimated based on the supervised information of sample categories. The effectiveness of the proposed algorithm, along with a comparison with state-of-the-art multimodal data integration methods, is established on several real-life multiblock data sets.

Both the ReDMiCA algorithm presented in Chapter 5 and the SeFGeIM algorithm presented in Chapter 6 are based on the sum of correlations (SUMCOR) criterion of the MCCA. The SUMCOR is an NP-hard problem, whereas the maximum variance (MAXVAR) criterion of the MCCA reduces the number of constraints that are associated with SUMCOR to a single constraint. Thus, MAXVAR provides a conceptually simple algebraic solution, which reduces the computational cost. Also, in real-life high-dimensional data analysis, the geometry of the multi-view data can provide structural information about the data sets, which facilitates efficient extraction of significant and relevant features from a multi-view source. However, the MCCA based approaches, namely, ReDMiCA and SeFGeIM, do not exploit the geometry of the data set. In this regard, a new supervised feature extraction algorithm, termed as GraDiM, is presented in Chapter 7, which integrates dynamic multi-view data sets by using the MAXVAR criterion and the knowledge of the graph. The proposed algorithm is dynamic in nature, that is, it incrementally updates the existing solutions, whenever a new view is available for the analysis. On the other hand, the algorithm is designed in such a way that if all the views are present at the beginning of the data analysis, the algorithm starts with the three most relevant modalities, and the remaining modalities are added sequentially according to their relevance. The proposed GraDiM algorithm addresses the singularity issue of the covariance matrices by using the ridge regression optimization technique. The optimum regularization parameters for the proposed algorithm are estimated based on the supervised information of sample categories. An analytical formulation demonstrates that the proposed algorithm can generate the required number of relevant and significant features from multi-view dynamic data sets, without extracting all possible features. In fact, all the views may not be required to extract different features. If the new view has relevant and significant information with respect to earlier views, then only the new view is incorporated in the integration process. The effectiveness of the proposed multi-view data integration algorithm, along with a comparison with other existing algorithms, is demonstrated on several benchmarks and real-life cancer data sets.

Finally, the thesis is concluded in Chapter 8, where the future directions and improvements of the proposed research work are also discussed.

Chapter 2

Survey on Multi-View Data Analysis

This chapter presents the basic notions of multi-view data analysis, along with a brief literature survey.

2.1 Multi-View Data

Recent years have spotted a growing interest in searching for various complementary data associated with a specific problem. Different data sources are likely to contain distinct and thus partly independent information. Combining those complementary pieces of information can be expected to enhance the total information about the problem. The effective integration and utilization of multiple data sources become an increasingly important problem in many applications. On the other hand, unimodal-based pattern recognition systems usually provide insufficient pattern representation due to the radical variation and noisy nature of the acquired signals. Combining data derived from multiple sources has the potential to significantly increase the intrinsic characteristics of the pattern, which leads to improved system performance compared to a single modality [151]. For example, a large number of diverse complementary biomedical data streams are being routinely acquired as part of the standard clinical workflow for patients. This research area leads to the direction of the future of personalized medicine, which will be dependent on leveraging the vast amount of medical data available to us to predict better treatments for patients. The integration of orthogonal features from a wide range of modalities can result in better predictors of disease aggressiveness and patient outcome, compared to any individual modality [41, 134, 190, 207]. In this background, there has been an increasing interest in data integration methods in recent times, both for supervised learning [92, 146, 151, 166] and unsupervised learning [6, 121, 224, 233]. Throughout the thesis, the term "modality" and "view" are used interchangeably. Thus, the "multimodal data set" is also mentioned as a "multi-view data set".

The modalities or views are represented in either relational form or feature vector-based form. There are $\mathcal{M}$ matrices $\{\mathcal{X}_i \in \Re^{m_i \times n}\}_{i=1}^{\mathcal{M}}$ to represent $\mathcal{M}$ modalities in feature vector-based representation, where n is the number of samples and m_i denotes the dimension of the i-th modality. Each matrix $\mathcal{X}_i$ may have different scale, unit, variance, dimension, and data distribution. On the other hand, $\mathcal{M}$ similarity matrices $\{W_i \in \Re^{n \times n}\}_{i=1}^{\mathcal{M}}$ represent $\mathcal{M}$

views in the case of relational data.

The easiest process to analyze the information of multi-view data sets using traditional machine learning algorithms is to join all multiple views into one single view. But, the naive integration of different views can create a concatenated feature set, which intensifies the "curse of dimensionality" problem [225]. Any feature extraction methods, like principal component analysis (PCA) [111] or linear discriminant analysis (LDA) [78], can be applied to this concatenated feature set. Both PCA and LDA are dimensionality reduction methods that jointly project the different attribute vectors into a low dimensional space of eigenvectors. However, PCA has two limitations, firstly, it assumes that the individual data streams lie on a linear manifold or subspace, and secondly, the data which has more variation may dominate other multidimensional datasets. On the other hand, LDA works efficiently when the assumption of equal population covariance structures for classes is satisfied.

The combination of the data interpretations approach [225], which is dependent on the decisions obtained from classifiers, is also used as another way of data fusion. The similarity matrices created by multiple data clusterings are used to combine various types of information in the evidence accumulation [80] problem. The information used to combine different views is lost due to the conversion of an input feature vector to a class label or decision attribute. Thus, the integration of interpretations approach may not be sufficient to establish a relation between different modalities [154, 270].

2.2 Multi-View Data Integration Approaches

Conventional machine learning algorithms, such as support vector machines, kernel machines, spectral clustering, and discriminant analysis concatenate all multiple views into one single view to adapt to the learning setting. However, this naive integration of different views intensifies the "curse of dimensionality" problem [225]. The small number of training samples causes the overfitting of the model. As each view has a unique statistical property, naive integration of multiple views does not have any physical meaning. Moreover, multi-view learning judiciously integrates relevant and non-redundant views by discarding noisy ones. Hence, it has been receiving increased recognition in recent years. The existing algorithms can be roughly grouped into five categories, namely, subspace learning, multiple kernel learning, co-training, embedding, and deep multi-view learning.

2.2.1 Subspace Learning

The main objective of subspace learning-based methods is to obtain a latent subspace shared by multiple modalities, where each input view can be generated from this latent subspace. The subspace learning effectively addresses the "curse of dimensionality" problem, as the latent subspace has lower dimensionality than that of any input view. Canonical correlation analysis (CCA) [112] finds linear relationships between two multidimensional views. It obtains two-directional weight vectors, also termed as basis vectors, and the empirical correlation between the respective projections onto these weight vectors is maximum. The CCA has been widely applied in many important scientific fields, such as brain MRI data analysis [194, 218], integration of omics data [91, 174], imaging genomics [116, 139], facial expression recognition, text mining [62, 157, 323] and image retrieval [88, 96]. There are several variants of CCA that exist in the literature, which are discussed below.

- **Regularized CCA:** CCA suffers from a computational issue due to a large number of features and the relatively small number of samples present in real-life data sets. The maximum correlation is 1 when the dimension of the feature is large. Thus, the recovering of canonical subspaces is not possible. Moreover, when the dimension of the features increases, all the features become highly correlated. This leads to ill-conditioned covariance matrices of different views. Because of this reason, their inverses are no longer reliable, resulting in an invalid computation of CCA and an unreliable meta-space. Regularized CCA (RCCA) [93, 156, 278] addresses this colinearity issue of each view by considering an adaptation of the ridge regression model to CCA. Moreover, real-life data sets are often plagued with noise. RCCA is used to correct these noises.

- **Constrained CCA:** In constrained CCA, some penalties are added to the basis vectors. According to the problem statement, these penalties are added to either one of the basis vectors or both of them. Thus, the constrained CCA problem can be formulated in terms of the constrained optimization problem. As analytical solutions do not exist, some numerical solutions through iterative optimization techniques are used to solve the constrained optimization CCA problem. Multiple optimization techniques, such as, the augmented-Lagrangian algorithm, sequential quadratic programming, Broyden-Fletcher-Goldfarb-Shanno algorithm, and reduced gradient method can be applied. In [309, 327], the solving of constrained CCA problems through optimization techniques is reported. Constrained CCA has been used to establish the association between neuropsychological, behavioral, or clinical data with brain imaging data in [59, 95]. Some relations between brain imaging data and task design have been established in [52, 53, 66, 82, 327, 328] using constrained CCA.

- **Sparse CCA:** The L1-norm penalty added to one of the basis vectors is the most frequently used penalty in constrained CCA. As the L1-norm penalty instigates sparsity on canonical coefficients, this constrained optimization problem is termed as sparse CCA. It performs reasonably with high-dimensional co-linear data by removing non-informative features. The penalty function working on individual views forms the element-level sparse CCA [257, 294], while the penalty function acting on the data group structure produces the group-level sparse CCA [46, 159, 160, 316, 322]. Sparse CCA can be further modified to structure sparse CCA, according to the known prior information about observations or features, such as characterizing connections between features [139] or categorizing features into different groups [159]. Although sparse CCA is widely used in emotion recognition [322], data fusion [195], and data clustering [42], originally it has been introduced to analyze omics data where the number of features is very large compared to the number of samples [208, 209, 294, 295]. In [322], the group sparse CCA has been used to select electroencephalogram (EEG) channels and to recognize EEG-based emotion. To preserve the spatial structure of images, a multimodal data fusion model has been developed in [195] by using structured and sparse CCA. A sparsity-aware CCA framework has been introduced to cluster sensor measurements [42].

- **Discriminant CCA:** CCA is generally unsupervised in nature, it does not take class information during the learning process. To learn the correlation matrix, discriminative CCA uses the label information by utilizing intra-class and inter-class similarity, thus it enhances the classification performance [15,70,119,252]. According to whether the local scattering is explored, the discriminative CCA models can be divided into two groups, namely, the local discriminative CCA [215,241,330] and the global discriminative CCA [140,255,256]. A discriminant model has been developed in [15] to address the face-sequence matching problem. In [252], a deep learning based multi-view linear discriminant analysis algorithm has been introduced, where both between-view and within-view class structures are preserved. In [70], a deep discriminative CCA has been proposed to classify speech-based emotion data. To preserve the inter-class and intra-class discriminative structure, a CCA based local discriminant embedding model has been developed in [119].

- **Kernel CCA:** If the two input views are non-linear, the correlation coefficient tends to be small, as the classical CCA finds the linear combination of input views. The kernel CCA (KCCA) [8] is an extension of the classical linear CCA to a general non-linear setting via a kernelization procedure. It maps the non-linear views into a higher dimensional Hilbert feature space. There are two types of KCCA that exist in the literature, namely, regularized KCCA [30,102] and non-regularized KCCA [118,323,326]. The KCCA has been widely applied in many important scientific fields, such as speech recognition [16], domain adaption [188], public surveillance [162,163], and neuroscientific field [28]. In [16], an incremental singular value decomposition approach has been introduced that makes computations of KCCA feasible to recognize the phonetic frame with typical speech data size. To address the domain adaption problem with semi-paired data, a regularized semi-paired KCCA model has been proposed in [188]. The problem of person re-identification in multicamera networks is addressed in [163], where the exponential KCCA model has been used. An algorithm based on KCCA that computes a multivariate temporal filter that links the correlation between brain activity and functional magnetic resonance imaging has been proposed in [28] to address the dynamic time-delay problem. The KCCA has been further developed in the finite sample and consistency analysis [34,75,86,101].

- **Multiset CCA:** Multiset canonical correlation analysis (MCCA) [110] extends the CCA for more than two views by finding a linear subspace that maximizes the correlations among all the views. Based on the definition of cross-view correlation for multiview learning, the MCCA models can be divided into two groups, namely, pairwise-correlation or zero-order-correlation based models [17, 37, 83, 110, 135, 229, 276] and high-order-correlation based models [169].

  1. **Pairwise-Correlation or Zero-Order-Correlation Based Models:** While correlation-based MCCA methods [109, 110, 135] consider only between-block information, covariance-based methods [61, 99, 100, 262, 263] take into account both the between-block and within-block information. There are several ways to measure the correlation in multi-view learning, such as maximization of the sum of all elements (SUMCOR) or the sum of squares of all elements (SSQCOR) in the correlation matrix, maximization of the largest eigenvalue (MAXVAR) or

minimization of the smallest eigenvalue (MINVAR) of the correlation matrix, minimization of the determinant (GENVAR) of the correlation matrix [110,135]. Another criterion, namely, sum of absolute value correlations (SABSCOR), has also been considered in [99]. The sum of covariance (SUMCOV) criterion has been proposed in [61]. In [100], sum of squared covariance (SSQCOV) criterion has also been introduced. Some modifications of SUMCOR, SSQCOR, and SABSCOR have been proposed in [262], which take into account some hypotheses on the connections between sets of variables. The sum of absolute value covariances (SABSCOV) has also been considered in this article. The SUMCOR, MAXVAR, SSQCOR, MINVAR, GENVAR, and SABSCOR are based on maximizing a function of the correlation between canonical variates [263], while the analysis of SUMCOV, SSQCOV, and SABSCOV is based on covariance between canonical variates [263].

2. **High-Order-Correlation Based Models:** A tensor is the extension of matrix factorization in multi-view data analysis. It is used to capture higher-order correlations among multiple views [47,297,298]. Tensor based generalization of CCA (TCCA) for more than two views has been introduced in [169]. Instead of calculating the pairwise correlation matrix, TCCA estimates the correlation of all views by constructing a covariance tensor.

MCCA has been extended into deep learning framework [24], probabilistic model [57,144,280], and kernel approach [18,228].

- **Probabilistic CCA:** The classical CCA model provides a linear algebraic solution, while the probabilistic CCA approach has a probabilistic interpretation of that solution [19,33]. In [19], a theoretical analysis has been given, which proves that posterior expectations of the maximum likelihood estimation for a latent variable are identical to the subspaces derived from CCA. Based on the observation reported in [19], several extensions are done using prior distributions of Bayesian analysis [141–143,279,281,283]. There are two main contributions associated with Bayesian analysis, these models are robust toward small sample size and modification is easy when the assumption of distribution is changed. The algorithms proposed in [141,283] introduced the automatic relevance determination model using inverse Wishart distribution. The probabilistic CCA is widely used in biomedicine data analysis, such as prioritization of cancer genes [143], a study of drug responses [281], and analysis of rare diseases [121]. A dynamic probabilistic CCA model has been proposed in [204] to identify the temporal dependencies on latent subspaces both for individual and shared information of views. In [325], a bilinear extension of probabilistic CCA has been reported for photo-sketch and face matching.

- **Locality Preserving CCA:** The locality preserving CCA (LPCCA) has been proposed in [254], where similarity matrices are incorporated in CCA to identify the local manifold structure. The basic idea of LPCCA is that the data points are closed in the low-dimensional projected subspace if they are close enough in the input high-dimensional space [130,254,268,284,305]. According to the analysis, LPCCA can be divided into two groups. One of them obtains a local neighbor graph by providing cross-correlation information between neighbors [268,284], while the other group discards the trivial correlation between non-neighbors and provides a local manifold structure [130, 254, 305]. In [305], a supervised LPCCA model has been developed to improve the classification performance. The neighborhood information is incorporated to improve the robustness of the model in [284].

- **Deep CCA:** Both LPCCA and KCCA are non-linear extensions of CCA, but their representation is bounded to either local information or a predefined kernel. Deep CCA (DCCA) obtain more complex non-linear transformations of different views by passing them through a deep network, such as, convolutional neural networks (CNN) [306, 307], auto-encoder [38, 287], and deep neural networks [14, 168, 288, 289, 304]. In recent years, DCCA has achieved immense success in representation learning [25], cross-domain retrieval [238], word embedding [168], and image annotation [196]. In [25], generalized CCA is combined with DCCA to make a deep generalized CCA model. It has been applied for three tasks, namely, articulatory measurements, phonetic transcription, and information recommendation for Twitter users. A hypergraph regularizer-based DCCA model has been developed in [238], where the image-to-text or text-to-image retrieval problem is addressed. A multilingual non-linear correlation problem has been addressed in [168] using the DCCA model to improve the standard of word embeddings. In [196], a DCCA-based model has been developed to study the image-tag annotation problem, where the tag is generated using Word2Vec network [191] and deep CNN is used to extract features from the image. A deep MCCA algorithm has been proposed in [244], where feed-forward networks have been used to map the input views to a shared subspace. In [55], a DMCCA model has been introduced that focuses on task-driven objectives using CCA.

The CCA has also been used in multi-view regression [127] and clustering [39] fields. A generalization of Fisher's discriminant analysis has been proposed in [64] to explore the latent subspace spanned by a multimodal data set. This generalization is supervised although CCA does not incorporate the class information. Multi-view metric learning [219, 313] has been developed to construct projections from multi-view data. The latent subspace is used to infer another view from the observation view. To establish the connections between the two views through latent subspaces, the Markov network [45], maximization of mutual information [189], and Gaussian process [242] have been used. In [124, 230], a latent subspace is used to factorize private and shared information from different views. The main objective of factor analysis is to obtain latent factors, which summarize the input data. Inter-battery factor analysis (IBFA) [273], a model closely related to CCA, extends this notion in multi-view learning. In recent years, several multi-view learning algorithms have been developed based on IBFA [58, 60, 123, 234].

Partial least squares (PLS) [296] is another popular statistical technique that has been used to find fundamental relations between two views. There exist three types of PLS methods in literature, namely, partial least squares correlation (PLSC) [329], partial least squares regression (PLSR) [51, 226], and partial least squares path modeling (PLS-PM) [264]. PLSC is a correlational model that analyzes associations between two views, while PLSR is a regression model that predicts one view from another. On the other hand, PLS-PM is a variance-based structural equation modelling that can be used to model complex relationships among different views.

2.2.2 Multiple Kernel Learning

The main objective of multiple kernel learning (MKL) is to control the search space capacity of possible kernel matrices to achieve good generalization. The kernels in MKL correspond to different views, and the integration of different kernels may improve the learning performance. Thus, MKL is widely used to analyze multi-view data sets. Over the past few years, MKL has become one of the important techniques to analyze multi-view data sets. It achieves attention due to the utilization of various optimization techniques [7,11,152,245] as well as the recognization ability by exploring possible combinations of base kernels [145,275,301,324]. MKL techniques are further extended to several models, such as, localized MKL [98], sample-adaptive MKL [167], Bayesian MKL [67], multiple empirical kernel learning [73,292], two-stage MKL [54,200,285,286], and function approximation MKL [147,240]. In [152], MKL has been formulated as a semi-definite programming problem. MKL is used to develop a dual formulation of the quadratically-constrained quadratic program as a second-order cone program problem in [20], where a sequential minimal optimization algorithm has been developed to efficiently obtain the optimal solution. Some efficient semi-infinite linear programs have been proposed in [245,246], where MKL addresses large-scale problems.

2.2.3 Co-training

Co-training [31] is one of the earliest models to integrate multimodal data. It learns alternately by maximizing the mutual correspondence between two unlabeled views. There are many modifications which have been done in the recent past. In [206], generalized expectation-maximization has been done, where adjustable probabilistic labels are assigned to unlabeled data. Some robust semi-supervised learning algorithms have been proposed in [197–199], where active learning is combined with co-training. In [311,312], Bayesian undirected graphical models are developed for co-training and a novel co-training kernel for Gaussian process classifiers. A graph-based and disagreement-based semi-supervised learning has been proposed in [290], where the co-training process is viewed as a combinative label propagation over two views. In [243], a co-regularization framework has been introduced where classifiers are learned in each view through forms of multi-view regularization. Some co-training based multi-view clustering algorithms have been proposed in [26, 148, 149]. There are mainly three reasons behind the success of co-training algorithms, namely, each view is self-sufficient to classify the patterns properly, there is a high probability that both the views predict the same labels, and each view is conditionally independent given the label. But, in real-world applications, it is very difficult to satisfy the conditional independence of views. Hence, several weaker alternatives have been proposed in [4, 21, 291].

2.2.4 Embedding

To overcome the representational differences, an alternative transformed representation has to be created for each view. An algorithm has been proposed in [154], where each view is projected into a homogeneous meta-space. The dimension of the projected space is the same for each view with the same scale. In [282], consensus embedding has been introduced, where according to the minimum predictive value, embeddings selected from different views have to be combined. The boosted embedding concatenation has been reported in [270], where supervised information is used in the fusion process. In [81], an algorithm of boosted embedding concatenation has been proposed based on the Adaboost classifier, which evaluates and provides weight on each embedding to integrate different views. However, these methods may provide redundant and noisy latent features, which deteriorate the final outcome [91]. High-dimensional kernels are also used to combine embeddings of different views [150].

2.2.5 Deep Multi-View Learning

Due to the powerful feature extraction capability, deep learning methods have gained attention in recent years. By using multiple hierarchical layers, deep learning models can learn non-linear, subtle, complex, and abstract representations of the target data from multiple views. Several deep multi-view learning algorithms exist in the literature, such as multi-view convolutional neural network [77,132,153,186,249,308], multi-view auto-encoder [76, 108, 203, 317], multi-view generative adversarial network [65, 117, 267, 271], multi-view graph neural network [74,103,138,302], multi-view deep belief net [9,12,247,259,315], and multi-view recurrent neural network [3, 231]. Apart from CCA other conventional learning methods are also extended into the deep framework, such as, deep multi-view matrix factorization [318], deep multi-view spectral learning network [120], and deep multi-view information bottleneck [10].

- **Multi-View Convolutional Neural Network:** In [308], a hybrid framework of multi-view convolutional neural network and extreme learning machine auto-encoder has been proposed to learn features for classification and retrieval of three-dimensional objects. Another three-dimensional multi-view convolutional neural network has been introduced in [132], which is based on the multi-view-one-network strategy. Both directed acyclic graph architecture and chain architecture are used including three-dimensional Inception-ResNet and three-dimensional Inception. A neurophysiologically inspired multi-view convolutional neural network has been proposed in [186] to classify motor imagery from electroencephalography signals. A group-view convolutional neural network has been reported in [77], where hierarchical view-group-shape architecture is used to identify three-dimensional objects.

- **Multi-View Auto-Encoder:** A model involving correspondence auto-encoder has been proposed in [76] for cross-modal retrieval problems. The human pose recovery problem based on video has been addressed in [108], where a multi-layered deep neural network has been used to construct low-rank hypergraph Laplacian. A discriminative margin-sensitive auto-encoder has been introduced in [317] to diagnose Alzheimer's disease and for recognition of protein folds accurately. In [203], a novel bimodal auto-encoder has been proposed to reconstruct both video and audio views.

- **Multi-View Generative Adversarial Network:** In [271], a disentangled representation learning-generative adversarial network has been proposed to synthesize images. The encoder-decoder structure of the generator makes the architecture effective. Generally, a generative adversarial network does not bother to learn the inverse mapping, but a bidirectional generative adversarial network has been reported in [65], where inverse mapping has been done. A two-pathway generative adversarial network has been proposed in [267] to preserve the completeness of the learned embedding space. In [117], another two-pathway generative adversarial network has been introduced where two distinct encoder-decoder structures have been used to capture both local and global information.

- **Multi-View Graph Neural Network:** A multi-view learning algorithm using a graph neural network followed by a multi-layer perceptron has been introduced in [103]. In [74], a novel task-guided multi-view graph auto-encoder clustering framework has been reported, which can learn node embeddings by applying the content information. To analyze the global poverty problem, a graph structure based on the convolutional network has been proposed [138]. This model can be applied to predict whether a person is living below the poverty line, to predict the adoption of economic inclusion, or to predict the gender of mobile phone subscribers. A redesigned graph neural network, collaborated with a convolutional neural network, has been introduced in [302], to obtain a feature representation of multi-view images.

- **Multi-View Deep Belief Net:** The deep belief net (DBN) [106] adopts the restricted Boltzmann machine (RBM) as its fundamental component. A hybrid model based on RBM has been reported in [12], and cross-modality as well as inter-modality features are extracted to detect the sequential event. A multi-view face recognition approach has been proposed in [9] based on DBN to capture the complementary representation of deep and local features. In [247], a multimodal deep Boltzmann machine has been introduced to learn a joint density model over the space of multi-view data set. Another multimodal deep Boltzmann machine algorithm has been proposed in [259], where several patient phenotypes and gene expression data are processed simultaneously to identify the importance of different genes. In [315], a multi-view DBN has been introduced, where RBM is used to model each view.

- **Multi-View Recurrent Neural Network:** The recurrent neural network (RNN) [258] is used to deal with the time series data. In [3], a multi-view RNN model has been presented to address the indoor scene recognition problem. An algorithm based on multi-view RNN has been proposed in [231], which detects the wake and sleep state of a person by analyzing the data generated from his/her smartphones and wearable technologies.

2.3 Conclusion

One of the important challenges associated with multi-view data integration is to extract the most relevant and significant set of features from multiple views. In this context, the next chapter presents a novel algorithm that judiciously integrates the merit of supervised regularized CCA and the theory of rough sets, to extract a set of new features from two views by maximizing their relevance with respect to the class labels and significance among them.

Chapter 3

Supervised Canonical Correlation Analysis Using Max Relevance-Max Significance Criterion

3.1 Introduction

In present days, there is a scope of getting complementary multiple data corresponding to a given problem or task, and the main challenge is to extract features, which are most relevant, significant, and nonredundant for the given problem. The effective utilization and integration of multiple data sources or multimodal information are becoming an increasingly important problem in many applications. Due to the noisy nature and drastic variation of the acquired signals, unimodal based pattern recognition and analysis systems usually provide low level of performance, which leads to inaccurate and insufficient pattern representation of the perception of interest. On the other hand, multimodal data contains more information. The integration of multimodal data is expected to provide potentially a more discriminatory and complete description of the intrinsic characteristics of the pattern, which leads to improved system performance compared to a single modality [151].

The simultaneous analysis of multimodal data is an important task in integrative systems biology approach, which gives a better understanding of the relationships among different biological functional levels [321]. For example, integration of heterogeneous omics data, namely, transcriptomics, metabolomics, and proteomics, may provide a better understanding of biological systems. The Cancer Genome Atlas (TCGA) (https://cancergenome.nih.gov/) helps to provide multiple types of data from the same individual. In TCGA, gene and microRNA expression arrays, copy number variation, DNA methylation data, and protein expression array are obtained from most of the tumor samples. By using multiple types of data of unique samples, it is possible to make the linkages between attributes within each type of data. It maximizes the information content and makes a model, which uses all the available data. It is intrinsically more powerful than the models that use only single data type. Given this background, there has been an increasing interest in data integration methods in biomedical sciences, both for supervised learning [92,129,146,151,166] and unsupervised learning [6, 121, 224, 233].

Canonical correlation analysis (CCA) [112] provides an efficient way of measuring the linear relationship between two multidimensional data sets. For two multidimensional variables, it finds the best linear transformation to achieve the maximum correlation between them. The CCA has been widely applied in many important scientific fields, such as brain MRI data analysis [194,218], integration of omics data [91,174], imaging genomics [116,139], facial expression recognition, text mining [62, 157, 323] and image retrieval [88, 96]. In recent years, some variants of CCA, such as generalized CCA [216], kernel CCA [323], sparse CCA [49], and locality preserving CCA [254] have also been developed. The CCA is also popular for integration of different omics data [35]. To map genes or proteins onto the Euclidean space, kernel CCA has been used in [303]. On the other hand, sparse CCA has been used in [36, 160] to study the mutual relation among different types of omics data. Besides the integration of two data sets, CCA can help to analyze gene expression dynamics geometrically [223]. Phylogenetic CCA [222], another variant of CCA, gives continuous valued character data obtained from biological species related by a phylogenetic tree. Hence, CCA can be used to capture the underlying genetic background of a complex disease, by associating two data sets containing information about a patient's phenotypical and genetic details. It gives those relevant variables or features from both data types, which are related to each other and provide more insight into the biological experimental hypotheses.

However, CCA suffers from a computational issue due to 'large p (number of features) and small n (number of samples)'. Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be two multivariate data sets having m_1 and m_2 number of features, respectively, and n is the number of samples in both $\mathcal{X}_1$ and $\mathcal{X}_2$. The features in $\mathcal{X}_1$ and $\mathcal{X}_2$ tend to be highly collinear if $n << m_1$ and $n << m_2$. This leads to ill-conditioned covariance matrices of $\mathcal{X}_1$ and $\mathcal{X}_2$, that is, $\mathcal{C}_{11}$ and $\mathcal{C}_{22}$. Because of this, their inverses are no longer reliable, resulting in an invalid computation of CCA and an unreliable meta-space. The covariance matrices $\mathcal{C}_{11}$ and $\mathcal{C}_{22}$ will be invertible if $n \geqslant m_1 + m_2 + 1$ [68]. However, this condition is usually not possible in the bioinformatics domain, where number of samples 'n' is usually limited. On the other hand, modern technology has enabled very high dimensional data streams to be routinely acquired, which results in very high dimensional feature spaces m_1 and m_2. To overcome this problem, a regularized version of CCA has been introduced in [94]. Regularized CCA (RCCA) [93,278] is an improved version of CCA. It uses a ridge regression optimization scheme to prevent over-fitting of insufficient training data [27]. It works by adding small positive quantities to the diagonals of $\mathcal{C}_{11}$ and $\mathcal{C}_{22}$ to guarantee their invertibility [107]. The RCCA has been successfully used to study gene expressions in liver cells and compare them with concentrations of hepatic fatty acids in mice [93]. Regularized sparse CCA is used in expression quantitative trait loci to detect genetic loci mapped to a disease [133]. However, RCCA is computationally very expensive because of this regularization process. Also, both CCA and RCCA are unsupervised in nature and fail to take complete advantage of available class label information.

To perform the regularization process, supervised RCCA (SRCCA) uses a supervised feature selection algorithm [91]. The available class label information is included in SRCCA to select maximally correlated features. In SRCCA, regularization is done by considering the most discriminatory score of the first pair of canonical variables, based on a feature selection method, and then the remaining dimensions are adjusted [91]. One of the important applications of SRCCA in functional genomics is to classify samples, such as to classify cancer versus normal samples or to classify different types or subtypes of cancer, according to the maximally correlated features or biomarkers. The SRCCA also helps in developing diagnostic tools for delivering precise, reliable, and interpretable results. With the supervised feature selection results, the cost of biological experiment and decision can be greatly reduced by analyzing only maximally correlated relevant biomarkers. However, existing SRCCA considers only correlation of the first pair of canonical variables. But, it may happen that other canonical variable pairs have insignificant relation with the first pair of canonical variables, or there may be some irrelevant features in the whole extracted feature set, which should not be considered in further processing [173].

In integrative omics data analysis, another important problem is uncertainty. This uncertainty may arise from vagueness in response variables of samples and imprecision in computations. The *t*-test, Wilcoxon rank sum test or Wilks's lambda test, used to capture supervised class information in existing SRCCA [91], are unable to handle this uncertainty. To model and propagate this uncertainty, the theory of rough sets has become successful, which can deal with incompleteness and vagueness [214]. It is proposed for indiscernibility in classification according to some relation and acts as an effective means for dimensionality reduction of discrete valued data [176]. Rough set theory has also been used for analyzing omics data [171, 175–179, 211, 212]. Usually, there are continuous valued data in real world applications. In rough set theory, the continuous valued features are divided into several discrete partitions for feature selection. However, the inherent error that exists in the discretization process is of major concern in the feature selection. The hypercuboid equivalence partition matrix [172] of rough hypercuboid approach is found to be suitable for feature selection of numerical data. It has been applied successfully for analyzing omics data [173, 174, 210].

In this regard, this chapter presents a new feature extraction algorithm, termed as CuRSaR (CCA using maximum Relevance-maximum Significance criterion and Rough sets), from two multidimensional data sets. It judiciously integrates the merits of SRCCA and the theory of rough sets. The proposed algorithm extracts a set of new features by maximizing their relevance with respect to the class labels and significance among them [177]. Both the relevance and significance measures are computed based on the concept of hypercuboid equivalence partition matrix of rough hypercuboid approach [172]. In the proposed algorithm, the regularization parameters do not only depend on the first pair of canonical variables, rather the whole extracted feature set is considered to optimize the regularization parameters. An analytical formulation is presented to establish the relation between the covariance matrices of different regularization parameters, which makes the computational cost of the proposed algorithm significantly lower than that of the existing algorithms. The effectiveness of the proposed algorithm, along with a comparison with other algorithms, is demonstrated on several real-life cancer data sets. Some of the results of this chapter are reported in [174, 182].

The rest of this chapter is organized as follows: Section 3.2 outlines the basic principles of CCA, RCCA, and SRCCA. Section 3.3 presents the proposed algorithm. A theoretical analysis is presented in this section to establish the relation between the covariance matrices of different regularization parameters, which drastically reduces the computational complexity of existing RCCA. The effectiveness of the proposed data integration algorithm, along with a comparative performance analysis with state-of-the-art algorithms on different data sets, is presented in Section 3.4. Concluding remarks are provided in Section 3.5.

3.2 Basics of Canonical Correlation Analysis and its Variants

This section presents the fundamental concepts in the theories of CCA, RCCA, and SR-CCA.

3.2.1 CCA: Canonical Correlation Analysis

Canonical correlation analysis (CCA) [112] obtains a linear relationship between two multidimensional variables. The objective of CCA is to extract latent features from two data sets $\mathcal{X}_1 \in \Re^{m_1 \times n}$ and $\mathcal{X}_2 \in \Re^{m_2 \times n}$, which are most correlated, where each column in $\mathcal{X}_1$ and $\mathcal{X}_2$ corresponds to one of the n samples, and each row represents one variable. Let us assume that each variable is centered to have zero mean across the samples. CCA obtains two directional weight vectors, also termed as basis vectors, $w_1 \in \Re^{m_1}$ and $w_2 \in \Re^{m_2}$ such that the empirical correlation between the respective projections onto these weight vectors, that is, between $\mathcal{X}_1^T w_1$ and $\mathcal{X}_2^T w_2$ is maximum. The correlation coefficient $\tilde{\rho}$ is given as follows:

$$\tilde{\rho} = \max_{w_1, w_2} \frac{\mathcal{E}\left[w_1^T \mathcal{X}_1 \mathcal{X}_2^T w_2\right]}{\sqrt{\mathcal{E}\left[w_1^T \mathcal{X}_1 \mathcal{X}_1^T w_1\right]}\sqrt{\mathcal{E}\left[w_2^T \mathcal{X}_2 \mathcal{X}_2^T w_2\right]}} = \max_{w_1, w_2} \frac{w_1^T \mathcal{C}_{12} w_2}{\sqrt{w_1^T \mathcal{C}_{11} w_1 w_2^T \mathcal{C}_{22} w_2}} \tag{3.1}$$

where $\mathcal{E}[f]$ denotes empirical expectation of function f, $\mathcal{C}_{12} \in \Re^{m_1 \times m_2}$ is the cross-covariance matrix of $\mathcal{X}_1$ and $\mathcal{X}_2$, which is given as follows:

$$[\mathcal{C}_{12}]_{m_1 \times m_2} = [\mathcal{X}_1]_{m_1 \times n}[\mathcal{X}_2^T]_{n \times m_2}; \tag{3.2}$$

while $\mathcal{C}_{11} \in \Re^{m_1 \times m_1}$ and $\mathcal{C}_{22} \in \Re^{m_2 \times m_2}$ are the covariance matrices of $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively, and are as follows:

$$[\mathcal{C}_{11}]_{m_1 \times m_1} = [\mathcal{X}_1]_{m_1 \times n}[\mathcal{X}_1^T]_{n \times m_1}; \tag{3.3}$$

$$[\mathcal{C}_{22}]_{m_2 \times m_2} = [\mathcal{X}_2]_{m_2 \times n}[\mathcal{X}_2^T]_{n \times m_2}. \tag{3.4}$$

Since $\tilde{\rho}$ is invariant to the scaling of w_1 and w_2, CCA can be formulated equivalently as

$$\max_{w_1, w_2} \quad w_1^T \mathcal{C}_{12} w_2;$$

$$\text{subject to} \quad w_1^T \mathcal{C}_{11} w_1 = 1; \quad \text{and} \quad w_2^T \mathcal{C}_{22} w_2 = 1. \tag{3.5}$$

To calculate w_1 and w_2, the eigenvectors of $\Sigma\Sigma^T$ and $\Sigma^T\Sigma$ are needed, where the matrix $\Sigma \in \Re^{m_1 \times m_2}$ is given as follows:

$$\Sigma = \mathcal{C}_{11}^{-1/2} \mathcal{C}_{12} \mathcal{C}_{22}^{-1/2}. \tag{3.6}$$

Without loss of generality, it is assumed that $m_1 \leqslant m_2$. Suppose $\rho_1 \geqslant \cdots \geqslant \rho_t \geqslant \cdots \geqslant \rho_{m_1}$ be the eigenvalues of $\Sigma\Sigma^T$ and $\xi_{1_1}, \cdots, \xi_{1_t}, \cdots, \xi_{1_{m_1}}$ are the orthonormalized eigenvectors corresponding to $\rho_1, \cdots, \rho_t, \cdots, \rho_{m_1}$. As non-zero eigenvalues of $\Sigma\Sigma^T$ are same as non-zero eigenvalues of $\Sigma^T\Sigma$ [89], either $\Sigma\Sigma^T$ or $\Sigma^T\Sigma$ is enough to calculate the eigenvectors. Furthermore, let say, $\rho_1 \geqslant \cdots \geqslant \rho_t \geqslant \cdots \geqslant \rho_{m_1}$ are the m_1 largest eigenvalues of $\Sigma^T\Sigma$ with orthonormalized eigenvectors $\xi_{2_1}, \cdots, \xi_{2_t}, \cdots, \xi_{2_{m_1}}$. Then, the t-th pair of basis vectors are given by

$$w_{1_t} = \mathcal{C}_{11}^{-1/2}\xi_{1_t}; \qquad \text{and} \qquad w_{2_t} = \mathcal{C}_{22}^{-1/2}\xi_{2_t}. \tag{3.7}$$

As ξ_{1_t} and ξ_{2_t} are the t-th eigenvectors of $\Sigma\Sigma^T$ and $\Sigma^T\Sigma$, respectively, with eigenvalue ρ_t, the characteristic polynomials of $\Sigma\Sigma^T$ and $\Sigma^T\Sigma$ can be written as

$$\begin{aligned}
\Sigma\Sigma^T\xi_{1_t} &= \rho_t\xi_{1_t}\\
\Rightarrow (\mathcal{C}_{11}^{-1/2}\mathcal{C}_{12}\mathcal{C}_{22}^{-1/2})(\mathcal{C}_{11}^{-1/2}\mathcal{C}_{12}\mathcal{C}_{22}^{-1/2})^T\xi_{1_t} &= \rho_t\xi_{1_t}\\
\Rightarrow \mathcal{C}_{11}^{-1/2}\mathcal{C}_{12}\mathcal{C}_{22}^{-1/2}\mathcal{C}_{22}^{-1/2}\mathcal{C}_{12}^{T}\mathcal{C}_{11}^{-1/2}\xi_{1_t} &= \rho_t\xi_{1_t}\\
\Rightarrow \mathcal{C}_{11}^{-1/2}\mathcal{C}_{12}\mathcal{C}_{22}^{-1}\mathcal{C}_{21}\mathcal{C}_{11}^{-1/2}\xi_{1_t} &= \rho_t\xi_{1_t}\\
\Rightarrow \mathcal{C}_{11}^{-1/2}\mathcal{C}_{11}^{-1/2}\mathcal{C}_{12}\mathcal{C}_{22}^{-1}\mathcal{C}_{21}\mathcal{C}_{11}^{-1/2}\xi_{1_t} &= \rho_t\mathcal{C}_{11}^{-1/2}\xi_{1_t}\\
\Rightarrow \mathcal{C}_{11}^{-1}\mathcal{C}_{12}\mathcal{C}_{22}^{-1}\mathcal{C}_{21}w_{1_t} &= \rho_t w_{1_t}; \qquad (3.8)\\
\text{and} \qquad \Sigma^T\Sigma\xi_{2_t} &= \rho_t\xi_{2_t}\\
\Rightarrow \mathcal{C}_{22}^{-1}\mathcal{C}_{21}\mathcal{C}_{11}^{-1}\mathcal{C}_{12}w_{2_t} &= \rho_t w_{2_t}. \qquad (3.9)
\end{aligned}$$

From (3.2.1) and (3.2.1), it can be seen that the basis vectors w_{1_t} and w_{2_t} are the eigenvectors of matrix $\mathcal{H}$ and $\tilde{\mathcal{H}}$, respectively, with eigenvalue ρ_t, where

$$\mathcal{H} = \mathcal{C}_{11}^{-1}\mathcal{C}_{12}\mathcal{C}_{22}^{-1}\mathcal{C}_{21}; \qquad \text{and} \qquad \tilde{\mathcal{H}} = \mathcal{C}_{22}^{-1}\mathcal{C}_{21}\mathcal{C}_{11}^{-1}\mathcal{C}_{12}. \tag{3.10}$$

The t-th pair of canonical variables $\{\mathcal{U}_{1_t}, \mathcal{U}_{2_t}\}$ is as follows:

$$\mathcal{U}_{1_t} = w_{1_t}^T\mathcal{X}_1; \qquad \text{and} \qquad \mathcal{U}_{2_t} = w_{2_t}^T\mathcal{X}_2. \tag{3.11}$$

Here, $\{\mathcal{U}_{1_1}, \mathcal{U}_{2_1}\}$ is the first pair of canonical variables, which provides the maximum correlation $\tilde{\rho} = \sqrt{\rho_1}$. The t-th pair of canonical variables $\{\mathcal{U}_{1_t}, \mathcal{U}_{2_t}\}$ is the linear combinations of t-th basis vectors and data set. It maximizes the correlation among all possible linear combinations and is uncorrelated with the previous $(t-1)$ canonical variable pairs. From (3.11), the t-th feature $\mathcal{F}_t$ is extracted as follows:

$$\mathcal{F}_t = \mathcal{U}_{1_t} + \mathcal{U}_{2_t}; \tag{3.12}$$

where $\forall t \in \{1, 2, \cdots, \mathcal{D}\}$ and $\mathcal{D} \leqslant \min(m_1, m_2)$.

3.2.2 RCCA: Regularized CCA

Real-life data sets are often plagued with noise. Regularized CCA (RCCA) [93, 156, 278] is used to correct these noises in $\mathcal{X}_1$ and $\mathcal{X}_2$. Let us assume that $\mathcal{X}_1$ and $\mathcal{X}_2$ are contaminated with Gaussian, independent and identically distributed noise $\mathcal{N}_1 \in \Re^{m_1 \times n}$ and $\mathcal{N}_2 \in \Re^{m_2 \times n}$. As these noises are Gaussian, independent and identically distributed, all possible combinations of the covariances of the m_1 and m_2 rows of $\mathcal{N}_1$ and $\mathcal{N}_2$, respectively, will be 0 except the covariance of a particular row vector with itself. Let the variances of each row of $\mathcal{N}_1$ and $\mathcal{N}_2$ be $\mathfrak{r}_1$ and $\mathfrak{r}_2$, respectively, which are known as regularization parameters. The cross-covariance matrix $\mathcal{C}_{12}$ of $\mathcal{X}_1$ and $\mathcal{X}_2$ will not be affected. But, the matrices $\mathcal{C}_{11}$ and $\mathcal{C}_{22}$ become $[\mathcal{C}_{11} + \mathfrak{r}_1 I]$ and $[\mathcal{C}_{22} + \mathfrak{r}_2 I]$, respectively, where I is the identity matrix of appropriate order. So, (3.10) becomes

$$\mathcal{H} = [\mathcal{C}_{11} + \mathfrak{r}_1 I]^{-1}\mathcal{C}_{12}[\mathcal{C}_{22} + \mathfrak{r}_2 I]^{-1}\mathcal{C}_{21}; \tag{3.13}$$

$$\text{and} \quad \tilde{\mathcal{H}} = [\mathcal{C}_{22} + \mathfrak{r}_2 I]^{-1}\mathcal{C}_{21}[\mathcal{C}_{11} + \mathfrak{r}_1 I]^{-1}\mathcal{C}_{12}. \tag{3.14}$$

In RCCA, the regularization parameters are varied in a certain range $\mathfrak{r}_{\mathit{min}} \leqslant \mathfrak{r}_1, \mathfrak{r}_2 \leqslant \mathfrak{r}_{\mathit{max}}$ and chosen by a grid search optimization technique [97]. Every pair of $\mathfrak{r}_1$ and $\mathfrak{r}_2$ will produce a pair of first canonical variables, which are maximally correlated. The optimal parameters $\mathfrak{r}_1$ and $\mathfrak{r}_2$ are considered for which the Pearson's correlation is maximum, that is,

$$\max_{\mathfrak{r}_1, \mathfrak{r}_2} \frac{w_1^T \mathcal{C}_{12} w_2}{\sqrt{w_1^T(\mathcal{C}_{11} + \mathfrak{r}_1 I)w_1 w_2^T(\mathcal{C}_{22} + \mathfrak{r}_2 I)w_2}}. \tag{3.15}$$

3.2.3 SRCCA: Supervised RCCA

Both CCA and RCCA are unsupervised in nature. They do not incorporate the information of class label or sample category even if it is present in the given data sets. To overcome this limitation of both CCA and RCCA, Golugula et al. [91] introduced the concept of supervised RCCA (SRCCA), which is a supervised version of RCCA. Similar to RCCA, SRCCA chooses the optimal regularization parameters $\mathfrak{r}_1$ and $\mathfrak{r}_2$ using grid search optimization by a feature selection method based on either t-test, Wilks's lambda test, or Wilcoxon rank sum test. The optimal regularization parameters are obtained by maximizing the discriminatory score of the feature corresponding to first pair of canonical variables, and then the remaining dimensions are extracted for the optimal parameters.

3.3 Proposed Method

This section presents a new feature extraction algorithm, termed as CuRSaR, integrating judiciously the information of two multidimensional data sets. Prior to describing the proposed algorithm for multimodal data analysis, some important analytical formulations are introduced next, which reduce the computational complexity of existing RCCA.

3.3.1 Covariance Matrices for Different Regularization Parameters

The spectral decomposition [269] can be used to calculate $[\mathcal{C}_{11}+\mathfrak{r}_1 I]^{-1}$ and $[\mathcal{C}_{22}+\mathfrak{r}_2 I]^{-1}$ for the computation of the $\mathcal{H}$ matrix of (3.13). The spectral decomposition can be described in terms of eigenvalue-eigenvector pairs of $[\mathcal{C}_{11}+\mathfrak{r}_1 I]$ and $[\mathcal{C}_{22}+\mathfrak{r}_2 I]$. An $m_1 \times m_1$ symmetric matrix $[\mathcal{C}_{11}+\mathfrak{r}_1 I]$ can be expressed in terms of its m_1 eigenvalue-eigenvector pairs (Λ_1, Ψ_1) as follows [269]:

$$[\mathcal{C}_{11}+\mathfrak{r}_1 I] = \Psi_1 \Lambda_1 \Psi_1^T = \sum_{i=1}^{m_1} \lambda_{1_i} \psi_{1_i} \psi_{1_i}^T; \tag{3.16}$$

where the i-th element λ_{1_i} of diagonal matrix Λ_1 denotes the i-th eigenvalue of the matrix $[\mathcal{C}_{11}+\mathfrak{r}_1 I]$. The i-th column of matrix Ψ_1 represents the orthonormalized eigenvector ψ_{1_i} corresponding to eigenvalue λ_{1_i}, $\forall i \in \{1, 2, \cdots, m_1\}$, and

$$\Psi_1 \Psi_1^T = \Psi_1^T \Psi_1 = I. \tag{3.17}$$

The computation of the inverse of matrix $[\mathcal{C}_{11}+\mathfrak{r}_1 I]$ is performed as follows [122]:

$$[\mathcal{C}_{11}+\mathfrak{r}_1 I]^{-1} = \Psi_1 \Lambda_1^{-1} \Psi_1^T = \sum_{i=1}^{m_1} \frac{1}{\lambda_{1_i}} \psi_{1_i} \psi_{1_i}^T. \tag{3.18}$$

In RCCA and SRCCA, the regularization parameters $\mathfrak{r}_1$ and $\mathfrak{r}_2$ are varied within a specified range $[\mathfrak{r}_{min}, \mathfrak{r}_{max}]$, where $\mathfrak{r}_{min} \leqslant \mathfrak{r}_1, \mathfrak{r}_2 \leqslant \mathfrak{r}_{max}$. It can be assumed that these regularization parameters follow an arithmetic progression. Each parameter starts with an initial value $\mathfrak{r}_{min}$. After every iteration, a constant value or a common difference is added with the previous value, and finally, it reaches $\mathfrak{r}_{max}$. Let us assume that d_1 and d_2 are the common differences for $\mathfrak{r}_1$ and $\mathfrak{r}_2$, respectively. So, the arithmetic progression series can be thought as follows:

$$\mathfrak{r}_1, \mathfrak{r}_1 + d_1, \cdots, \mathfrak{r}_1 + i d_1, \cdots, \mathfrak{r}_1 + (\mathfrak{t}_1 - 1) d_1$$

$$\mathfrak{r}_2, \mathfrak{r}_2 + d_2, \cdots, \mathfrak{r}_2 + j d_2, \cdots, \mathfrak{r}_2 + (\mathfrak{t}_2 - 1) d_2 \tag{3.19}$$

where initially $\mathfrak{r}_1 = \mathfrak{r}_{min}$ and $\mathfrak{r}_2 = \mathfrak{r}_{min}$ and at final step $\mathfrak{r}_1 + (\mathfrak{t}_1 - 1) d_1 = \mathfrak{r}_{max}$ and $\mathfrak{r}_2 + (\mathfrak{t}_2 - 1) d_2 = \mathfrak{r}_{max}$. The parameters $\mathfrak{t}_1$ and $\mathfrak{t}_2$ denote the number of possible values of regularization parameters $\mathfrak{r}_1$ and $\mathfrak{r}_2$, respectively. It is clearly seen that the diagonal elements of the covariance matrices are only changed by adding regularization parameters. Let us assume that $[\mathcal{C}_{11}+\mathfrak{r}_1 I]$ has dominant eigenvalue λ_{1_1} and the corresponding eigenvector ψ_{1_1}. So,

$$[\mathcal{C}_{11}+\mathfrak{r}_1 I] \psi_{1_1} = \lambda_{1_1} \psi_{1_1}. \tag{3.20}$$

Let us also assume that a scalar d_1 is added on the diagonal elements of the matrix $[\mathcal{C}_{11}+\mathfrak{r}_1 I]$. Multiplying this new matrix by the vector ψ_{1_1}, we get

$$[\mathcal{C}_{11}+(\mathfrak{r}_1 + d_1) I] \psi_{1_1} = [\mathcal{C}_{11}+\mathfrak{r}_1 I] \psi_{1_1} + d_1 I \psi_{1_1}$$

$$= \lambda_{1_1}\psi_{1_1} + d_1\psi_{1_1} = (\lambda_{1_1} + d_1)\psi_{1_1}. \tag{3.21}$$

Hence, if a regularization parameter is added on the diagonal elements of the covariance matrix, the eigenvalues are changed, but the eigenvectors remain same.

Let $\Lambda_{1_1}, \cdots, \Lambda_{1_{(i+1)}}, \cdots, \Lambda_{1_{t_1}}$ be the diagonal matrices, where diagonal elements are the eigenvalues of $[C_{11} + \mathfrak{r}_1 I], \cdots, [C_{11} + (\mathfrak{r}_1 + id_1)I], \cdots, [C_{11} + (\mathfrak{r}_1 + (\mathfrak{t}_1 - 1)d_1)I]$. Similarly, $\Lambda_{2_1}, \cdots, \Lambda_{2_{(j+1)}}, \cdots, \Lambda_{2_{t_2}}$ are the diagonal matrices with eigenvalues of $[C_{22}+\mathfrak{r}_2 I], \cdots, [C_{22}+ (\mathfrak{r}_2 + jd_2)I], \cdots, [C_{22} + (\mathfrak{r}_2 + (\mathfrak{t}_2 - 1)d_2)I]$ on the diagonal elements. The corresponding orthonormal eigenvectors are in the columns of Ψ_1 and Ψ_2. So, eigenvalue-eigenvector equations can be written as follows:

$$[C_{11} + (\mathfrak{r}_1 + (i-1)d_1)I]\Psi_1 = \Psi_1\Lambda_{1_i}; \tag{3.22}$$

$$\text{and} \quad [C_{22} + (\mathfrak{r}_2 + (j-1)d_2)I]\Psi_2 = \Psi_2\Lambda_{2_j}; \tag{3.23}$$

where $\forall i \in \{1, 2, \cdots, \mathfrak{t}_1\}$ and $\forall j \in \{1, 2, \cdots, \mathfrak{t}_2\}$. From (3.22), we get

$$[C_{11} + \mathfrak{r}_1 I]\Psi_1 + (i-1)d_1 I\Psi_1 = \Psi_1\Lambda_{1_i}$$

$$\Rightarrow \Psi_1\Lambda_{1_1} + (i-1)d_1\Psi_1 = \Psi_1\Lambda_{1_i}$$

$$\Rightarrow \Psi_1(\Lambda_{1_i} - \Lambda_{1_1} - (i-1)d_1 I) = 0$$

$$\Rightarrow \Lambda_{1_i} = \Lambda_1 + (i-1)d_1 I; \tag{3.24}$$

where $\Lambda_1 = \Lambda_{1_1}$. Similarly, from (3.23), we get

$$\Lambda_{2_j} = \Lambda_2 + (j-1)d_2 I; \tag{3.25}$$

where $\Lambda_2 = \Lambda_{2_1}$. Combining (3.22), (3.3.1) and (3.23), (3.25), we get

$$[C_{11} + (\mathfrak{r}_1 + id_1)I]\Psi_1 = \Psi_1(\Lambda_1 + id_1 I); \tag{3.26}$$

$$\text{and} \quad [C_{22} + (\mathfrak{r}_2 + jd_2)I]\Psi_2 = \Psi_2(\Lambda_2 + jd_2 I). \tag{3.27}$$

From (3.26) and (3.27), it is clearly seen that there is no need to calculate eigenvalue of the covariance matrices corresponding to every pair of regularization parameters $\mathfrak{r}_1$ and $\mathfrak{r}_2$. It is sufficient to calculate eigenvalues Λ_1 and Λ_2 of the covariance matrices corresponding to the initial values of $\mathfrak{r}_1$ and $\mathfrak{r}_2$, respectively. The eigenvalues of the covariance matrices corresponding to other values of $\mathfrak{r}_1$ and $\mathfrak{r}_2$ can be computed from the initial values using (3.3.1) and (3.25). On the other hand, relations (3.3.1), (3.26), and (3.27) establish the fact that eigenvectors of the covariance matrices remain unchanged irrespective of the values of regularization parameters. So, the eigenvalues and eigenvectors of the covariance matrices can be used to compute eigenvalues and eigenvectors of the covariance matrices corresponding to other values $\mathfrak{r}_1$ and $\mathfrak{r}_2$, using (3.3.1), (3.25), (3.26), and (3.27).

Based on the above analysis, it can be shown that if the regularization parameters $\mathfrak{r}_1$ and $\mathfrak{r}_2$ follow an arithmetic progression, the matrix $\mathcal{H}$ of (3.13) and the matrix $\tilde{\mathcal{H}}$ of (3.14) become

$$\mathcal{H}_{ij} = \left[\mathcal{C}_{11} + (\mathfrak{r}_1 + (i-1)d_1)I\right]^{-1} \mathcal{C}_{12}[\mathcal{C}_{22} + (\mathfrak{r}_2 + (j-1)d_2)I]^{-1}\mathcal{C}_{21}; \tag{3.28}$$

$$\text{and} \quad \tilde{\mathcal{H}}_{ij} = \left[\mathcal{C}_{22} + (\mathfrak{r}_2 + (j-1)d_2)I\right]^{-1} \mathcal{C}_{21}[\mathcal{C}_{11} + (\mathfrak{r}_1 + (i-1)d_1)I]^{-1}\mathcal{C}_{12}. \tag{3.29}$$

Combining (3.18), (3.26), (3.27), and (3.28), we get

$$\mathcal{H}_{ij} = \Psi_1[\Lambda_1 + (i-1)d_1 I]^{-1}\Psi_1^T\mathcal{C}_{12}\Psi_2[\Lambda_2 + (j-1)d_2 I]^{-1}\Psi_2^T\mathcal{C}_{21}. \tag{3.30}$$

Similarly, combining (3.18), (3.26), (3.27), and (3.29), we get

$$\tilde{\mathcal{H}}_{ij} = \Psi_2[\Lambda_2 + (j-1)d_2 I]^{-1}\Psi_2^T\mathcal{C}_{21}\Psi_1[\Lambda_1 + (i-1)d_1 I]^{-1}\Psi_1^T\mathcal{C}_{12}. \tag{3.31}$$

Suppose $\rho_1 \geqslant \cdots \geqslant \rho_t \geqslant \cdots \geqslant \rho_{\mathcal{D}}$ are the eigenvalues of $\mathcal{H}_{ij}$ and $w_{1_1}, \cdots, w_{1_t}, \cdots, w_{1_{\mathcal{D}}}$ are the orthonormalized eigenvectors corresponding to $\rho_1, \cdots, \rho_t, \cdots, \rho_{\mathcal{D}}$. Furthermore, let say, $\mathcal{D} \leqslant \min(m_1, m_2)$ and $\rho_1 \geqslant \cdots \geqslant \rho_t \geqslant \cdots \geqslant \rho_{\mathcal{D}}$ are the $\mathcal{D}$ largest eigenvalues of $\tilde{\mathcal{H}}_{ij}$ with orthonormalized eigenvectors $w_{2_1}, \cdots, w_{2_t}, \cdots, w_{2_{\mathcal{D}}}$. So,

$$\mathcal{H}_{ij} w_{1_t} = \rho_t w_{1_t}$$

$$\Rightarrow \Psi_1[\Lambda_1 + (i-1)d_1 I]^{-1}\Psi_1^T\mathcal{C}_{12}\Psi_2[\Lambda_2 + (j-1)d_2 I]^{-1}\Psi_2^T\mathcal{C}_{21}w_{1_t} = \rho_t w_{1_t}$$

$$\Rightarrow \Psi_2[\Lambda_2 + (j-1)d_2 I]^{-1}\Psi_2^T\mathcal{C}_{21}\Psi_1[\Lambda_1 + (i-1)d_1 I]^{-1}\Psi_1^T\mathcal{C}_{12}\Psi_2[\Lambda_2 + (j-1)d_2 I]^{-1}\Psi_2^T\mathcal{C}_{21}w_{1_t}$$

$$= \rho_t\Psi_2[\Lambda_2 + (j-1)d_2 I]^{-1}\Psi_2^T\mathcal{C}_{21}w_{1_t}$$

$$\Rightarrow \tilde{\mathcal{H}}_{ij}\Psi_2[\Lambda_2 + (j-1)d_2 I]^{-1}\Psi_2^T\mathcal{C}_{21}w_{1_t} = \rho_t\Psi_2[\Lambda_2 + (j-1)d_2 I]^{-1}\Psi_2^T\mathcal{C}_{21}w_{1_t}$$

$$\Rightarrow \tilde{\mathcal{H}}_{ij} w_{2_t} = \rho_t w_{2_t}; \tag{3.32}$$

The t-th eigenvector w_{2_t} of $\tilde{\mathcal{H}}_{ij}$ is proportional to $\Psi_2[\Lambda_2 + (j-1)d_2 I]^{-1}\Psi_2^T\mathcal{C}_{21}w_{1_t}$, that is, $w_{2_t} = \Psi_2[\Lambda_2 + (j-1)d_2 I]^{-1}\Psi_2^T\mathcal{C}_{21}w_{1_t}$. From (3.3.1), it can also be seen that either $\mathcal{H}_{ij}$ or $\tilde{\mathcal{H}}_{ij}$ is enough to calculate the eigenvector of $\mathcal{H}_{ij}$ and $\tilde{\mathcal{H}}_{ij}$.

Assuming $p = \min(m_1, m_2)$, p eigenvalue-eigenvector pairs of $\mathcal{H}_{ij}$ and $\tilde{\mathcal{H}}_{ij}$, which are the basis vectors, can be calculated using Jacobi method [90]. Then, p pairs of canonical variables are computed using (3.11). Finally, p features can be extracted using (3.12). The computational complexity of Jacobi method to compute p eigenvalue-eigenvector pairs is $\mathcal{O}(p^3)$.

3.3.2 CuRSaR: Proposed Algorithm

One of the main problems in real life high dimensional multimodal data analysis is how to extract relevant and significant features. In general, the extracted feature set may contain a huge number of irrelevant and insignificant features. The presence of such features may lead to a reduction in the useful information and degrade the prediction capability. Thus, the extracted feature subset should contain the features which have high relevance and high significance in the feature set. Such features are expected to be able to predict the classes of the samples. Accordingly, a measure is required that can assess the effectiveness of a feature set. In this work, hypercuboid equivalence partition matrix of rough hypercuboid approach [172] is used to select relevant and significant features, which are extracted from two multidimensional data sets by calculating their maximum correlation and variation.

Let $\mathcal{X}_1 \in \Re^{m_1 \times n}$ and $\mathcal{X}_2 \in \Re^{m_2 \times n}$ be two multidimensional data sets with m_1 and m_2 variables or attributes, respectively, and n samples. Let us assume that each variable is centered to have zero mean across the samples. Let $\mathfrak{t}_1$ and $\mathfrak{t}_2$ be the number of possible values of regularization parameters $\mathfrak{r}_1$ and $\mathfrak{r}_2$, respectively. The value of each regularization parameter is varied within a certain range $[\mathfrak{r}_{min}, \mathfrak{r}_{max}]$ as per (3.3.1), where $\mathfrak{r}_{min} \leqslant \mathfrak{r}_1, \mathfrak{r}_2 \leqslant \mathfrak{r}_{max}$. Let $\mathcal{F}_{t\,ij}$ be the t-th extracted feature with (i,j)-th regularization parameters of $\mathfrak{r}_1$ and $\mathfrak{r}_2$ and $\gamma_{\mathcal{F}_t}(\mathbb{D})$ be the relevance of the feature $\mathcal{F}_t$ with respect to the class labels $\mathbb{D}$. Define $\sigma_{\{\mathcal{F}_t,\mathcal{F}_\ell\}}(\mathbb{D}, \mathcal{F}_t)$ as the significance of the feature $\mathcal{F}_t$ with respect to another feature $\mathcal{F}_\ell \in \mathbb{S}$, where $\mathbb{S}$ is the set of $\mathcal{D}$ selected features and $\mathcal{D} \leqslant \min(m_1, m_2)$. The change in dependency when a feature is removed from the set of features, is a measure of the significance of the feature. To what extent a feature is contributing to the dependency on class labels can be determined by the significance of that feature. The significance of the feature $\mathcal{F}_t$ with respect to the feature set $\{\mathcal{F}_t, \mathcal{F}_\ell\}$ is given by

$$\sigma_{\{\mathcal{F}_t,\mathcal{F}_\ell\}}(\mathbb{D}, \mathcal{F}_t) = \gamma_{\{\mathcal{F}_t,\mathcal{F}_\ell\}}(\mathbb{D}) - \gamma_{\mathcal{F}_\ell}(\mathbb{D}). \tag{3.33}$$

Hence, the higher the change in dependency, the more significant the feature $\mathcal{F}_t$ is. If significance is 0, then the feature is dispensable. Therefore, the total relevance of all selected features for (i,j)-th regularization parameters of $\mathfrak{r}_1$ and $\mathfrak{r}_2$ is given by

$$R(i,j) = \sum_{\mathcal{F}_{t\,ij} \in \mathbb{S}} \gamma_{\mathcal{F}_{t\,ij}}(\mathbb{D}), \tag{3.34}$$

while the total significance among the selected features is as follows:

$$S(i,j) = \sum_{\mathcal{F}_{t\,ij} \neq \mathcal{F}_{\ell\,ij} \in \mathbb{S}} \sigma_{\{\mathcal{F}_{t\,ij},\mathcal{F}_{\ell\,ij}\}}(\mathbb{D}, \mathcal{F}_{t\,ij}) + \sigma_{\{\mathcal{F}_{t\,ij},\mathcal{F}_{\ell\,ij}\}}(\mathbb{D}, \mathcal{F}_{\ell\,ij}). \tag{3.35}$$

Therefore, the problem of extracting a set $\mathbb{S}$ of relevant and significant features from all possible combinations of regularization parameters $\mathfrak{r}_1$ and $\mathfrak{r}_2$ is equivalent to maximize both $R(i,j)$ and $S(i,j)$, that is, to maximizing the objective function $J(i,j)$, where

$$J(i,j) = \omega \times \frac{R(i,j)}{\mathcal{D}} + (1-\omega) \times \frac{S(i,j)}{\mathcal{D}(\mathcal{D}-1)} \tag{3.36}$$

where ω is a weight parameter. The criterion combining the above two constraints is called *maximum relevance-maximum significance* [172, 177]. The problem of generating a set of most significant and relevant feature set $\mathbb{S}$ from two multiblock data sets is addressed by Algorithm 3.1.

Algorithm 3.1 CuRSaR: Supervised CCA Using Max Relevance-Max Significance Criterion

```
Input: Two multidimensional variables $\mathcal{X}_1$ and $\mathcal{X}_2$.
Output: A set $\mathbb{S}$ of $\mathcal{D}$ selected features.
Calculate the cross-covariance matrix $\mathcal{C}_{12} \in \Re^{m_1 \times m_2}$ of $\mathcal{X}_1$ and $\mathcal{X}_2$ using (3.2).
Calculate the covariance matrix $\mathcal{C}_{11} \in \Re^{m_1 \times m_1}$ and $\mathcal{C}_{22} \in \Re^{m_2 \times m_2}$ of $\mathcal{X}_1$ and $\mathcal{X}_2$ using (3.3) and (3.4), respectively.
Calculate the eigenvalues $\Lambda_1 \in \Re^{m_1}$ and $\Lambda_2 \in \Re^{m_2}$ of $\mathcal{C}_{11}$ and $\mathcal{C}_{22}$, along with corresponding eigenvectors $\Psi_1$ and $\Psi_2$ using Jacobi method.
Initialize $\mathbb{S} \leftarrow \emptyset$ and $J_{\text{optimal}} = 0$.
for each $(i, j)$-th regularization parameters, of $\mathfrak{r}_1$ and $\mathfrak{r}_2$, where $\forall i \in \{1, 2, \cdots, \mathfrak{t}_1\}$ and $\forall j \in \{1, 2, \cdots, \mathfrak{t}_2\}$ do
    (I) If $m_1 \leqslant m_2$ (respectively, $m_1 > m_2$), calculate $\mathcal{H}_{ij}$ using (3.30) (respectively, $\tilde{\mathcal{H}}_{ij}$ using (3.31)).
    (II) Calculate the eigenvectors $w_{1_{ij}}$ (respectively, $w_{2_{ij}}$) of $\mathcal{H}_{ij}$(respectively, $\tilde{\mathcal{H}}_{ij}$) using Jacobi method and take first $\mathcal{D}$ eigenvectors.
    (III) Calculate $w_{2_{ij}} = \Psi_2[\Lambda_2 + (j-1)d_2 I]^{-1}\Psi_2^T \mathcal{C}_{21} w_{1_{ij}}$ (respectively, $w_{1_{ij}} = \Psi_1[\Lambda_1 + (i-1)d_1 I]^{-1}\Psi_1^T \mathcal{C}_{12} w_{2_{ij}}$).
    (IV) Calculate $\mathcal{D}$ pairs of canonical variables $\{\mathcal{U}_{1_{ij}}, \mathcal{U}_{2_{ij}}\}$ using (3.11).
    (V) Extract $\mathcal{D}$ features $\{\mathcal{F}_{ij}\}$ corresponding to $(i, j)$-th pair of regularization parameters using (3.12) and store them in $\mathbb{C}$.
    (VI) Compute the objective function $J(i, j)$ using (3.36).
    (VII) If $J(i, j) > J_{\text{optimal}}$, then $\mathbb{S} \leftarrow \mathbb{C}$, and $J_{\text{optimal}} = J(i, j)$.
end for
Stop.
```

3.3.3 Computation of Relevance and Significance

Generally, an m-dimensional hypercuboid or hyperrectangle is defined in the m-dimensional Euclidean space, where the space is defined by the m variables measured for each sample or object. In geometry, a hypercuboid or hyperrectangle is the generalization of a rectangle for higher dimensions, formally defined as the Cartesian product of orthogonal intervals. A d-dimensional hypercuboid with d attributes as its dimensions is defined as the Cartesian product of d orthogonal intervals. It encloses a region in the d-dimensional space, where each dimension corresponds to a certain attribute. The value domain of each dimension is the value range or interval that corresponds to a particular class. For all hypercuboids, any two objects belonging to a same class hypercuboid are said to be indiscernible with respect to that particular class.

Let $\mathbb{U} = \{\mathcal{O}_1, \cdots, \mathcal{O}_j, \cdots, \mathcal{O}_n\}$ be the set of n samples or objects with condition attribute or feature set $\mathbb{C} = \{\mathcal{F}_1, \cdots, \mathcal{F}_t, \cdots, \mathcal{F}_{\mathcal{D}}\}$, where $\mathcal{D} \leqslant \min(m_1, m_2)$ is the total number of extracted candidate features, for each regularization parameters $\mathfrak{r}_1$ and $\mathfrak{r}_2$, having non-zero significance values with respect to the already-selected features of $\mathbb{S}$. Let $\mathbb{D}$ be the class label or decision attribute set. If $\mathbb{U}/\mathbb{D} = \{\beta_1, \cdots, \beta_i, \cdots, \beta_c\}$ denotes c equivalence classes or granules of the universe $\mathbb{U}$ created by the equivalence relation induced from $\mathbb{D}$, then c information granules of $\mathbb{U}$ can also be created by the equivalence relation induced from each condition attribute $\mathcal{F}_t \in \mathbb{C}$. If $\mathbb{U}/\mathcal{F}_t = \{\delta_1, \cdots, \delta_i, \cdots, \delta_c\}$ denotes c equivalence classes or information granules of $\mathbb{U}$ induced by the condition attribute $\mathcal{F}_t$ and n is the number of objects in $\mathbb{U}$, then c-partitions of $\mathbb{U}$ are the sets of (cn) values $\{\hbar_{ij}(\mathcal{F}_t)\}$, which are arrayed as a matrix $\mathbb{H}(\mathcal{F}_t) = [\hbar_{ij}(\mathcal{F}_t)]_{c \times n}$. The matrix $\mathbb{H}(\mathcal{F}_t)$ is termed as hypercuboid equivalence partition matrix of the condition attribute $\mathcal{F}_t$ [172], where

$$\hbar_{ij}(\mathcal{F}_t) = \begin{cases} 1 & \text{if } \mathcal{L}_i \leqslant \mathcal{O}_j(\mathcal{F}_t) \leqslant \mathcal{U}_i \\ 0 & \text{otherwise} \end{cases} \tag{3.37}$$

represents the membership of object $\mathcal{O}_j$ in the i-th equivalence partition or class β_i satisfying following two conditions:

$$1 \leqslant \sum_{j=1}^{n} \hbar_{ij}(\mathcal{F}_t) \leqslant n, \forall i; \qquad 1 \leqslant \sum_{i=1}^{c} \hbar_{ij}(\mathcal{F}_t) \leqslant c, \forall j. \tag{3.38}$$

Here, $[\mathcal{L}_i, \mathcal{U}_i]$ represents the interval of i-th class β_i according to the class labels $\mathbb{D}$. The interval $[\mathcal{L}_i, \mathcal{U}_i]$ is spanned by the objects with class β_i with respect to the condition attribute $\mathcal{F}_t$. In other words, the value of each object $\mathcal{O}_j \in \beta_i$ with respect to $\mathcal{F}_t$ falls within $[\mathcal{L}_i, \mathcal{U}_i]$. A $c \times n$ hypercuboid equivalence partition matrix $\mathbb{H}(\mathcal{F}_t)$ represents the c-hypercuboid equivalence partitions of the universe generated by an equivalence relation. Each row of this matrix represents a hypercuboid equivalence class or partition. The i-th hypercuboid partition is represented as follows [172]:

$$\beta_i = \{\hbar_{i1}(\mathcal{F}_t)/\mathcal{O}_1 + \hbar_{i2}(\mathcal{F}_t)/\mathcal{O}_2 + \cdots + \hbar_{in}(\mathcal{F}_t)/\mathcal{O}_n\}. \tag{3.39}$$

However, every two intervals or hypercuboids may intersect with each other. These intersections form the implicit hypercuboids, which encompass objects those are misclassified. The degree of dependency of a condition attribute or a subset of attributes on decision attribute is estimated based on the cardinality of implicit hypercuboids. The misclassified objects belonging to implicit hypercuboids are identified using the confusion vector, which is defined based on hypercuboid equivalence partition matrix as follows [172]:

$$\mathbb{V}(\mathcal{F}_t) = [v_1(\mathcal{F}_t), v_2(\mathcal{F}_t), \cdots, v_n(\mathcal{F}_t)] \tag{3.40}$$

$$\text{where} \quad v_j(\mathcal{F}_t) = \min\{1, \sum_{i=1}^{c} h_{ij}(\mathcal{F}_t) - 1\}. \tag{3.41}$$

In other words, $v_j(\mathcal{F}_t) = 1$ if the j-th object O_j belongs to the implicit hypercuboid, which represents the boundary region of more than one classes. On the other hand, if the object O_j is encompassed by the lower approximation of any class, then $v_j(\mathcal{F}_t) = 0$ and the object O_j does not belong to the lower or upper approximations of any other classes. Hence, the confusion vector and hypercuboid equivalence partition matrix corresponding to feature $\mathcal{F}_t$ can be used for defining upper and lower approximations of the class β_i. Let $\beta_i \subseteq \mathbb{U}$. The information of the attribute $\mathcal{F}_t$ can be used to approximate β_i, by constructing $\mathcal{R}$-lower approximation and $\mathcal{R}$-upper approximation of β_i:

$$\underline{\mathcal{R}}(\beta_i) = \{O_j | \ h_{ij}(\mathcal{F}_t) = 1 \text{ and } v_j(\mathcal{F}_t) = 0\}; \tag{3.42}$$

$$\overline{\mathcal{R}}(\beta_i) = \{O_j | \ h_{ij}(\mathcal{F}_t) = 1\}; \tag{3.43}$$

where the attribute $\mathcal{F}_t$ induces equivalence relation $\mathcal{R}$. Hence, the cardinality of lower approximation of class β_i is computed as follows:

$$|\underline{\mathcal{R}}(\beta_i)| = \sum_{j=1}^{n} h_{ij}(\mathcal{F}_t)[1 - v_j(\mathcal{F}_t)]. \tag{3.44}$$

Based on the definition of lower approximation of rough sets, the positive region of decision attribute set $\mathbb{D}$ is defined as:

$$POS_{\mathcal{R}}(\mathbb{D}) = \bigcup_{\beta_i \in \mathbb{U}/\mathbb{D}} \underline{\mathcal{R}}(\beta_i). \tag{3.45}$$

The positive region, $POS_{\mathcal{R}}(\mathbb{D})$, contains all objects of $\mathbb{U}$ that can be classified to classes of $\mathbb{U}/\mathbb{D}$ using the knowledge in attribute $\mathcal{F}_t$. Combining (3.37), (3.40), and (3.45), the cardinality of positive regions of decision attribute $\mathbb{D}$, in terms of hypercuboid equivalence partition matrix and confusion vector of condition attribute $\mathcal{F}_t$, is given by

$$|POS_{\mathcal{R}}(\mathbb{D})| = \sum_{i=1}^{c} \sum_{j=1}^{n} h_{ij}(\mathcal{F}_t)[1 - v_j(\mathcal{F}_t)]. \tag{3.46}$$

Hence, the dependency between condition attribute $\mathcal{F}_t$ and decision attribute $\mathbb{D}$ is as follows:

$$\gamma_{\mathcal{F}_t}(\mathbb{D}) = \frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n} h_{ij}(\mathcal{F}_t)[1 - v_j(\mathcal{F}_t)] = 1 - \frac{1}{n} \sum_{j=1}^{n} v_j(\mathcal{F}_t) \tag{3.47}$$

where $\gamma_{\mathcal{F}_t}(\mathbb{D}) \in [0, 1]$. If $\mathbb{D}$ depends totally on $\mathcal{F}_t$, then $\gamma_{\mathcal{F}_t}(\mathbb{D}) = 1$; if $\mathbb{D}$ depends partially on $\mathcal{F}_t$, then $\gamma_{\mathcal{F}_t}(\mathbb{D}) \in (0, 1)$; and if $\mathbb{D}$ does not depend on $\mathcal{F}_t$, then $\gamma_{\mathcal{F}_t}(\mathbb{D}) = 0$.

The relevance of a feature $\mathcal{F}_t$ with respect to the class label or decision attribute $\mathbb{D}$ is computed using (3.47), while the joint relevance $\gamma_{\{\mathcal{F}_t,\mathcal{F}_\ell\}}(\mathbb{D})$ is to be computed to calculate the significance of the feature $\mathcal{F}_t$ with respect to the set $\{\mathcal{F}_t, \mathcal{F}_\ell\}$ using (3.33). The joint relevance depends on the $c \times n$ hypercuboid equivalence partition matrix corresponding to the set $\{\mathcal{F}_t, \mathcal{F}_\ell\}$, which is computed from two $c \times n$ equivalence partition matrices $\mathbb{H}(\mathcal{F}_t)$ and $\mathbb{H}(\mathcal{F}_\ell)$ as follows:

$$\mathbb{H}(\{\mathcal{F}_t, \mathcal{F}_\ell\}) = \mathbb{H}(\mathcal{F}_t) \cap \mathbb{H}(\mathcal{F}_\ell); \tag{3.48}$$

$$\text{where} \quad h_{ij}(\{\mathcal{F}_t, \mathcal{F}_\ell\}) = h_{ij}(\mathcal{F}_t) \times h_{ij}(\mathcal{F}_\ell). \tag{3.49}$$

3.3.4 Complexity Analysis

Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be the two data sets with n samples and c classes, and m_1 and m_2 represent the number of features in $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively. Let us assume that the regularization parameters $\mathfrak{r}_1$ and $\mathfrak{r}_2$ have $\mathfrak{t}_1$ and $\mathfrak{t}_2$ possible values. Let $q = \max(m_1, m_2)$ and $p = \min(m_1, m_2)$, where the number of extracted features $\mathcal{D} << p$. The computational complexity to calculate cross-covariance matrix $\mathcal{C}_{12}$ is $\mathcal{O}(pqn)$, whereas that of covariance matrices $\mathcal{C}_{11}$ and $\mathcal{C}_{22}$ is $\mathcal{O}(p^2n + q^2n)$. In step 3, the eigenvalues Λ_1 and Λ_2 with corresponding eigenvectors Ψ_1 and Ψ_2 can be calculated with complexity $\mathcal{O}(p^3 + q^3)$ using Jacobi method. Hence, the total time complexity of these three steps is $\mathcal{O}(pqn + p^2n + q^2n + p^3 + q^3) \approx \mathcal{O}(q^3)$. Step 4 has constant time complexity, which is $\mathcal{O}(1)$.

There is a loop in step 5, which is executed $(\mathfrak{t}_1 \times \mathfrak{t}_2)$ times. The computational complexity to calculate $\mathcal{H}_{ij}$ using (3.30) is $\mathcal{O}(p^3 + p^2q + pq^2 + q^3) \approx \mathcal{O}(q^3)$. The computational complexity to calculate first $\mathcal{D}$ eigenvectors in step 5.(II) is $\mathcal{O}(\mathcal{D}p^2)$. Step 5.(III) has complexity $\mathcal{O}(q^3 + pq^2 + \mathcal{D}pq)$. The canonical variables $\mathcal{U}_1$ and $\mathcal{U}_2$ have total $\mathcal{O}(\mathcal{D}pn + \mathcal{D}qn)$ time complexity. The computational complexity to extract first $\mathcal{D}$ features $\{\mathcal{F}\}$ is $\mathcal{O}(\mathcal{D}n)$. The time complexity to compute both relevance and significance of a feature is the same, which is $\mathcal{O}(cn)$. In effect, the total complexity to compute both relevance and significance of $\mathcal{D}$ features is $\mathcal{O}(\mathcal{D}cn)$. Hence, step 5.(VI) has computational complexity $\mathcal{O}(\mathcal{D})$. Finally, step 5.(VII) has $\mathcal{O}(\mathcal{D}n)$ time complexity. Hence, the total complexity to execute the loop $(\mathfrak{t}_1 \times \mathfrak{t}_2)$ times is $\mathcal{O}(\mathfrak{t}_1\mathfrak{t}_2(q^3 + \mathcal{D}p^2 + pq^2 + \mathcal{D}pq + \mathcal{D}pn + \mathcal{D}qn + \mathcal{D}n + \mathcal{D}cn + \mathcal{D} + \mathcal{D}n)) \approx \mathcal{O}(\mathfrak{t}_1\mathfrak{t}_2 q(q^2 + \mathcal{D}n))$.

Hence, the overall computational complexity of the proposed algorithm to extract relevant and significant features, which are linearly correlated, is $\mathcal{O}(q^3 + \mathfrak{t}_1\mathfrak{t}_2 q(q^2 + \mathcal{D}n)) \approx \mathcal{O}(\mathfrak{t}_1\mathfrak{t}_2 q(q^2 + \mathcal{D}n))$. On the other hand, the existing CCA, RCCA, and SRCCA algorithms have time complexity $\mathcal{O}(p!)$, $\mathcal{O}(\mathfrak{t}_1\mathfrak{t}_2 np!)$, and $\mathcal{O}(\mathfrak{t}_1\mathfrak{t}_2 p!)$, respectively, based on the analysis reported in [91].

3.4 Performance Analysis

In this section, the performance of the proposed feature extraction algorithm, termed as CuRSaR, is extensively studied and compared with that of some existing CCA based algorithms..

3.4.1 Data Sets and Experimental Setup

Five multimodal omics data sets, namely, glioblastoma multiforme (GBM), lung (LUNG), kidney (KIDNEY), lower grade glioma (LGG), and ovarian serous cystadenocarcinoma (OV), are used in the current research work. In Chapter 3 and Chapter 4, two modalities, namely, DNA methylation (mDNA) and RNA, are used, while in Chapter 5, Chapter 6, and Chapter 7, other modalities are used to validate the effectiveness of different multi-view data integration algorithms. The details of the gene (RNA) have been taken from RNA sequences in LUNG, KIDNEY, and LGG data sets, while gene expression provides gene-related information in the GBM and OV data sets. These data sets are downloaded from The Cancer Genome Atlas (TCGA) (`https://cancergenome.nih.gov/`). All five data sets with RNA and mDNA modalities are summarized in Table 3.1 and all modalities are encapsulated in Table 5.2 of Chapter 5. A detailed description of the data sets is reported in Appendix A.

Table 3.1: Description of Omics Data Sets Used

Different	Number of			
Data Sets	Classes	Samples	RNAs	mDNAs
GBM	5	213	12042	21422
LUNG	2	546	20502	294668
KIDNEY	2	305	20502	300451
LGG	3	374	11973	293965
OV	4	206	12042	20311

The performance of the proposed algorithm is compared with that of principal component analysis (PCA), CCA, RCCA, and several variants of SRCCA using t-test (SRCCA$_{\mathrm{TT}}$) [91], Wilcoxon rank sum test (SRCCA$_{\mathrm{WR}}$) [91], Wilks's lambda test (SRCCA$_{\mathrm{WL}}$) [91], mutual information (SRCCA$_{\mathrm{MI}}$), and rough hypercuboid (SRCCA$_{\mathrm{RH}}$). The performance of rough hypercuboid (RH) approach is also compared with that of mutual information (MI) in the proposed feature extraction framework. The value of ω in (3.36) is set to 0.5, while $\mathfrak{r}_1$ and $\mathfrak{r}_2$ are varied within $[0.0, 1.0]$ with 0.1 as common difference. The proposed algorithm is implemented in C language and run in Ubuntu 14.04 LTS having machine configuration Intel(R) Core(TM) i7-4790 CPU @ 3.60GHz×8 and 32 GB RAM. The source code of the CuRSaR algorithm is available at `https://www.isical.ac.in/~bibl/results/cursar/cursar.html`.

To evaluate the performance of different algorithms, support vector machine (SVM) [274] is used in the current study. Being a maximum margin classifier, the SVM defines the boundary between data samples of different classes by drawing an optimal hyperplane. The hyperplane leads to good generalization properties as it maximizes the margin between different classes. In the current work, linear kernels are used. Both 10-fold cross-validation (CV) and training-testing are performed to assess the performance of different algorithms. To analyze the statistical significance of the derived results in 10-fold CV, paired-t test (one-tailed), Wilcoxon signed rank test (one-tailed), and Friedman test (one-tailed), with a 95% confidence level, are used to compute the p-values. For training-testing, the randomly selected 50% samples from each class are used for training and the rest are used for testing purposes for each of the data sets. For each data set, 25 top-ranked correlated features are selected for the analysis, as in most of the cases the accuracy does not increase with the increase in number of features after 18-20 features. Thus, 25 features are taken to compare the accuracy and statistical significance analysis in the tables.

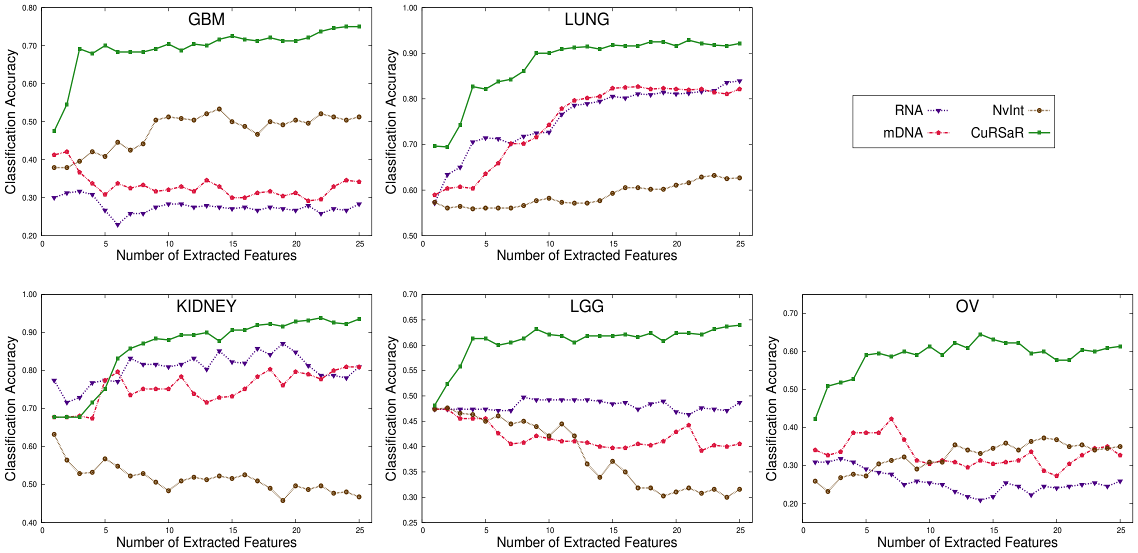


Figure 3.1: Variation of classification accuracy with respect to number of extracted features obtained using the PCA on individual modalities and concatenated data matrix (NvInt), and the proposed (CuRSaR) algorithm for 10-fold CV.

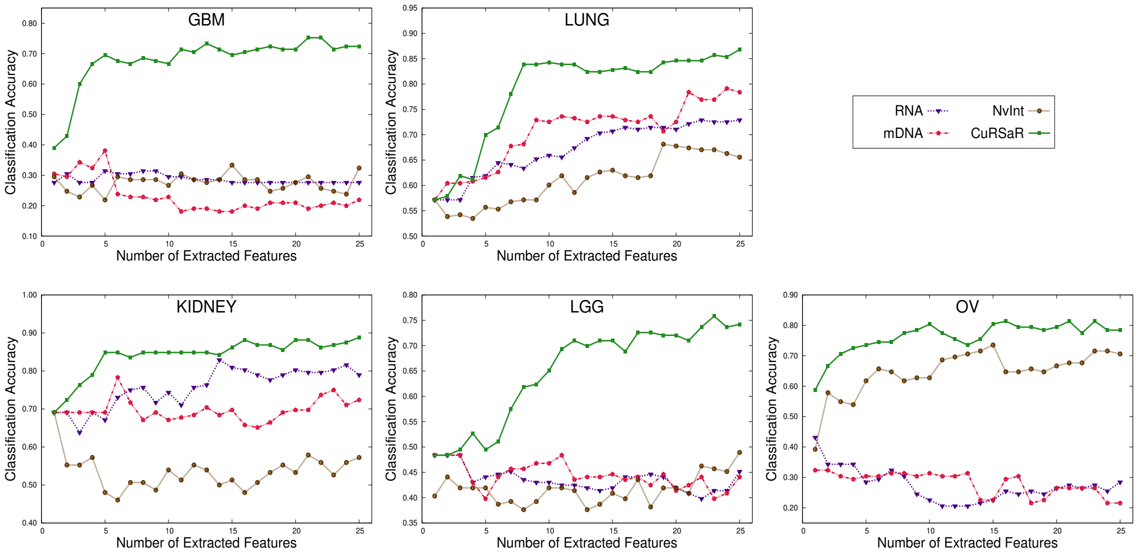


Figure 3.2: Variation of classification accuracy with respect to number of extracted features obtained using the PCA on individual modalities and concatenated data matrix (NvInt), and the proposed (CuRSaR) algorithm for training-testing.

3.4.2 Effectiveness of Proposed Algorithm

This section presents the performance of the proposed data integration algorithm, termed as CuRSaR, and its comparison with that of PCA on individual modalities and concatenated data matrix. Corresponding results are reported in Figure 3.1 and Figure 3.2 considering both 10-fold CV and training-testing, respectively. From the results reported in Figure 3.1 and Figure 3.2, it is observed that the classification accuracy of multiple modalities using naive integration (NvInt) is better than that of a single modality on GBM and OV data sets, while this is not the case for LUNG, KIDNEY, and LGG data sets. These results also infer that the integration of multiple modalities may provide better performance than a single modality if the integration is done efficiently.

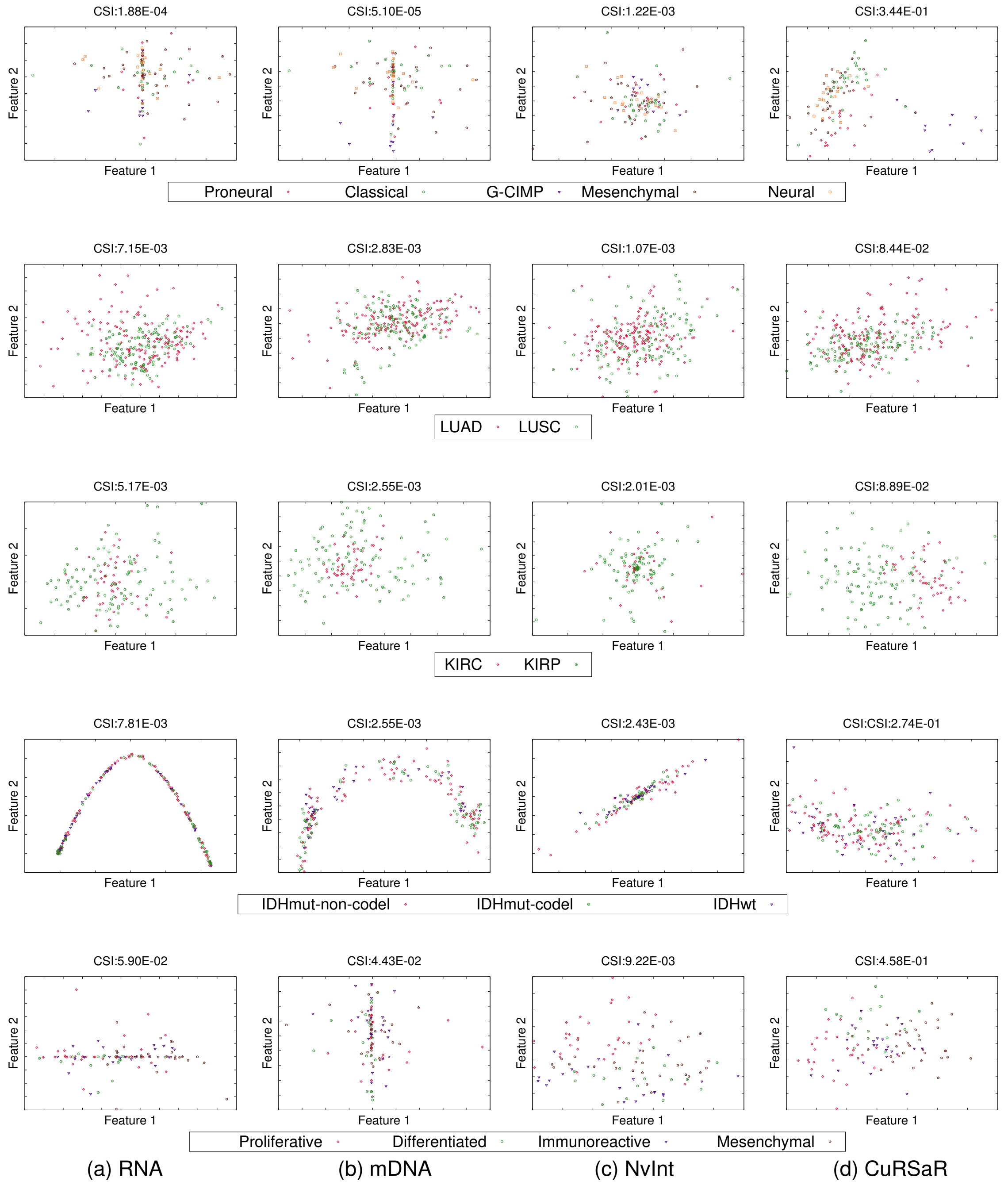
Figure 3.3 shows the scatter plots, along with the class separability index (CSI) on five data sets. The x-axis and y-axis of each plot represent the first and second extracted features, respectively. The class separability index (CSI) is defined as

$$\text{CSI} = \frac{\text{tr}(S_b)}{\text{tr}(S_w)}; \tag{3.50}$$

where $\text{tr}(A)$ represents the trace of matrix A. S_b and S_w indicate the between-class scatter matrix and within-class scatter matrix, respectively. A larger value of between-class scatter and smaller value of within-class scatter signifies better separation between the classes. Hence, the higher value of CSI indicates more separation. The value of the CSI is reported at the top of each figure. The qualitative results in Figure 3.3 show that PCA on individual modalities and concatenated data matrix cannot separate the classes properly. The CSI of PCA on single view and naive integrated matrix is lower compare to that of the proposed CuRSaR algorithm. All the results show that the naive integration, by direct concatenation of multiple modalities, is not sufficient for integrating the knowledge of all the modalities. Because of the radical imbalance and noisy nature of different modalities, naive integration does not perform well. Moreover, multiple modalities of a unique sample may provide complementary knowledge. The different modalities of the unique sample can make a connection between the characteristics of each sample. Hence, the combination of different modalities of a unique sample would have more discriminatory and absolute knowledge of the intrinsic properties of that sample than a single modality. Thus, to integrate the knowledge acquired in different modalities, an appropriate fusion method is required.

3.4.3 Importance of Rough Hypercuboid Approach

In the proposed CuRSaR algorithm, both the relevance and significance of an extracted feature are calculated based on the theory of hypercuboid equivalence partition matrix of rough hypercuboid approach. The relevance of a feature with respect to the class labels is calculated using (3.47), while the significance of a feature with respect to the already-extracted features is computed using (3.33). However, other measures such as mutual information can also be used to compute both relevance and significance of a feature [170, 177, 213]. In order to establish the importance of rough hypercuboid approach over mutual information, extensive experimental results are reported in Figure 3.4 and Figure 3.5 considering five data sets. Subsequent discussions analyze the results with respect to the classification accuracy of both the 10-fold CV and training-testing. All the results



(a) RNA (b) mDNA (c) NvInt (d) CuRSaR

Figure 3.3: Scatter plots for PCA on individual modalities and concatenated data matrix (NvInt), and proposed (CuRSaR) algorithm, along with class separability index (top to bottom: GBM, LUNG, KIDNEY, LGG, OV).

reported in Figure 3.4 and Figure 3.5 confirm that the performance of the hypercuboid equivalence partition matrix is better than that of mutual information in all the cases, irrespective of the data sets used.

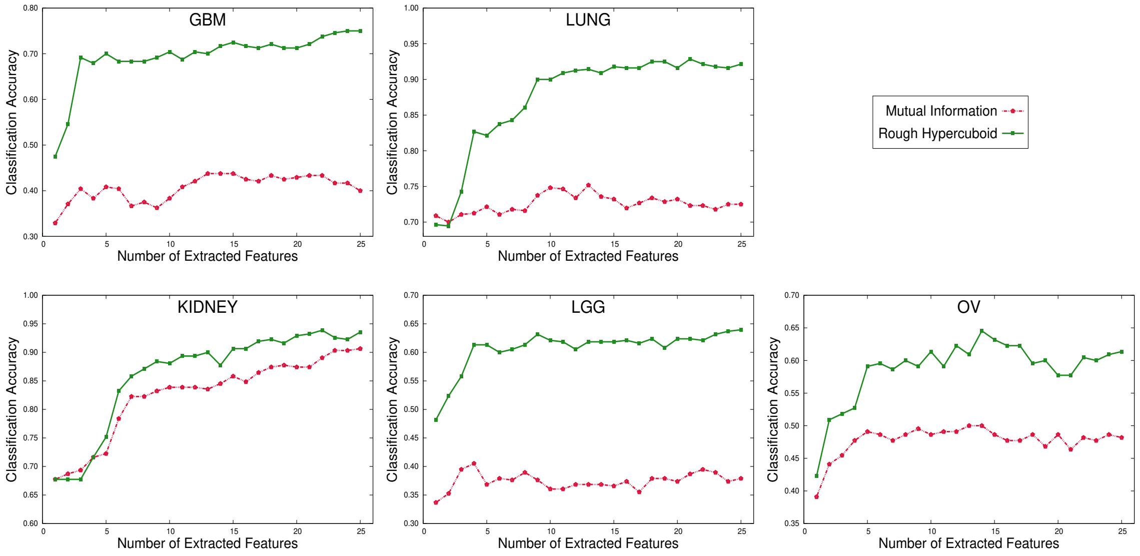


Figure 3.4: Variation of classification accuracy with respect to number of extracted features using mutual information and rough hypercuboid in the proposed algorithm for 10-fold CV.

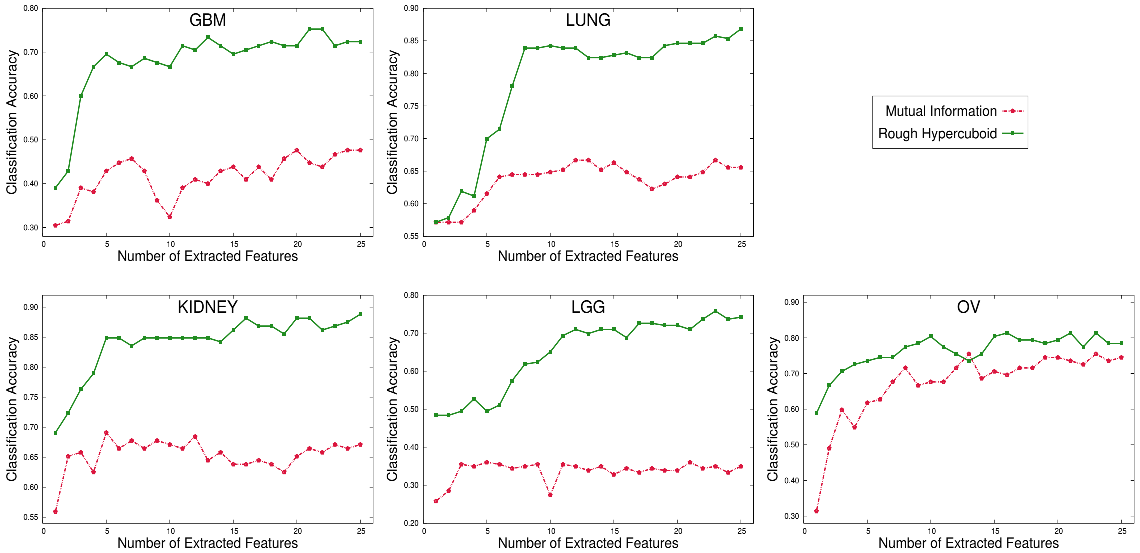
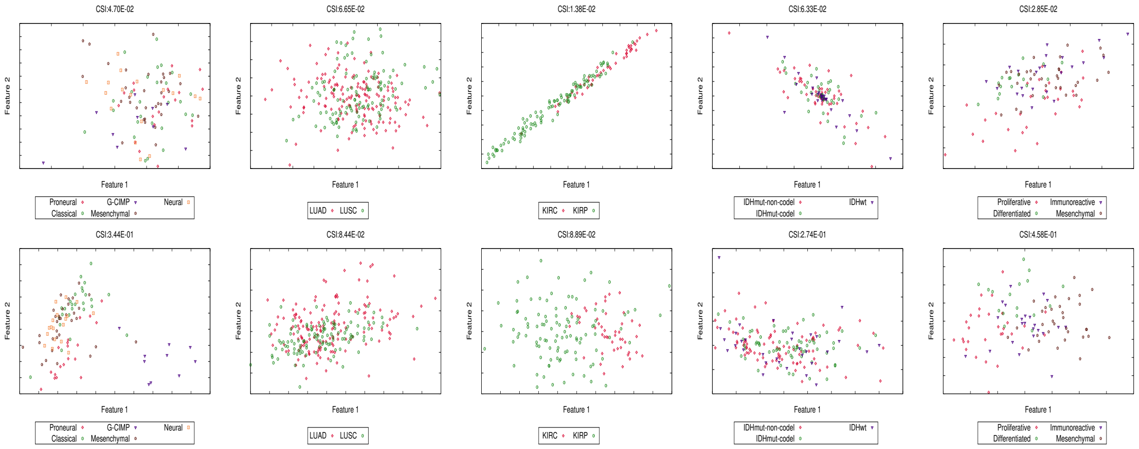


Figure 3.5: Variation of classification accuracy with respect to number of extracted features using mutual information and rough hypercuboid in the proposed algorithm for training-testing.

Figure 3.6 presents the scatter plots, along with the CSI for five data sets, where both rough hypercuboid equivalence partition matrix (RH) and mutual information (MI) are used to compute both significance and relevance of an extracted feature. From the results reported in Figure 3.6, it is noticeable that the CSI of the extracted features using mutual information is lower compared to that of hypercuboid equivalence partition matrix. Analyz-



(a) GBM (b) LUNG (c) KIDNEY (d) LGG (e) OV

Figure 3.6: Scatter plots for mutual information (top row) and rough hypercuboid (bottom row) in the proposed framework, along with class separability index.

ing the results of Figure 3.6, it is evident that the rough hypercuboid approach outperforms mutual information in the proposed framework. The significantly better performance of the rough hypercuboid based approach is obtained due to the fact that the quality of an extracted feature set, in rough hypercuboid approach, is evaluated by the hypercuboid equivalence partition matrix that makes use of supervised information of sample categories in the granulation process. Also, it provides an efficient way to calculate relevance and significance in approximation spaces. In effect, a reduced set of features having maximum relevance and significance is being obtained using the proposed CuRSaR algorithm.

Table 3.2: Classification Accuracy and Execution Time for Mutual Information and Rough Hypercuboid

Data Sets	Measure	Accuracy	Accuracy and Significance Analysis for 10-Fold CV						Time
		(Train-Test)	Mean	Median	StdDev	Paired-t:p	Wilcoxon:p	Friedman:p	(in sec.)
GBM	MI	0.476	0.400	0.354	0.174	**7.80E-05**	**2.50E-03**	**1.57E-03**	2996.0
	RH	**0.724**	**0.750**	**0.750**	0.079	-	-	-	2989.5
LUNG	MI	0.656	0.725	0.705	0.139	**7.91E-04**	**3.82E-03**	**2.70E-03**	2914.3
	RH	**0.868**	**0.921**	**0.920**	0.040	-	-	-	2781.2
KIDNEY	MI	0.671	0.906	**0.968**	0.172	*3.16E-01*	8.57E-01	*1.57E-01*	3519.1
	RH	**0.888**	**0.935**	0.935	0.034	-	-	-	2352.2
LGG	MI	0.349	0.379	0.408	0.153	**1.03E-03**	**4.65E-03**	**1.14E-02**	3027.6
	RH	**0.742**	**0.639**	**0.632**	0.061	-	-	-	3029.5
OV	MI	0.745	0.482	0.500	0.169	**1.75E-02**	**2.96E-02**	*2.06E-01*	3597.0
	RH	**0.784**	**0.614**	**0.682**	0.221	-	-	-	2331.5

Table 3.2 compares the classification accuracy, computed using rough hypercuboid approach and mutual information. The mean, median, and standard deviation of 10-fold CV are also reported in Table 3.2. To perform the statistical significance analysis, the p-values computed using different tests are reported in Table 3.2. Comparing the results reported in Table 3.2, it is evident that the proposed algorithm with rough hypercuboid attains higher

mean and median accuracy than with the mutual information, in almost all cases. The mutual information has achieved higher median accuracy (0.968) than rough hypercuboid on the KIDNEY data set only. Out of total 15 cases, the proposed CuRSaR algorithm, where hypercuboid equivalence partition matrix is used to compute both significance and relevance of an extracted feature, achieves significantly better (marked in bold) p-values than the mutual information based approach in 11 cases. On the other hand, the proposed algorithm provides better but not significant (marked in italics) p-values in only 4 cases, for all three significant tests on the KIDNEY data set and Friedman test on the OV data set.

3.4.4 Comparative Performance Analysis

Finally this section presents the comparative performance analysis of the proposed CuRSaR algorithm and various state-of-the-art data integration algorithms, namely, CCA, RCCA, SRCCA$_{\mathrm{TT}}$, SRCCA$_{\mathrm{WR}}$, SRCCA$_{\mathrm{WL}}$, SRCCA$_{\mathrm{MI}}$, and SRCCA$_{\mathrm{RH}}$, on five data sets, namely, GBM, LUNG, KIDNEY, LGG, and OV. Corresponding results are reported in Figure 3.7 and Figure 3.8, along with Table 3.3 and Table 3.4. From the results reported in Figure 3.7 and Figure 3.8, it is seen that the classification accuracy of the proposed CuRSaR algorithm is significantly higher than that of the existing data integration algorithms. Figure 3.9 shows the scatter plots of several existing algorithms on five data sets. The value of the CSI is also reported at the top of each figure.

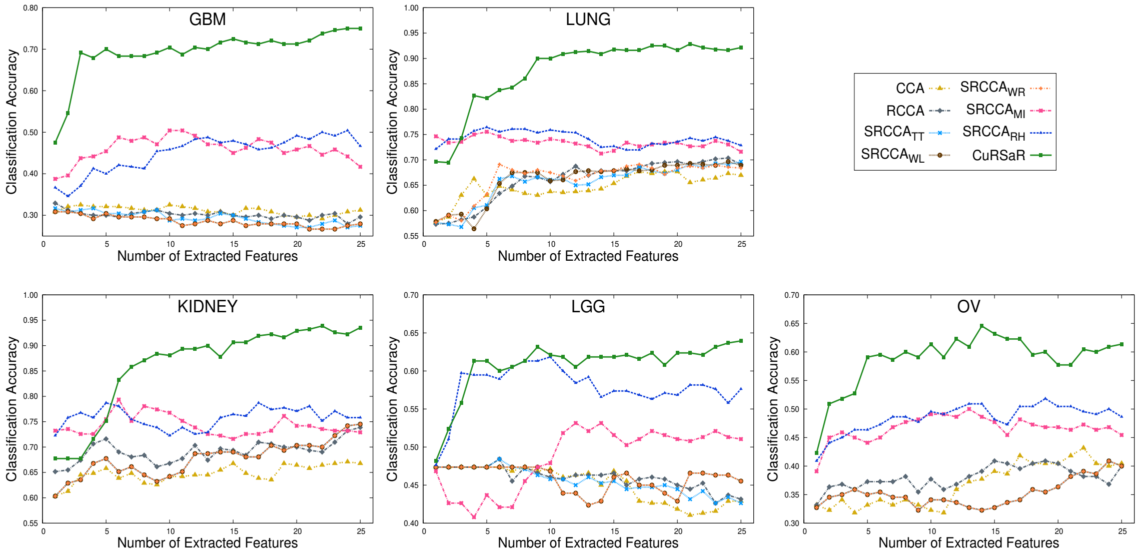


Figure 3.7: Variation of classification accuracy with respect to number of extracted features for several existing algorithms and the proposed (CuRSaR) algorithm using 10-fold CV.

From the results reported in Figure 3.9, it is observable that the CSI of the extracted features using the proposed CuRSaR algorithm are higher than that of the several existing algorithms. Comparing the results of Figure 3.9, it is also noticeable that the proposed algorithm is able to separate different classes of GBM and KIDNEY data sets using the first two extracted features only, which is also evident from the corresponding class separability

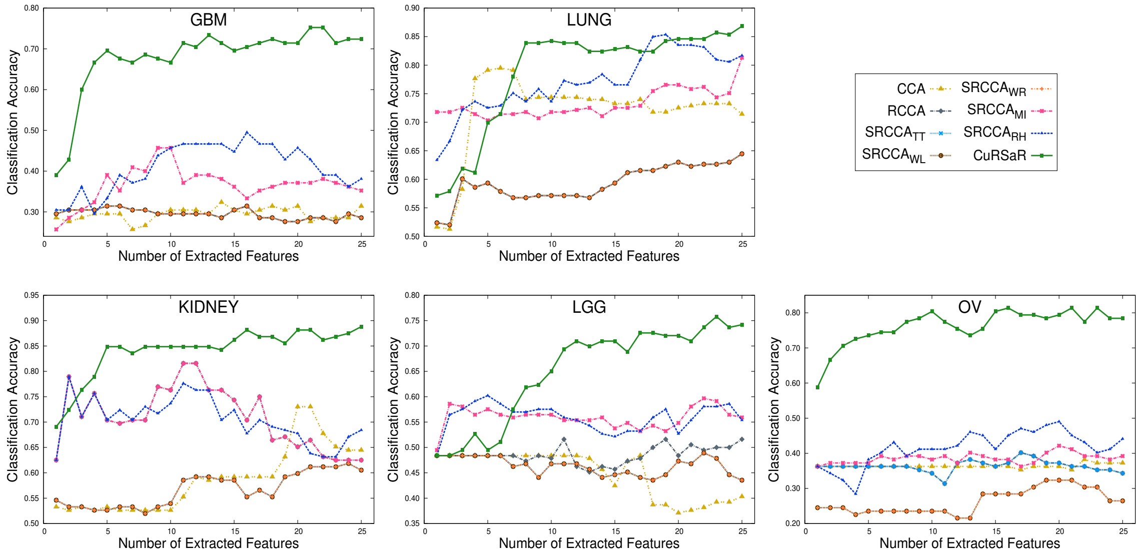


Figure 3.8: Variation of classification accuracy with respect to number of extracted features for several existing algorithms and the proposed (CuRSaR) algorithm using training-testing.

index values; though there is some overlap between the classes on LUNG, LGG, and OV data sets. On the other hand, the classes are hardly separable using all existing algorithms on each data set.

All the results reported in Table 3.3 confirm that the proposed CuRSaR algorithm attains the highest mean and median accuracy, in almost all the cases. The SRCCA_{RH} has achieved higher median accuracy of 0.737 than the proposed CuRSaR on the LGG data set only. To perform the statistical significance analysis, the p-values computed using different tests are reported in Table 3.4. The proposed algorithm attains significantly better p-values (marked in bold) than several existing data integration algorithms in 93 cases, out of a total of 105 cases, considering 95% confidence level. On the other hand, the proposed CuRSaR algorithm provides better but not significant (marked in italics) p-values in only 12 cases. The significantly better performance of the proposed CuRSaR algorithm is achieved due to the fact that the CuRSaR algorithm extracts features by maximizing the relevance and significance of the features. Both relevance and significance measures depend on the information of sample categories. On the other hand, CCA and RCCA extract features from two different modalities without considering the supervised information of class labels. In effect, the proposed algorithm is able to extract more relevant and significant features from a pair of modalities.

The existing SRCCA algorithms consider only the correlation of the first pair of canonical variables [91]. In effect, other canonical variable pairs may have insignificant relation with the first pair of canonical variables or may introduce some irrelevant features in the whole extracted feature set, which may degrade the prediction capability of the classifiers used. Also, the existing SRCCA algorithms fail to address the problem of uncertainty associated with data analysis. On the other hand, the proposed algorithm considers both relevance and significance measures of all extracted features while optimizing the regular-

Table 3.3: Classification Accuracy and Execution Time of Different Algorithms

Different	Data	Accuracy	Accuracy for 10-Fold CV			Time
Algorithms	Sets	(Train-Test)	Mean	Median	StdDev	(in sec.)
CCA	GBM	0.314	0.313	0.292	0.086	2130.2
RCCA		0.286	0.296	0.271	0.093	3061.4
SRCCA$_{\mathrm{TT}}$		0.286	0.275	0.250	0.110	3086.0
SRCCA$_{\mathrm{WL}}$		0.286	0.279	0.250	0.100	3041.8
SRCCA$_{\mathrm{WR}}$		0.286	0.279	0.250	0.100	3018.1
SRCCA$_{\mathrm{MI}}$		0.352	0.417	0.313	0.283	3044.6
SRCCA$_{\mathrm{RH}}$		0.381	0.467	0.396	0.273	3039.7
CuRSaR		**0.724**	**0.750**	**0.750**	0.079	2989.5
CCA	LUNG	0.714	0.670	0.661	0.118	2054.4
RCCA		0.645	0.691	0.714	0.097	2924.5
SRCCA$_{\mathrm{TT}}$		0.645	0.696	0.696	0.117	2914.0
SRCCA$_{\mathrm{WL}}$		0.645	0.689	0.643	0.112	2915.8
SRCCA$_{\mathrm{WR}}$		0.645	0.684	0.679	0.117	2880.7
SRCCA$_{\mathrm{MI}}$		0.813	0.716	0.723	0.108	2925.0
SRCCA$_{\mathrm{RH}}$		0.817	0.729	0.732	0.116	2895.9
CuRSaR		**0.868**	**0.921**	**0.920**	0.040	2781.2
CCA	KIDNEY	0.645	0.668	0.710	0.139	2159.7
RCCA		0.625	0.739	0.790	0.155	3618.6
SRCCA$_{\mathrm{TT}}$		0.605	0.745	0.790	0.129	3494.0
SRCCA$_{\mathrm{WL}}$		0.605	0.745	0.790	0.129	2984.6
SRCCA$_{\mathrm{WR}}$		0.605	0.745	0.790	0.129	3609.9
SRCCA$_{\mathrm{MI}}$		0.625	0.729	0.774	0.140	3599.1
SRCCA$_{\mathrm{RH}}$		0.684	0.758	0.790	0.155	3630.3
CuRSaR		**0.888**	**0.935**	**0.935**	0.034	2352.2
CCA	LGG	0.403	0.429	0.421	0.079	2116.7
RCCA		0.516	0.432	0.421	0.060	3216.1
SRCCA$_{\mathrm{TT}}$		0.435	0.426	0.408	0.063	3149.6
SRCCA$_{\mathrm{WL}}$		0.435	0.455	0.461	0.045	3184.6
SRCCA$_{\mathrm{WR}}$		0.435	0.455	0.461	0.045	3218.3
SRCCA$_{\mathrm{MI}}$		0.559	0.511	0.461	0.187	3102.0
SRCCA$_{\mathrm{RH}}$		0.554	0.576	**0.737**	0.283	3096.3
CuRSaR		**0.742**	**0.639**	0.632	0.061	3029.5
CCA	OV	0.373	0.405	0.386	0.124	2164.3
RCCA		0.343	0.400	0.409	0.098	3019.2
SRCCA$_{\mathrm{TT}}$		0.343	0.400	0.409	0.111	3679.4
SRCCA$_{\mathrm{WL}}$		0.265	0.400	0.409	0.111	3643.0
SRCCA$_{\mathrm{WR}}$		0.265	0.400	0.409	0.111	3457.7
SRCCA$_{\mathrm{MI}}$		0.392	0.455	0.432	0.132	3529.8
SRCCA$_{\mathrm{RH}}$		0.441	0.486	0.477	0.174	3627.7
CuRSaR		**0.784**	**0.614**	**0.682**	0.221	2331.5

ization parameters. The rough hypercuboid approach, employed in the proposed algorithm, can also efficiently handle the uncertainty due to imprecision in computation and vagueness in the class definition. In effect, the proposed algorithm provides significantly better results as compared to existing algorithms in most of the cases. Moreover, the analytical formulation introduced in this chapter makes the computational complexity of the proposed

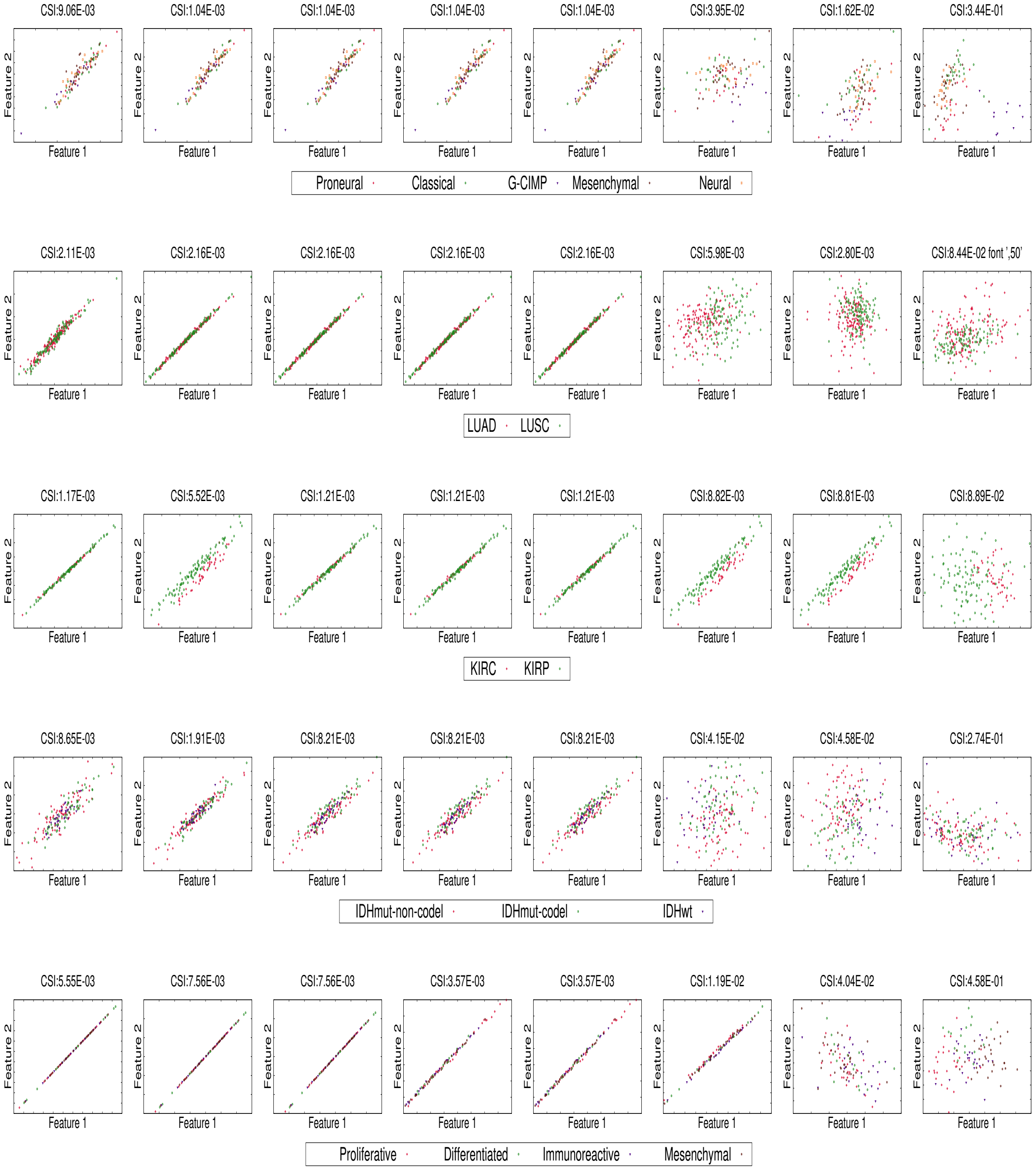
Table 3.4: Statistical Significance Analysis of Different Algorithms

Different Algorithms	Data Sets	p-values for 10-Fold CV		
		Paired-t	Wilcoxon	Friedman
CCA	GBM	**3.31E-07**	**2.50E-03**	**1.57E-03**
RCCA		**5.33E-08**	**2.46E-03**	**1.57E-03**
SRCCA$_{\mathrm{TT}}$		**2.97E-07**	**2.50E-03**	**1.57E-03**
SRCCA$_{\mathrm{WL}}$		**3.61E-08**	**2.45E-03**	**1.57E-03**
SRCCA$_{\mathrm{WR}}$		**3.61E-08**	**2.45E-03**	**1.57E-03**
SRCCA$_{\mathrm{MI}}$		**1.17E-03**	**5.86E-03**	**4.68E-03**
SRCCA$_{\mathrm{RH}}$		**3.36E-03**	**7.58E-03**	**1.96E-02**
CCA	LUNG	**4.36E-05**	**2.52E-03**	**1.57E-03**
RCCA		**5.25E-05**	**2.52E-03**	**1.57E-03**
SRCCA$_{\mathrm{TT}}$		**1.43E-04**	**2.50E-03**	**1.57E-03**
SRCCA$_{\mathrm{WL}}$		**2.70E-05**	**2.53E-03**	**1.57E-03**
SRCCA$_{\mathrm{WR}}$		**7.94E-05**	**2.52E-03**	**1.57E-03**
SRCCA$_{\mathrm{MI}}$		**2.40E-04**	**2.52E-03**	**1.57E-03**
SRCCA$_{\mathrm{RH}}$		**2.91E-04**	**2.52E-03**	**1.57E-03**
CCA	KIDNEY	**2.15E-04**	**2.53E-03**	**1.57E-03**
RCCA		**3.21E-03**	**3.98E-03**	**1.14E-02**
SRCCA$_{\mathrm{TT}}$		**1.22E-03**	**2.52E-03**	**1.57E-03**
SRCCA$_{\mathrm{WL}}$		**1.22E-03**	**2.52E-03**	**1.57E-03**
SRCCA$_{\mathrm{WR}}$		**1.22E-03**	**2.52E-03**	**1.57E-03**
SRCCA$_{\mathrm{MI}}$		**1.21E-03**	**3.79E-03**	**2.70E-03**
SRCCA$_{\mathrm{RH}}$		**4.54E-03**	**4.58E-03**	**1.14E-02**
CCA	LGG	**8.38E-05**	**3.76E-03**	**2.70E-03**
RCCA		**7.31E-05**	**2.50E-03**	**1.57E-03**
SRCCA$_{\mathrm{TT}}$		**6.72E-05**	**2.50E-03**	**1.57E-03**
SRCCA$_{\mathrm{WL}}$		**1.33E-04**	**2.52E-03**	**1.57E-03**
SRCCA$_{\mathrm{WR}}$		**1.33E-04**	**2.52E-03**	**1.57E-03**
SRCCA$_{\mathrm{MI}}$		**2.31E-02**	**2.32E-02**	*2.06E-01*
SRCCA$_{\mathrm{RH}}$		*2.61E-01*	*3.80E-01*	5.27E-01
CCA	OV	**9.35E-03**	**2.06E-02**	*2.06E-01*
RCCA		**4.85E-03**	**1.04E-02**	*9.56E-02*
SRCCA$_{\mathrm{TT}}$		**2.04E-03**	**6.23E-03**	*5.78E-02*
SRCCA$_{\mathrm{WL}}$		**2.04E-03**	**6.23E-03**	*5.78E-02*
SRCCA$_{\mathrm{WR}}$		**2.04E-03**	**6.23E-03**	*5.78E-02*
SRCCA$_{\mathrm{MI}}$		**1.21E-02**	**2.52E-02**	*3.17E-01*
SRCCA$_{\mathrm{RH}}$		**3.52E-02**	*6.93E-02*	*3.17E-01*

CuRSaR algorithm significantly lower than the existing RCCA and SRCCA.

3.5 Conclusion

This chapter presents a new feature extraction algorithm, termed as CuRSaR, for two multidimensional data sets. The merits of CCA and rough sets have been integrated judiciously to develop the proposed algorithm. To establish the relation between the covariance matrices of different regularization parameters, a theoretical formulation has been presented. It helps the proposed CuRSaR algorithm to extract correlated features, which are relevant



(a) CCA (b) RCCA (c) SRCCA$_{TT}$ (d) SRCCA$_{WL}$ (e) SRCCA$_{WR}$ (f) SRCCA$_{MI}$ (g) SRCCA$_{RH}$ (h) CuRSaR

Figure 3.9: Scatter plots for the proposed (CuRSaR) algorithm and several existing algorithms, along with class separability index (top to bottom: GBM, LUNG, KIDNEY, LGG, OV).

with respect to the class label and significant among them. The hypercuboid equivalence partition matrix has been used to compute both relevance and significance of a feature. The optimum regularization parameters of CCA have been determined using the equivalence partition matrix. The effectiveness of the proposed algorithm, along with a comparison with other algorithms, has been demonstrated considering two different modalities, namely, RNA and mDNA. The concept of hypercuboid equivalence partition matrix is found to be successful in extracting relevant and significant features from multimodal high dimensional real-life data sets.

One of the main problems associated with high dimensional multimodal real-life data sets is how to extract relevant and significant features sequentially. Instead of producing all canonical variables simultaneously, if each variable is computed sequentially, the quality of each generated feature can be evaluated independently, and eventually, a reduced set of features can be selected based on their quality. In this regard, a new feature extraction algorithm is presented in the next chapter, which extracts new features sequentially from two multidimensional data sets by maximizing their relevance with respect to the class label and significance with respect to already-extracted features. An analytical formulation is introduced, which enables the proposed algorithm to extract required number of correlated features sequentially with lesser computational cost as compared to existing algorithms.

Chapter 4

Fast and Robust Supervised CCA

4.1 Introduction

Due to the drastic variation and noisy nature of the acquired signals, unimodal based pattern analysis and recognition systems usually afford low level of performance, which leads to insufficient and inaccurate pattern representation of the perception of interest. On the other hand, multimodal data contain more information. By using multiple types of data of a unique sample, it is possible to make the linkages between attributes within each type of data. The combination of multimodal data may potentially provide a more complete and discriminatory description of the intrinsic characteristics of the pattern by producing improved system performance than single modality only.

As mentioned in Chapter 3, canonical correlation analysis (CCA) [112] is a bivariate feature extraction method, which provides an efficient way of measuring the linear relationship between two multidimensional variables. The goal of CCA is to find the best linear transformation for two multidimensional data sets so that the maximum correlation between them can be achieved. Regularized CCA (RCCA) [93,278] is an improved version of CCA. It prevents over-fitting of insufficient training data by using a ridge regression optimization scheme [27]. It works by adding small positive quantities to the diagonals of two covariance matrices $\mathcal{C}_{11}$ and $\mathcal{C}_{22}$ of two data sets $\mathcal{X}_1$ and $\mathcal{X}_2$ having m_1 and m_2 features, respectively, to guarantee their invertibility [107]. In [56], an alternative method to the existing RCCA has been presented, which is based on the estimates of the correlation matrices that minimize the mean squared error risk function. An et al. [13] proposed a robust CCA, which uses shrinkage estimation and smoothing technique to estimate the data covariance matrices with limited samples. However, RCCA is computationally very expensive because of this regularization process. Also, both CCA and RCCA are unsupervised and fail to take complete advantage of available class label information [91].

Supervised RCCA (SRCCA) incorporates a supervised feature selection scheme to perform the regularization [91,155]. It includes the information of available class label to select maximally correlated features. In SRCCA, regularization is done by embedding component with the most discriminatory score as chosen by feature selection scheme and then adjusted for the remaining dimensions [91,155]. However, existing SRCCA considers only correlation of first pair of canonical variables. It may happen that other canonical variable pairs have insignificant relation with first pair of canonical variables, or there may be some irrelevant features in the whole extracted feature set, which should not be considered for further processing [173]. In this regard, a new supervised RCCA, termed as CuRSaR (CCA using maximum Relevance-maximum Significance criterion and Rough sets), has been proposed recently in [174] and presented in Chapter 3, where whole extracted feature set is used to optimize the regularization parameters. However, both existing SRCCA and CuRSaR of Chapter 3 extract all possible features ($\min(m_1, m_2)$), which may not be needed at all. If features are extracted sequentially, then only the required number of relevant, significant, and nonredundant features can be extracted. In effect, it will be computationally less expensive. Moreover, uncertainty in omics data analysis is one of the major concerns. Some of the sources of this uncertainty include imprecision in computation and vagueness in class definition. Rough set theory has gained popularity in modeling and propagating uncertainty. It deals with vagueness and incompleteness, and is proposed for indiscernibility in classification, according to some similarity.

In this regard, this chapter presents a fast and robust feature extraction algorithm, termed as FaRoC (Fast and Robust CCA), for two multidimensional data sets. It integrates judiciously the merits of SRCCA and the theory of rough sets. While SRCCA addresses the problem of integrating heterogeneous sources of data, the rough hypercuboid approach of rough sets deals with vagueness in sample categories. The proposed algorithm extracts a new feature by maximizing the relevance with respect to sample categories or class labels and significance with respect to already-extracted features. Both the significance and relevance measures are computed based on the concept of hypercuboid equivalence partition matrix. In the proposed algorithm, the relevance and/or significance do not depend only on the first pair of canonical variables, rather the whole extracted feature set is considered to calculate these measures. A theoretical analysis is presented to establish the relation between CCA and RCCA, which drastically reduces the computational complexity of existing RCCA and helps to extract correlated features sequentially. As the features are extracted sequentially, only the required number of significant and relevant features can be generated without generating all possible features. In effect, the proposed algorithm has lower computational cost as compared to existing approaches. The efficacy of the proposed FaRoC algorithm, as well as comparative performance analysis with existing algorithms, is shown on real-life data sets. Some of the results of this chapter are reported in [181, 183].

The rest of the chapter is organized as follows: Section 4.2 presents the proposed algorithm. A theoretical analysis is presented in this section to establish the relation between CCA and RCCA, which drastically reduces the computational complexity of existing RCCA and helps to extract correlated features sequentially. The effectiveness of the proposed data integration algorithm, along with a comparative performance analysis with state-of-the-art algorithms on different data sets, is presented in Section 4.3. Concluding remarks are provided in Section 4.4.

4.2 Proposed Method

This section presents a fast and robust feature extraction algorithm, termed as FaRoC, integrating judiciously the information of two multidimensional data sets. Some important analytical formulations are reported next prior to describing the proposed algorithm.

4.2.1 Relation Between CCA and RCCA

Let $\mathcal{X}_1 \in \Re^{m_1 \times n}$ and $\mathcal{X}_2 \in \Re^{m_2 \times n}$ be two multivariate data sets having m_1 and m_2 number of features, respectively, and n is the number of samples in both $\mathcal{X}_1$ and $\mathcal{X}_2$. Let us assume that each multivariate data is centered to have zero mean across the samples. As explained in Section 3.2 of Chapter 3, the objective of CCA is to extract latent features from $\mathcal{X}_1$ and $\mathcal{X}_2$, which are most highly correlated. CCA obtains two directional weight vectors, also termed as basis vectors, $w_1 \in \Re^{m_1}$ and $w_2 \in \Re^{m_2}$ such that the empirical correlation between the respective projections onto these weight vectors, that is, between $\mathcal{X}_1^T w_1$ and $\mathcal{X}_2^T w_2$ is maximum. The correlation coefficient ρ is given in (3.1) of Chapter 3. The basis vectors w_1 and w_2 are the eigenvectors of matrices $\mathcal{H}$ and $\tilde{\mathcal{H}}$, respectively, with eigenvalue ρ. The matrix $\mathcal{H}$ and $\tilde{\mathcal{H}}$ are defined in (3.10) of Chapter 3.

To deal with the singularity issue of the covariance matrices $\mathcal{C}_{11}$ and $\mathcal{C}_{22}$ of $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively, regularization parameters $\mathfrak{r}_1$ and $\mathfrak{r}_2$ are added to the diagonal of $\mathcal{C}_{11}$ and $\mathcal{C}_{22}$, respectively. Hence, $\mathcal{H}$ and $\tilde{\mathcal{H}}$ become (3.13) and (3.14) of Chapter 3, respectively. In general, the regularized parameters $\mathfrak{r}_1$ and $\mathfrak{r}_2$ of both RCCA and SRCCA are varied within a specified range $[\mathfrak{r}_{min}, \mathfrak{r}_{max}]$. Let us assume that these $\mathfrak{r}_1$ and $\mathfrak{r}_2$ follow an arithmetic progression, with common differences d_1 and d_2, respectively, as mentioned in (3.3.1) of Chapter 3. Let the parameters $\mathfrak{t}_1$ and $\mathfrak{t}_2$ be the number of possible values of $\mathfrak{r}_1$ and $\mathfrak{r}_2$, respectively. As explained in Section 3.3.1 of Chapter 3, the spectral decomposition [269] can be used to calculate $[\mathcal{C}_{11} + \mathfrak{r}_1 I]^{-1}$ and $[\mathcal{C}_{22} + \mathfrak{r}_2 I]^{-1}$ for the computation of $\mathcal{H}$ and $\tilde{\mathcal{H}}$. Hence, $\mathcal{H}_{ij}$ and $\tilde{\mathcal{H}}_{ij}$ become (3.30) and (3.31) of Chapter 3, respectively, $\forall i \in \{1, 2, \cdots, \mathfrak{t}_1\}$ and $\forall j \in \{1, 2, \cdots, \mathfrak{t}_2\}$. Here, each element of the diagonal matrices Λ_1 and Λ_2 are the eigenvalues of the matrices $[\mathcal{C}_{11} + \mathfrak{r}_1 I]$ and $[\mathcal{C}_{22} + \mathfrak{r}_2 I]$, respectively, where each column of the matrices Ψ_1 and Ψ_2 represent the corresponding orthonormalized eigenvectors.

Now, $[\Lambda_1 + (i-1)d_1 I]$ is a non-singular diagonal matrix, which is obtained by adding two diagonal matrices, namely, Λ_1 and $[(i-1)d_1 I]$. The diagonal elements of Λ_1 represent the eigenvalues of matrix $\mathcal{C}_{11}$, while that of $[(i-1)d_1 I]$ are $(i-1)d_1$. As Λ_1 and $[\Lambda_1 + (i-1)d_1 I]$ are non-singular matrices, and $[(i-1)d_1 I]$ has rank m_1 for $i > 1$, the inverse of matrix $[\Lambda_1 + (i-1)d_1 I]$ can be calculated using the inverse of matrix Λ_1 [192]. As matrix $[(i-1)d_1 I]$ has rank m_1, the matrix $[\Lambda_1 + (i-1)d_1 I]$ can be written as

$$\mathcal{G}_{m_1+1} = \Lambda_1 + (i-1)d_1 I = \Lambda_1 + [(i-1)d_1 I]_1 + [(i-1)d_1 I]_2 + \cdots + [(i-1)d_1 I]_{m_1}, \quad (4.1)$$

where each $[(i-1)d_1 I]_r, \forall r = 1, 2, \cdots, m_1$, has rank 1. So, the inverse of $\mathcal{G}_{m_1+1}$ can be expressed as follows:

$$\mathcal{G}_{m_1+1}^{-1} = \mathcal{G}_{m_1}^{-1} + g_{m_1} \mathcal{G}_{m_1}^{-1}[(i-1)d_1 I]_{m_1} \mathcal{G}_{m_1}^{-1} = \Lambda_1^{-1} + \sum_{r=1}^{m_1} g_r \mathcal{G}_r^{-1}[(i-1)d_1 I]_r \mathcal{G}_r^{-1} \quad (4.2)$$

considering $\mathcal{G}_1 = \Lambda_1$, where

$$g_r = \frac{1}{1 + \operatorname{trace}\left(\mathcal{G}_r^{-1}[(i-1)d_1 I]_r\right)}. \quad (4.3)$$

Similarly, the matrix $[\Lambda_2 + (j-1)d_2 I]$ can be written as

$$\tilde{\mathcal{G}}_{m_2+1} = \Lambda_2 + (j-1)d_2 I = \Lambda_2 + [(j-1)d_2 I]_1 + [(j-1)d_2 I]_2 + \cdots + [(j-1)d_2 I]_{m_2}, \tag{4.4}$$

where the matrix $[(j-1)d_2 I]$ has rank m_2 for $j > 1$ and each $[(j-1)d_2 I]_s, \forall s = 1, 2, \cdots, m_2$, has rank 1. So, the inverse of $\tilde{\mathcal{G}}_{m_2+1}$ can be expressed, considering $\tilde{\mathcal{G}}_1 = \Lambda_2$, as follows:

$$\tilde{\mathcal{G}}_{m_2+1}^{-1} = \Lambda_2^{-1} + \sum_{s=1}^{m_2} \tilde{g}_s \tilde{\mathcal{G}}_s^{-1}[(j-1)d_2 I]_s \tilde{\mathcal{G}}_s^{-1}, \tag{4.5}$$

$$\text{where} \;\; \tilde{g}_s = \frac{1}{1 + \operatorname{trace}\left(\tilde{\mathcal{G}}_s^{-1}[(j-1)d_2 I]_s\right)}. \tag{4.6}$$

Hence, using (4.2) and (4.5), the matrix $\mathcal{H}_{ij}$ of (3.30) in Chapter 3 becomes

$$\mathcal{H}_{ij} = \Psi_1(\Lambda_1^{-1} + \sum_{r=1}^{m_1} g_r \mathcal{G}_r^{-1}[(i-1)d_1 I]_r \mathcal{G}_r^{-1})\Psi_1^T \mathcal{C}_{12} \Psi_2(\Lambda_2^{-1} + \sum_{s=1}^{m_2} \tilde{g}_s \tilde{\mathcal{G}}_s^{-1}[(j-1)d_2 I]_s \tilde{\mathcal{G}}_s^{-1})\Psi_2^T \mathcal{C}_{21}$$

$$\Rightarrow \mathcal{H}_{ij} = \Psi_1 \Lambda_1^{-1} \Psi_1^T \mathcal{C}_{12} \Psi_2 \Lambda_2^{-1} \Psi_2^T \mathcal{C}_{21} + \mathcal{B}_{ij} = \mathcal{H}_{11} + \mathcal{B}_{ij}, \tag{4.7}$$

$$\text{where} \;\; \mathcal{B}_{ij} = \Theta_i \Psi_2 \Lambda_2^{-1} \Psi_2^T \mathcal{C}_{21} + \Psi_1 \Lambda_1^{-1} \Psi_1^T \mathcal{C}_{12} \Phi_j + \Theta_i \Phi_j = \Theta_i \mathcal{C}_{22}^{-1} \mathcal{C}_{21} + \mathcal{C}_{11}^{-1} \mathcal{C}_{12} \Phi_j + \Theta_i \Phi_j; \tag{4.8}$$

$$\Theta_i = \Psi_1 \sum_{r=1}^{m_1} g_r \mathcal{G}_r^{-1}[(i-1)d_1 I]_r \mathcal{G}_r^{-1} \Psi_1^T \mathcal{C}_{12}; \tag{4.9}$$

$$\text{and} \;\; \Phi_j = \Psi_2 \sum_{s=1}^{m_2} \tilde{g}_s \tilde{\mathcal{G}}_s^{-1}[(j-1)d_2 I]_s \tilde{\mathcal{G}}_s^{-1} \Psi_2^T \mathcal{C}_{21}. \tag{4.10}$$

Similarly, using (4.2) and (4.5), the matrix $\tilde{\mathcal{H}}_{ij}$ of (3.31) in Chapter 3 becomes

$$\tilde{\mathcal{H}}_{ij} = \Psi_2(\Lambda_2^{-1} + \sum_{s=1}^{m_2} \tilde{g}_s \tilde{\mathcal{G}}_s^{-1}[(j-1)d_2 I]_s \tilde{\mathcal{G}}_s^{-1})\Psi_2^T \mathcal{C}_{21} \Psi_1(\Lambda_1^{-1} + \sum_{r=1}^{m_1} g_r \mathcal{G}_r^{-1}[(i-1)d_1 I]_r \mathcal{G}_r^{-1})\Psi_1^T \mathcal{C}_{12}$$

$$\Rightarrow \tilde{\mathcal{H}}_{ij} = \Psi_2 \Lambda_2^{-1} \Psi_2^T \mathcal{C}_{21} \Psi_1 \Lambda_1^{-1} \Psi_1^T \mathcal{C}_{12} + \tilde{\mathcal{B}}_{ij} = \tilde{\mathcal{H}}_{11} + \tilde{\mathcal{B}}_{ij}, \tag{4.11}$$

$$\text{where} \;\; \tilde{\mathcal{B}}_{ij} = \Phi_j \Psi_1 \Lambda_1^{-1} \Psi_1^T \mathcal{C}_{12} + \Psi_2 \Lambda_2^{-1} \Psi_2^T \mathcal{C}_{21} \Theta_i + \Phi_j \Theta_i = \Phi_j \mathcal{C}_{11}^{-1} \mathcal{C}_{12} + \mathcal{C}_{22}^{-1} \mathcal{C}_{21} \Theta_i + \Phi_j \Theta_i. \tag{4.12}$$

From (4.2.1) and (4.2.1), it is clear that if eigenvalues and eigenvectors of $\mathcal{C}_{11}$ and $\mathcal{C}_{22}$ are calculated to compute $\mathcal{H}_{11}$ and $\tilde{\mathcal{H}}_{11}$ matrices for initial values of $\mathfrak{r}_1$ and $\mathfrak{r}_2$, there is no need to compute eigenvalues and eigenvectors for computing $\mathcal{H}_{ij}$ and $\tilde{\mathcal{H}}_{ij}$ at other values of $\mathfrak{r}_1$ and $\mathfrak{r}_2$, as initial eigenvalues and eigenvectors can be used to compute different $\mathcal{H}_{ij}$ and $\tilde{\mathcal{H}}_{ij}$ matrices. Also, if the minimum value of $\mathfrak{r}_1$ and $\mathfrak{r}_2$ is set to 0, then $\mathcal{H}$ and $\tilde{\mathcal{H}}$ of CCA can be used to compute different $\mathcal{H}_{ij}$ and $\tilde{\mathcal{H}}_{ij}$ matrices of RCCA corresponding to different values of regularization parameters.

4.2.2 Sequential Generation of Canonical Variables

From (3.2.1) and (3.2.1) of Chapter 3, it is evident that the non-zero eigenvalues of $\Sigma\Sigma^T$, $\Sigma^T\Sigma$, $\mathcal{H}$, and $\tilde{\mathcal{H}}$ are same [89]. So, either $\mathcal{H}$ or $\tilde{\mathcal{H}}$ is computed using (4.2.1) or (4.2.1), respectively, corresponding to a pair of regularization parameters $\mathfrak{r}_1$ and $\mathfrak{r}_2$ depending on whether $m_1 \leqslant m_2$ or $m_1 > m_2$. Let us assume that $\mathcal{H}$ has t-th eigenvalue ρ_t and corresponding eigenvector is w_{1_t}. So,

$$\mathcal{H}w_{1_t} = \rho_t w_{1_t}$$

$$\Rightarrow \mathcal{C}_{11}^{-1}\mathcal{C}_{12}\mathcal{C}_{22}^{-1}\mathcal{C}_{21}w_{1_t} = \rho_t w_{1_t}$$

$$\Rightarrow \mathcal{C}_{22}^{-1}\mathcal{C}_{21}\mathcal{C}_{11}^{-1}\mathcal{C}_{12}\mathcal{C}_{22}^{-1}\mathcal{C}_{21}w_{1_t} = \rho_t \mathcal{C}_{22}^{-1}\mathcal{C}_{21}w_{1_t}$$

$$\Rightarrow \tilde{\mathcal{H}}w_{2_t} = \rho_t w_{2_t}; \quad \text{where} \quad w_{2_t} = \mathcal{C}_{22}^{-1}\mathcal{C}_{21}w_{1_t}. \tag{4.13}$$

So, the t-th eigenvector w_{2_t} of $\tilde{\mathcal{H}}$ is proportional to $\mathcal{C}_{22}^{-1}\mathcal{C}_{21}$ and can be obtained from the t-th eigenvector w_{1_t} of $\mathcal{H}$ using (4.2.2). So, from (4.2.2), it is also clear that either $\mathcal{H}$ or $\tilde{\mathcal{H}}$ is enough to calculate the eigenvectors of $\mathcal{H}$ and $\tilde{\mathcal{H}}$. Assuming p=min(m_1, m_2), p eigenvalue-eigenvector pairs can be calculated using Jacobi method [90]. Then, p pairs of basis vectors and p pairs of canonical variables are computed using (3.2.1) or (3.2.1) and (3.11) of Chapter 3, respectively. Finally, p features can be extracted using (3.12) of Chapter 3. The computational complexity of Jacobi method to compute p eigenvalue-eigenvector pairs is $\mathcal{O}(p^3)$.

However, the value of p is large for real life high dimensional multimodal data analysis. So, a small fraction, among the huge amount of extracted features, is effective to perform a certain task. Furthermore, a small subset of extracted features is advisable to develop tools for delivering interpretable, reliable, and precise results. Hence, the goal of multimodal data analysis is to identify a reduced set of most relevant extracted features. This is referred to as feature selection, and an important problem in machine learning. So, instead of generating all p eigenvalue-eigenvector pairs using Jacobi method, if each eigenvalue-eigenvector pair of $\mathcal{H}$ is generated sequentially, the quality of each extracted feature can be evaluated, and finally, $\mathcal{D}$ features can be extracted for multimodal data analysis, where $\mathcal{D} << p$. In the proposed algorithm, each eigenvalue-eigenvector pair of $\mathcal{H}$ is calculated sequentially by using the Power method [90]. The t-th eigenvalue-eigenvector pair can be calculated with the help of the first eigenvalue-eigenvector pair as explained below. Following analysis establishes that there is a direct relation between t-th and $(t+1)$-th eigenvalue-eigenvector pairs, and using this relation, all correlated features can be extracted sequentially. Let us assume that ρ_t and w_{1_t} be the t-th eigenvalue and corresponding eigenvector, respectively, of $\mathcal{H}$ matrix. So,

$$\mathcal{H}w_{1_t} = \rho_t w_{1_t}$$

$$\Rightarrow \mathcal{H}w_{1_t}w_{1_t}^T = \rho_t w_{1_t}w_{1_t}^T$$

$$\Rightarrow \mathcal{H} - \mathcal{H}w_{1_t}w_{1_t}^T = \mathcal{H} - \rho_t w_{1_t} w_{\chi t}^T$$

$$\Rightarrow \left(\mathcal{H} - \mathcal{H}w_{1_t}w_{1_t}^T\right) w_{1_{(t+1)}} = \left(\mathcal{H} - \rho_t w_{1_t} w_{1_t}^T\right) w_{1_{(t+1)}}$$

$$\Rightarrow \mathcal{H}w_{1_{(t+1)}} - \mathcal{H}w_{1_t}w_{1_t}^T w_{1_{(t+1)}} = \left(\mathcal{H} - \rho_t w_{1_t} w_{1_t}^T\right) w_{1_{(t+1)}}$$

$$\Rightarrow \mathcal{H}w_{1_{(t+1)}} = \left(\mathcal{H} - \rho_t w_{1_t} w_{1_t}^T\right) w_{1_{(t+1)}}; \tag{4.14}$$

where $w_{1_{(t+1)}}$ is the $(t+1)$-th eigenvector of $\mathcal{H}$ corresponding to the eigenvalue $\rho_{(t+1)}$, that is,

$$\mathcal{H}w_{1_{(t+1)}} = \rho_{(t+1)} w_{1_{(t+1)}}. \tag{4.15}$$

Hence, from (4.2.2) and (4.15), we get

$$\left(\mathcal{H} - \rho_t w_{1_t} w_{1_t}^T\right) w_{1_{(t+1)}} = \rho_{(t+1)} w_{1_{(t+1)}}. \tag{4.16}$$

Hence, from (4.16), it is proved that the $(t+1)$-th eigenvalue-eigenvector pair $\left\{\rho_{(t+1)}, w_{1_{(t+1)}}\right\}$ of the matrix $\mathcal{H}$ is same as first eigenvalue-eigenvector pair of the matrix $\left(\mathcal{H} - \rho_t w_{1_t} w_{1_t}^T\right)$. For calculating $(t+1)$-th eigenvalue-eigenvector pair, the matrices $\mathcal{H}$ and $\tilde{\mathcal{H}}$ can be calculated, based on Deflation method [293], as follows:

$$\mathcal{H}(t+1) = \mathcal{H}(t) - \rho_t w_{1_t} w_{1_t}^T = \mathcal{H}(1) - \sum_{\ell=1}^{t} \rho_\ell w_{1_\ell} w_{1_\ell}^T; \tag{4.17}$$

$$\tilde{\mathcal{H}}(t+1) = \tilde{\mathcal{H}}(t) - \rho_t w_{2_t} w_{2_t}^T = \tilde{\mathcal{H}}(1) - \sum_{\ell=1}^{t} \rho_\ell w_{2_\ell} w_{2_\ell}^T. \tag{4.18}$$

Therefore, $\rho_{(t+1)}$ and $w_{1_{(t+1)}}$ can be calculated with the help of previously calculated eigenvalue-eigenvector pairs, that is, ρ_ℓ and w_{1_ℓ}, $\forall \ell = 1, 2, \cdots, t$. Hence, using (4.17), each eigenvalue-eigenvector pair of matrix $\mathcal{H}$ can be calculated sequentially. So, for RCCA with (i,j)-th regularization parameters of $\mathfrak{r}_1$ and $\mathfrak{r}_2$, to compute $(t+1)$-th basis eigenvector, the matrices $\mathcal{H}_{ij}$ and $\tilde{\mathcal{H}}_{ij}$ can be calculated by using (4.2.1), (4.17) and (4.2.1), (4.18) as follows:

$$\mathcal{H}_{ij}(t+1) = \mathcal{H}_{ij}(1) - \sum_{\ell=1}^{t} \rho_{\ell_{ij}} w_{1_{\ell_{ij}}} w_{1_{\ell_{ij}}}^T = \mathcal{H}_{11} + \mathcal{B}_{ij} - \sum_{\ell=1}^{t} \rho_{\ell_{ij}} w_{1_{\ell_{ij}}} w_{1_{\ell_{ij}}}^T; \tag{4.19}$$

$$\text{and} \quad \tilde{\mathcal{H}}_{ij}(t+1) = \tilde{\mathcal{H}}_{ij}(1) - \sum_{\ell=1}^{t} \rho_{\ell_{ij}} w_{2_{\ell_{ij}}} w_{2_{\ell_{ij}}}^T = \tilde{\mathcal{H}}_{11} + \tilde{\mathcal{B}}_{ij} - \sum_{\ell=1}^{t} \rho_{\ell_{ij}} w_{2_{\ell_{ij}}} w_{2_{\ell_{ij}}}^T; \tag{4.20}$$

where $\forall t \in \{1, 2, \cdots, p\}$, $p = \min(m_1, m_2)$, $\forall i \in \{1, 2, \cdots, \mathfrak{t}_1\}$, and $\forall j \in \{1, 2, \cdots, \mathfrak{t}_2\}$.

4.2.3 Relevance and Significance for Regularization

In the current work, significant and relevant features are extracted from two multidimensional data sets using the concept of hypercuboid equivalence partition matrix [172] of rough hypercuboid approach, described in Chapter 3. The regularization parameters are optimized through computing these two measures. Let $\mathcal{X}_1 \in \Re^{m_1 \times n}$ and $\mathcal{X}_2 \in \Re^{m_2 \times n}$ be two multidimensional data sets with m_1 and m_2 variables or attributes, respectively, and n samples. Let us assume that each attribute is centered to have zero mean across the samples. Let $\mathfrak{t}_1$ and $\mathfrak{t}_2$ be the number of possible values of regularization parameters $\mathfrak{r}_1$ and $\mathfrak{r}_2$, respectively. The value of each regularization parameter is varied within a certain range $[\mathfrak{r}_{min}, \mathfrak{r}_{max}]$, where $\mathfrak{r}_{min} \leqslant \mathfrak{r}_1, \mathfrak{r}_2 \leqslant \mathfrak{r}_{max}$. Let $\mathcal{F}_{t_{ij}}$ be the t-th extracted feature with (i,j)-th regularization parameters of $\mathfrak{r}_1$ and $\mathfrak{r}_2$ and $\gamma_{\mathcal{F}_t}(\mathbb{D})$ is the relevance of the feature $\mathcal{F}_t$ with respect to the class labels $\mathbb{D}$, which is given in (3.47) of Chapter 3. Define $\sigma_{\{\mathcal{F}_t, \mathcal{F}_\ell\}}(\mathbb{D}, \mathcal{F}_t)$ as the significance of the feature $\mathcal{F}_t$ with respect to another feature $\mathcal{F}_\ell \in \mathbb{S}$, where $\mathbb{S}$ is the set of $\mathcal{D}$ selected features and $\mathcal{D} \leqslant \min(m_1, m_2)$. The change in joint relevance or dependency when a feature is discarded from the set of features, is a measure of the significance of the feature. To what extent a feature contributes for computing the dependency on class labels can be computed by the significance of the feature. The significance $\sigma_{\{\mathcal{F}_t, \mathcal{F}_\ell\}}(\mathbb{D}, \mathcal{F}_t)$ of the feature $\mathcal{F}_t$ with respect to the feature set $\{\mathcal{F}_t, \mathcal{F}_\ell\}$ is given in (3.33) of Chapter 3.

Hence, the problem of extracting a relevant and significant feature set $\mathbb{S}$ from all possible combinations of regularization parameters $\mathfrak{r}_1$ and $\mathfrak{r}_2$ is equivalent to maximizing the average relevance of all extracted features as well as maximizing the average significance among them. The problem of generating the set $\mathbb{S}$ from two multiblock data sets is addressed by Algorithm 4.1.

4.2.4 Complexity Analysis

Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be the two datasets with n samples and c classes, where m_1 and m_2 represent the number of features in $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively. Let us assume that the regularization parameters $\mathfrak{r}_1$ and $\mathfrak{r}_2$ have $\mathfrak{t}_1$ and $\mathfrak{t}_2$ possible values. Let $q = \max(m_1, m_2)$ and $p = \min(m_1, m_2)$, where the number of extracted features $\mathcal{D} << p$. The computational complexity to calculate cross-covariance matrix $\mathcal{C}_{12}$ is $\mathcal{O}(pqn)$, whereas the total time complexity to compute covariance matrices $\mathcal{C}_{11}$ and $\mathcal{C}_{22}$ is $\mathcal{O}(p^2n + q^2n)$. In step 3, the eigenvalues Λ_1 and Λ_2, along with corresponding eigenvectors Ψ_1 and Ψ_2, can be calculated with complexity $\mathcal{O}(p^3 + q^3)$ using Jacobi method. Hence, the total time complexity of these three steps is $\mathcal{O}(pqn + p^2n + q^2n + p^3 + q^3) \approx \mathcal{O}(q^3)$. The total time complexity to compute $\mathcal{C}_{11}^{-1}$ and $\mathcal{C}_{22}^{-1}$ is $\mathcal{O}(p^3 + q^3)$. So, step 4, for computing the matrix $\mathcal{H}_{11}$, has computational complexity $\mathcal{O}(p^3 + p^2q + pq^2 + q^3) \approx \mathcal{O}(q^3)$. Step 5 has constant time complexity, which is $\mathcal{O}(1)$.

There is a loop in step 6, which is executed $\mathcal{D}$ times. The first step of this loop has constant complexity of $\mathcal{O}(1)$ and the next step has another loop, which is executed $(\mathfrak{t}_1 \times \mathfrak{t}_2)$ times. The computational complexity to calculate $\mathcal{B}_{ij}$ or $\tilde{\mathcal{B}}_{ij}$ is $\mathcal{O}(p^2 + q^2 + p^2q + pq^2) \approx \mathcal{O}(pq^2)$. Hence, the total complexity of step 6(b)(i) is $\mathcal{O}(pq^2 + p^2) \approx \mathcal{O}(pq^2)$. The next step has $\mathcal{O}(p^2)$ time complexity to calculate the eigenvalue and corresponding eigenvector (which is a basis vector) using the Power method. On the other hand, another basis vector can be calculated with time complexity $\mathcal{O}(pq^2 + pq)$. So, step 6(II)(ii) has total complexity $\mathcal{O}(p^2 + pq^2 + pq) \approx \mathcal{O}(pq^2)$. The total time complexity for computing canonical variables

Algorithm 4.1 FaRoC: Fast and Robust Supervised CCA

```
Input: Two multidimensional variables $\mathcal{X}_1$ and $\mathcal{X}_2$.
Output: A set $\mathbb{S}$ of $\mathcal{D}$ selected features.
Calculate the cross-covariance matrix $\mathcal{C}_{12} \in \Re^{m_1 \times m_2}$ of $\mathcal{X}_1$ and $\mathcal{X}_2$ using (3.2) of Chapter 3.
Calculate the covariance matrices $\mathcal{C}_{11} \in \Re^{m_1 \times m_1}$ and $\mathcal{C}_{22} \in \Re^{m_2 \times m_2}$ of $\mathcal{X}_1$ and $\mathcal{X}_2$ using (3.3) and (3.4) of Chapter 3, respectively.
Calculate the eigenvalues $\Lambda_1 \in \Re^{m_1}$ and $\Lambda_2 \in \Re^{m_2}$ of $\mathcal{C}_{11}$ and $\mathcal{C}_{22}$, along with corresponding eigenvectors $\Psi_1$ and $\Psi_2$ using Jacobi method.
If $m_1 \leqslant m_2$, calculate $\mathcal{H}_{11}$ using (3.30) of Chapter 3, otherwise calculate $\tilde{\mathcal{H}}_{11}$ using (3.31) of Chapter 3.
Initialize $\mathbb{S} \leftarrow \emptyset$ and $t = 1$.
for each $t \leqslant D$ do
    (I) Initialize $\mathbb{C} \leftarrow \emptyset$.
    (II) for each $(i, j)$-th regularization parameters of $\mathfrak{r}_1$ and $\mathfrak{r}_2$, where $\forall i \in \{1, 2, \cdots, \mathfrak{t}_1\}$ and $\forall j \in \{1, 2, \cdots, \mathfrak{t}_2\}$; if $m_1 \leqslant m_2$ (respectively, $m_1 > m_2$) do
        (i) If $t = 1$, calculate $\mathcal{H}_{ij}(t)$ using (4.2.1) (respectively, $\tilde{\mathcal{H}}_{ij}(t)$ using (4.2.1)), otherwise using (4.19) (respectively, using (4.20)).
        (ii) Calculate largest eigenvalue $\rho_{t_{ij}}$ and eigenvector $w_{1_{t_{ij}}}$ (respectively, $w_{2_{t_{ij}}}$) of matrix $\mathcal{H}_{ij}(t)$ (respectively, $\tilde{\mathcal{H}}_{ij}(t)$) using Power method and (4.2.2), where $w_{1_{t_{ij}}}$ and $w_{2_{t_{ij}}}$ are the $t$-th basis vectors.
        (iii) Calculate the $t$-th pair of canonical variables $\{\mathcal{U}_{1_{t_{ij}}}, \mathcal{U}_{2_{t_{ij}}}\}$ using (3.11) of Chapter 3.
        (iv) Compute the $t$-th extracted feature $\mathcal{F}_{t_{ij}}$ corresponding to $(i, j)$-th pair of regularization parameters using (3.12) of Chapter 3.
        (v) Calculate the relevance $\gamma_{\mathcal{F}_{t_{ij}}}(\mathbb{D})$ of $\mathcal{F}_{t_{ij}}$ using (3.47) of Chapter 3.
        (vi) Calculate the significance $\sigma_{\{\mathcal{F}_{t_{ij}}, \mathcal{F}_\ell\}}(\mathbb{D}, \mathcal{F}_{t_{ij}})$ of $\mathcal{F}_{t_{ij}}$ with respect to each $\mathcal{F}_\ell$ of the already-selected features of $\mathbb{S}$ using (3.33) of Chapter 3.
        (vii) Add $\mathcal{F}_{t_{ij}}$ to $\mathbb{C}$ if its significance is non-zero with respect to all of the selected features of $\mathbb{S}$. In effect, $\mathbb{C} = \mathbb{C} \bigcup \mathcal{F}_{t_{ij}}$.
    (III) end for
    (IV) If $\mathbb{C} \neq \emptyset$, select a feature as $t$-th feature $\mathcal{F}_t$ from all the features of $\mathbb{C}$, which maximizes the following condition:
         $$\begin{array}{ll} \gamma_{\mathcal{F}_{t_{ij}}}(\mathbb{D}) & \text{if } k = 1 \\ \gamma_{\mathcal{F}_{t_{ij}}}(\mathbb{D}) + \frac{1}{t-1} \sum\limits_{\mathcal{F}_\ell \in \mathbb{S}} \sigma_{\{\mathcal{F}_{t_{ij}}, \mathcal{F}_\ell\}}(\mathbb{D}, \mathcal{F}_{t_{ij}}) & \text{otherwise.} \end{array} \qquad (4.21)$$
         As a result of that, $\mathbb{S} = \mathbb{S} \bigcup \mathcal{F}_t$ and $t = t + 1$.
end for
Stop.
```

$\mathcal{U}_1$ and $\mathcal{U}_2$ in step 6(II)(iii) is $\mathcal{O}(pn + qn)$. The computational complexity to extract a feature $\mathcal{F}$ is $\mathcal{O}(n)$. The time complexity to compute both relevance and significance of a feature is same, which is $\mathcal{O}(cn)$. Hence, the total complexity to execute the loop $(\mathfrak{t}_1 \times \mathfrak{t}_2)$ times is $\mathcal{O}(\mathfrak{t}_1\mathfrak{t}_2(pq^2 + pq^2 + pn + qn + n + cn)) \approx \mathcal{O}(\mathfrak{t}_1\mathfrak{t}_2 pq^2)$. The selection of a feature from $(\mathfrak{t}_1 \times \mathfrak{t}_2)$ candidate features by maximizing relevance and significance, which is carried out in step 6(IV), has complexity $\mathcal{O}(\mathfrak{t}_1\mathfrak{t}_2)$. Hence, the total complexity to execute the loop $\mathcal{D}$ times is $\mathcal{O}(\mathcal{D}(\mathfrak{t}_1\mathfrak{t}_2 pq^2 + \mathfrak{t}_1\mathfrak{t}_2)) \approx \mathcal{O}(\mathcal{D}\mathfrak{t}_1\mathfrak{t}_2 pq^2)$. Hence, the overall computational complexity of the proposed algorithm is $\mathcal{O}(q^3 + \mathcal{D}\mathfrak{t}_1\mathfrak{t}_2 pq^2) \approx \mathcal{O}(q^2(q + \mathcal{D}\mathfrak{t}_1\mathfrak{t}_2 p))$.

4.3 Performance Analysis

The performance of the proposed feature extraction algorithm, termed as FaRoC, is extensively studied and compared with that of some existing CCA based algorithms. The algorithms compared are CCA, RCCA, several variants of SRCCA using t-test (SRCCA$_{\mathrm{TT}}$) [91], Wilcoxon rank sum test (SRCCA$_{\mathrm{WR}}$) [91], Wilks's lambda test (SRCCA$_{\mathrm{WL}}$) [91], mutual information (SRCCA$_{\mathrm{MI}}$), rough hypercuboid (SRCCA$_{\mathrm{RH}}$), and CuRSaR [174] presented in Chapter 3. The performance of the rough hypercuboid approach is also compared with that of mutual information in the proposed feature extraction framework. The support vector machine (SVM) [274] with linear kernels is used to compute this error. All the algorithms are implemented in C language and run in Ubuntu 14.04 LTS having machine configuration Intel(R) Core(TM) i7-4790 CPU @ 3.60GHz×8 and 32 GB RAM. Both $\mathfrak{r}_1$ and $\mathfrak{r}_2$ are varied within $[0.0, 1.0]$ with 0.1 as common difference. The source code of the proposed FaRoC algorithm, written in C language, is available at `https://www.isical.ac.in/~bibl/results/faroc/faroc.html`.

Both 10-fold cross-validation (CV) and training-testing are performed to assess the performance of different algorithms. To analyze the statistical significance of the derived results in 10-fold CV, paired-t test (one-tailed), Wilcoxon signed rank test (one-tailed) and Friedman test (one-tailed), with a 95% confidence level, are used to compute the p-values. For training-testing, the randomly selected 50% samples from each class are used for training and the rest are used for testing purpose for each of the data sets. For each data set, 25 top-ranked correlated features are selected for the analysis.

Five multimodal data sets, namely, glioblastoma multiforme (GBM), lung (LUNG), kidney (KIDNEY), lower grade glioma (LGG), and ovarian serous cystadenocarcinoma (OV), are used in the current research work, each having two different modalities, namely, DNA methylation (mDNA) and RNA. The other hand, the details of the gene (RNA) have been taken from RNA sequences in LUNG, KIDNEY, and LGG data sets, while gene expression provides gene-related information in the GBM and OV data sets. These data sets are downloaded from TCGA (`https://cancergenome.nih.gov/`). All five data sets are summarized in Table 3.1 of Chapter 3 and briefly described in Appendix A.

4.3.1 Importance of Rough Hypercuboid Approach

In the proposed FaRoC algorithm, both relevance and significance measures of an extracted feature are computed based on the concept of hypercuboid equivalence partition matrix. The relevance of an extracted feature with respect to decision attribute set or class labels is calculated as per (3.47) of Chapter 3, while the significance of a feature is calculated using (3.33) of Chapter 3 with respect to the already-extracted features. In this regard, it should be noted that other measures like mutual information can also be employed for computing both the significance and relevance of a feature. Figure 4.1 and Figure 4.2 establish the importance of the rough hypercuboid approach over mutual information considering all five data sets using both 10-fold CV and training-testing. All the results reported in Figure 4.1 and Figure 4.2 confirm that the performance of hypercuboid equivalence partition matrix of rough hypercuboid approach is significantly better than that of mutual information, irrespective of the data sets used and number of extracted features.

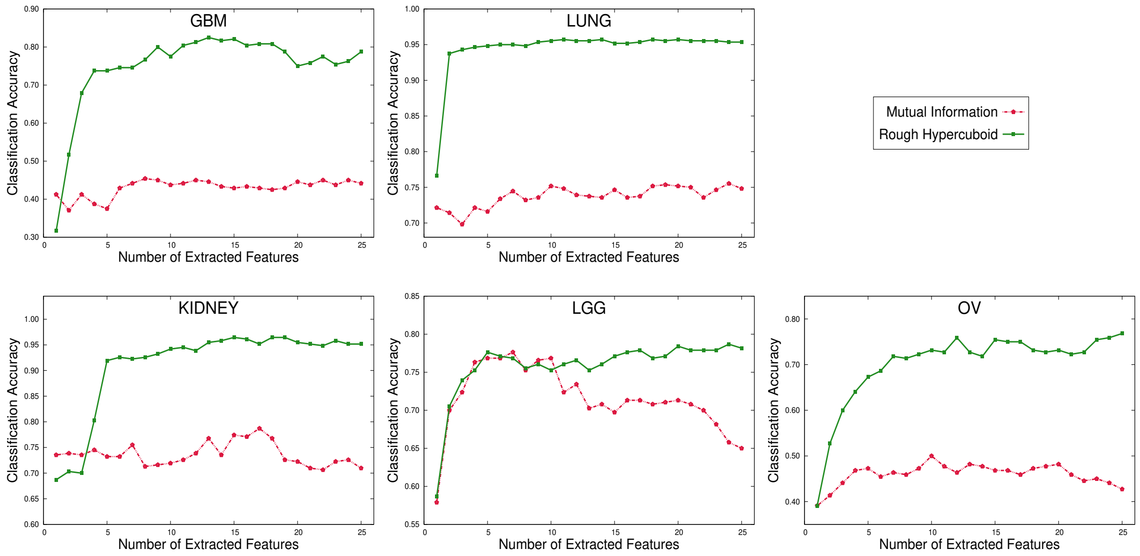


Figure 4.1: Variation of classification accuracy with respect to number of extracted features using mutual information and rough hypercuboid in the proposed framework for 10-fold CV.

Table 4.1: Classification Accuracy and Execution Time for Mutual Information and Rough Hypercuboid

Data	Measure	Accuracy	Accuracy and Significance Analysis for 10-Fold CV						Time
Sets		(Train-Test)	Mean	Median	StdDev	Paired-t:p	Wilcoxon:p	Friedman:p	(in sec.)
GBM	MI	0.343	0.442	0.458	0.145	**4.43E-05**	**2.47E-03**	**1.57E-03**	3057.4
	RH	**0.771**	**0.788**	**0.792**	0.050	-	-	-	3116.7
LUNG	MI	0.758	0.748	0.768	0.107	**7.28E-05**	**2.47E-03**	**1.57E-03**	3006.3
	RH	**0.872**	**0.954**	**0.946**	0.034	-	-	-	2340.3
KIDNEY	MI	0.618	0.710	0.694	0.126	**1.71E-04**	**3.42E-03**	**1.14E-02**	3952.9
	RH	**0.974**	**0.952**	**0.935**	0.023	-	-	-	3576.0
LGG	MI	0.624	0.650	0.618	0.145	**5.31E-03**	**6.31E-03**	**1.96E-02**	3150.7
	RH	**0.860**	**0.782**	**0.803**	0.074	-	-	-	2803.3
OV	MI	0.402	0.427	0.432	0.205	**3.33E-04**	**5.81E-03**	**4.68E-03**	3365.6
	RH	**0.951**	**0.768**	**0.773**	0.045	-	-	-	2921.0

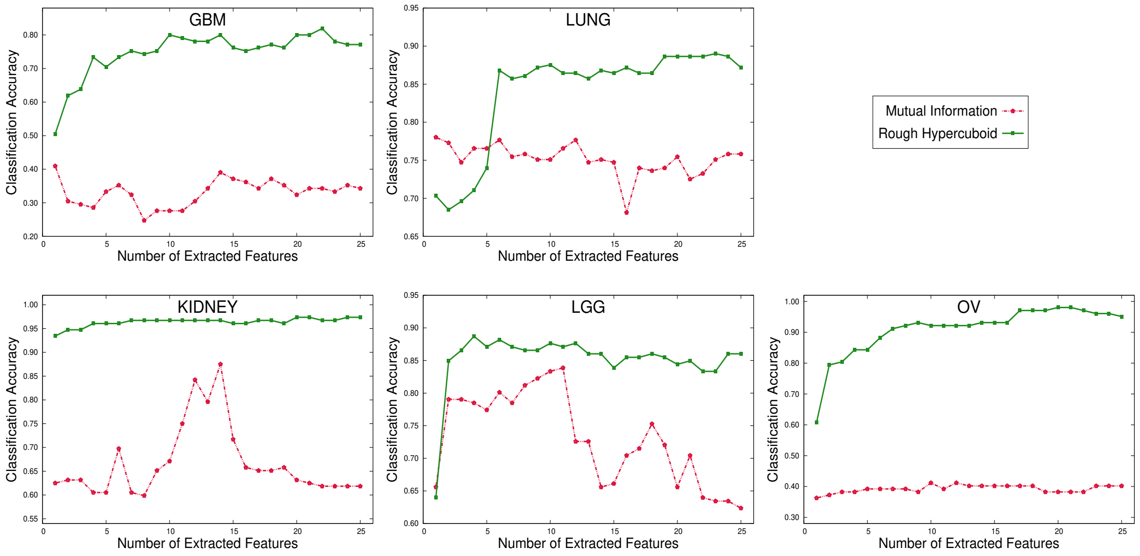
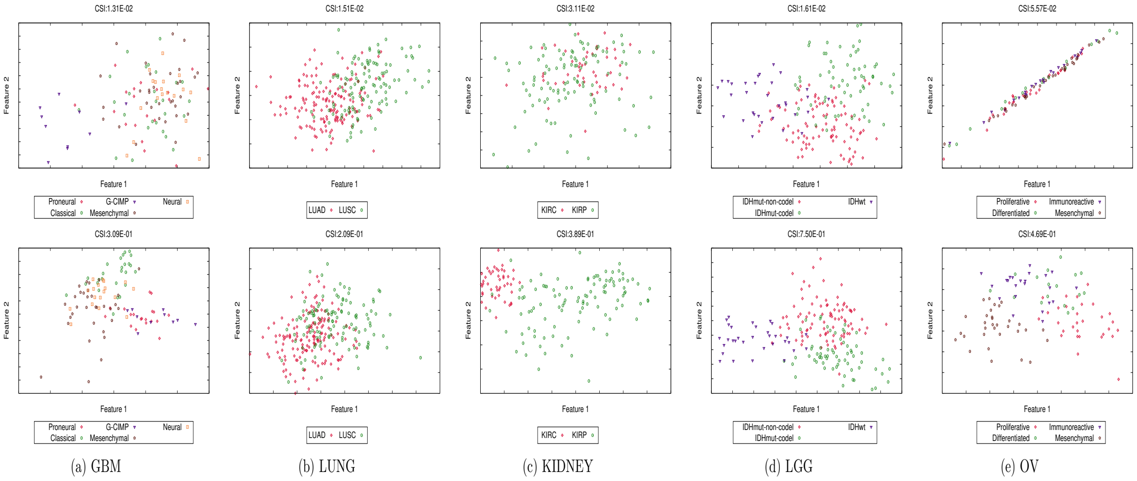


Figure 4.2: Variation of classification accuracy with respect to number of extracted features using mutual information and rough hypercuboid in the proposed framework for training-testing.



(a) GBM (b) LUNG (c) KIDNEY (d) LGG (e) OV

Figure 4.3: Scatter plots for mutual information (top row) and rough hypercuboid (bottom row) in the proposed framework, along with class separability index.

Moreover, Figure 4.3 compares the performance of the rough hypercuboid approach and mutual information using the scatter plots and class separability index, on five data sets. The x-axis and y-axis of each plot represent the first and second extracted features. The value of the class separability index (CSI) is also reported at the top of each figure. From the results reported in Figure 4.3, it is evident that the CSI of the extracted features using mutual information is lower compared to that of hypercuboid equivalence partition matrix. The qualitative results in Figure 4.3 show that the relevant and significant features extracted using mutual information cannot separate the classes properly, particularly for the OV data set. On the other hand, the results reported in Figure 4.3 demonstrate that the features which are most relevant and significant according to rough hypercuboid equivalence partition matrix can separate the classes properly, even on the OV data set.

Table 4.1 compares the performance of hypercuboid equivalence partition matrix and mutual information in terms of the classification accuracy, both for training-testing and 10-fold CV. To perform the statistical significance analysis, the p-values computed using different tests are also reported in Table 4.1. Comparing the results in Table 4.1, it is apparent that rough hypercuboid approach attains higher mean and median accuracy than mutual information, irrespective of the data sets used. In all 15 cases, the hypercuboid equivalence partition matrix achieves significantly better (marked in bold) p-values than mutual information.

All the results reported in Figure 4.1, Figure 4.2, Figure 4.3, and Table 4.1 demonstrate that the hypercuboid equivalence partition matrix determines more relevant and significant features than mutual information does. The significantly better performance of the rough hypercuboid based proposed approach is obtained due to the fact that the quality of an extracted feature set, in rough hypercuboid approach, is evaluated by the hypercuboid equivalence partition matrix that makes use of supervised information of sample categories in the granulation process. Also, it provides an efficient way to calculate relevance and significance in approximation spaces. The proposed FaRoC algorithm, in effect, is able to generate a reduced set of significant and relevant features from multimodal data sets.

4.3.2 Importance of Sequential Feature Generation

The proposed FaRoC algorithm extracts $\mathcal{D}$ features sequentially from five multidimensional data sets, based on their individual relevance with respect to class label and significance with respect to the already-extracted features. However, $\mathcal{D}$ features can be extracted simultaneously by the maximum relevance-maximum significance criterion as done in CuRSaR [174], presented in Chapter 3. In order to establish the importance of sequential feature generation of the proposed FaRoC algorithm over simultaneous feature generation by CuRSaR, extensive experimental results are reported in Figure 4.4 and Figure 4.5. The results reported in Figure 4.4 and Figure 4.5 establish the fact that the FaRoC outperforms CuRSaR in almost all the cases, irrespective of the data sets and the number of extracted features. Comparing the bottom row of Figure 3.6 of Chapter 3 with the bottom row of Figure 4.3, it is evident that the FaRoC can separate the classes more accurately than CuRSaR. Also, the features extracted using FaRoC have higher CSI values than that of CuRSaR, except for GBM data set.

Table 4.2 reports the statistical significance analysis of CuRSaR with respect to FaRoC. The proposed FaRoC algorithm attains significantly better p-values (marked in bold) than CuRSaR in 10 cases, out of total 15 cases, considering 95% confidence level. On the other hand, the FaRoC provides better but not significant (marked in italics) p-values in only 5 cases, for Friedman test on all data sets except the LGG data and paired-t test on GBM data set. The better performance of the FaRoC algorithm over CuRSaR is achieved due to the fact that the FaRoC considers different pairs of regularization parameters for different features, while the CuRSaR extracts a set of features for a fixed pair of parameters. In effect, the extracted features are more relevant and significant for the FaRoC than the CuRSaR.

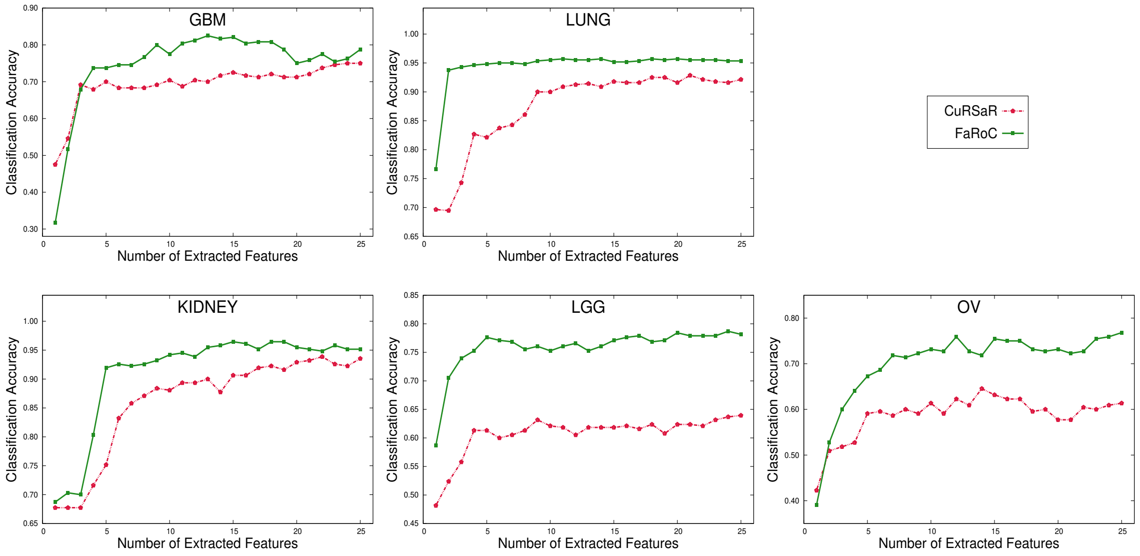


Figure 4.4: Variation of classification accuracy with respect to number of extracted features for CuRSaR and FaRoC in case of 10-fold CV.

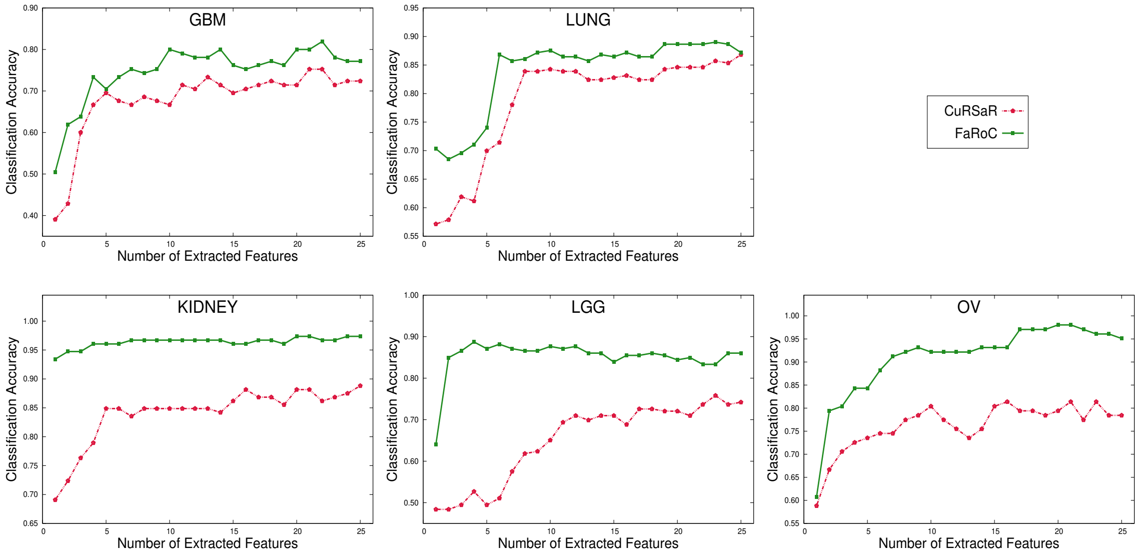


Figure 4.5: Variation of classification accuracy with respect to number of extracted features for CuRSaR and FaRoC in case of training-testing.

4.3.3 Comparative Performance Analysis

Finally, Figure 4.6 and Figure 4.7, and Table 4.2 compare the performance of the proposed FaRoC algorithm with that of several existing SRCCA algorithms, namely, SRCCA$_{\mathrm{TT}}$ [91], SRCCA$_{\mathrm{WR}}$ [91], SRCCA$_{\mathrm{WL}}$ [91], SRCCA$_{\mathrm{MI}}$, and SRCCA$_{\mathrm{RH}}$. Results are reported in Figure 4.6 and Figure 4.7 for different number of extracted features on all five data sets, while Table 4.2 reports the p-values to analyze the statistical significance of the results obtained

Table 4.2: Statistical Significance Analysis of Different Algorithms

Different	Data	p-values for 10-Fold CV		
Algorithms	Sets	Paired-t	Wilcoxon	Friedman
CCA	GBM	**2.36E-07**	**2.50E-03**	**1.57E-03**
RCCA		**2.93E-07**	**2.46E-03**	**1.57E-03**
SRCCA$_{\mathrm{TT}}$		**6.28E-07**	**2.50E-03**	**1.57E-03**
SRCCA$_{\mathrm{WL}}$		**1.77E-07**	**2.49E-03**	**1.57E-03**
SRCCA$_{\mathrm{WR}}$		**1.77E-07**	**2.49E-03**	**1.57E-03**
SRCCA$_{\mathrm{MI}}$		**7.23E-04**	**2.52E-03**	**1.57E-03**
SRCCA$_{\mathrm{RH}}$		**2.02E-03**	**7.19E-03**	*5.78E-02*
CuRSaR		*5.40E-02*	**4.49E-02**	*5.88E-02*
CCA	LUNG	**1.85E-05**	**2.50E-03**	**1.57E-03**
RCCA		**1.11E-05**	**2.53E-03**	**1.57E-03**
SRCCA$_{\mathrm{TT}}$		**6.62E-05**	**2.52E-03**	**1.57E-03**
SRCCA$_{\mathrm{WL}}$		**8.17E-06**	**2.53E-03**	**1.57E-03**
SRCCA$_{\mathrm{WR}}$		**4.20E-05**	**2.52E-03**	**1.57E-03**
SRCCA$_{\mathrm{MI}}$		**4.50E-05**	**2.53E-03**	**1.57E-03**
SRCCA$_{\mathrm{RH}}$		**7.84E-05**	**2.52E-03**	**1.57E-03**
CuRSaR		**1.76E-02**	**1.71E-02**	*5.88E-02*
CCA	KIDNEY	**1.05E-04**	**2.52E-03**	**1.57E-03**
RCCA		**1.16E-03**	**3.79E-03**	**2.70E-03**
SRCCA$_{\mathrm{TT}}$		**5.12E-04**	**2.49E-03**	**1.57E-03**
SRCCA$_{\mathrm{WL}}$		**5.12E-04**	**2.49E-03**	**1.57E-03**
SRCCA$_{\mathrm{WR}}$		**5.12E-04**	**2.49E-03**	**1.57E-03**
SRCCA$_{\mathrm{MI}}$		**4.11E-04**	**2.49E-03**	**1.57E-03**
SRCCA$_{\mathrm{RH}}$		**1.78E-03**	**3.42E-03**	**1.14E-02**
CuRSaR		**4.79E-02**	**4.78E-02**	*1.03E-01*
CCA	LGG	**9.35E-07**	**2.50E-03**	**1.57E-03**
RCCA		**2.39E-08**	**2.50E-03**	**1.57E-03**
SRCCA$_{\mathrm{TT}}$		**5.51E-08**	**2.50E-03**	**1.57E-03**
SRCCA$_{\mathrm{WL}}$		**1.61E-07**	**2.39E-03**	**1.57E-03**
SRCCA$_{\mathrm{WR}}$		**1.61E-07**	**2.39E-03**	**1.57E-03**
SRCCA$_{\mathrm{MI}}$		**1.68E-03**	**6.23E-03**	**1.14E-02**
SRCCA$_{\mathrm{RH}}$		**1.91E-02**	**3.31E-02**	*2.06E-01*
CuRSaR		**6.86E-04**	**3.92E-03**	**1.14E-02**
CCA	OV	**7.62E-06**	**2.53E-03**	**1.57E-03**
RCCA		**8.12E-07**	**2.52E-03**	**1.57E-03**
SRCCA$_{\mathrm{TT}}$		**4.67E-07**	**2.53E-03**	**1.57E-03**
SRCCA$_{\mathrm{WL}}$		**4.67E-07**	**2.53E-03**	**1.57E-03**
SRCCA$_{\mathrm{WR}}$		**4.67E-07**	**2.53E-03**	**1.57E-03**
SRCCA$_{\mathrm{MI}}$		**2.12E-05**	**3.82E-03**	**2.70E-03**
SRCCA$_{\mathrm{RH}}$		**5.03E-04**	**5.40E-03**	**1.96E-02**
CuRSaR		**2.03E-02**	**2.92E-02**	*1.57E-01*

using existing algorithms with respect to the proposed FaRoC algorithm. Comparing Table 3.3 of Chapter 3 with Table 4.1, it is evident that the proposed FaRoC algorithm attains highest mean and median accuracy, irrespective of the data sets. The proposed algorithm attains significantly better p-values (marked in bold) than several existing data integration algorithms in 103 cases of total 105 cases, considering 95% confidence level. On the other

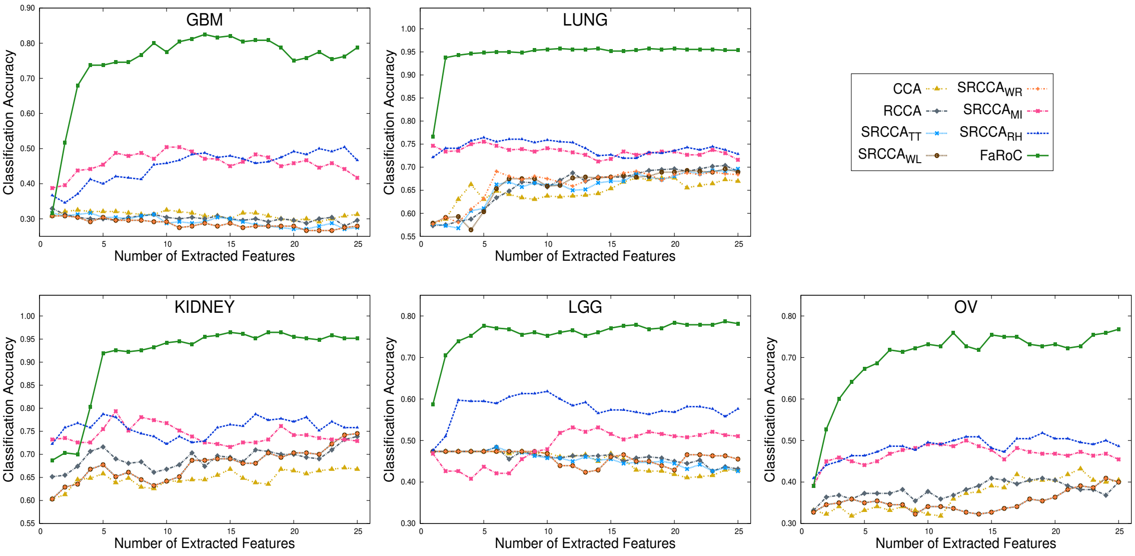


Figure 4.6: Variation of classification accuracy with respect to number of extracted features for several existing algorithms and the proposed (FaRoC) algorithm on 10-fold CV.

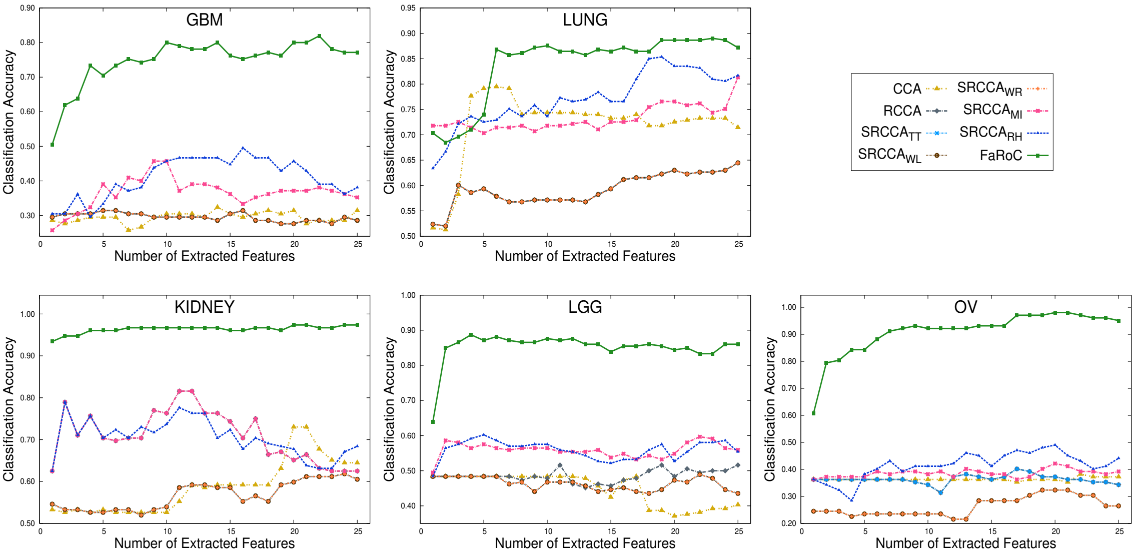


Figure 4.7: Variation of classification accuracy with respect to number of extracted features for several existing algorithms and the proposed (FaRoC) algorithm on training-testing.

hand, the proposed FaRoC algorithm provides better but not significant (marked in italics) p-values in only 2 cases. Comparing the scatter plots at bottom row of Figure 4.3 and the first seven columns of Figure 3.9 of Chapter 3, it is noticeable that the proposed algorithm is able to separate different classes using the first two extracted features only for all five data sets, which is also evident from the corresponding class separability index values. On the other hand, the classes are hardly separable using all existing algorithms.

4.4 Conclusion

This chapter presents a new feature extraction algorithm, termed as FaRoC, for two multidimensional data sets. The merits of CCA and rough sets have been integrated judiciously to develop the proposed algorithm. To establish the relation between regularization parameters and CCA, a theoretical formulation has been presented based on spectral decomposition, which helps the proposed FaRoC algorithm to extract the required number of correlated features sequentially. The proposed algorithm extracts a new feature from two multidimensional data sets by maximizing its relevance with respect to class label and significance with respect to already-extracted features. The hypercuboid equivalence partition matrix of rough hypercuboid approach has been used to compute both the relevance and significance of a feature. The optimum regularization parameters of CCA have been determined using the equivalence partition matrix. The effectiveness of the proposed algorithm, along with a comparison with other algorithms, has been demonstrated considering two different modalities, namely, RNA and mDNA. The hypercuboid equivalence partition matrix is found to be successful in extracting relevant and significant features from multimodal high dimensional real-life data sets. The current formulation shows the utility of rough hypercuboid approach and canonical correlation analysis with respect to knowledge discovery tasks.

Both CuRSaR and FaRoC, presented in Chapter 3 and Chapter 4, respectively, can only account for two sets of variables. The multiset canonical correlation analysis (MCCA) is a well-known statistical method for multi-view data integration. However, the existing algorithms to find the multiset canonical variables are computationally very expensive, which restricts the application of the MCCA in real-life big data analysis. The covariance matrix of each high-dimensional view may also suffer from the singularity problem due to the limited number of samples. Moreover, the MCCA based existing feature extraction algorithms are, in general, unsupervised in nature. In this regard, a new supervised feature extraction algorithm is introduced in the next chapter, which integrates multimodal multidimensional data sets by solving maximal correlation problem of the MCCA. The analytical formulation enables efficient computation of the multiset canonical variables under supervised ridge regression optimization technique.

Chapter 5

Regularized Discriminant Multiset CCA

5.1 Introduction

As mentioned in Chapter 3 and Chapter 4, there has been a growing interest in multi-view learning in recent years. In many real-world applications, multiple views or modalities provide relevant and complementary information about a specific problem. The integration of relevant and non-redundant features from a wide range of modalities is expected to result in better predictors, as compared to any individual modality [85,128,158,247,250,300]. The naive integration of multiblock data generates a concatenated feature set, which intensifies the "curse of dimensionality" problem. To overcome the issues associated with dimensionality, scale, and kernel-based weighting, canonical correlation analysis (CCA) [112] can be used to analyze the inter-dependency between two multidimensional variables. In real-world data analysis, the dimension of feature space p is significantly higher compared to the limited number of samples n. This 'large p and small n' problem makes the features of a data set highly collinear, which leads to ill-conditioning of the covariance matrix of the multidimensional variable. To deal with the singularity issue of covariance matrices, regularized CCA (RCCA) has been introduced in [278]. However, both CCA and RCCA fail to take complete advantage of available information of sample categories, as they are unsupervised in nature [266]. To achieve better classification performance, both discriminant CCA [253] and supervised RCCA [91, 174, 183] utilize the supervised class information.

The CCA [112], RCCA [278], and different variants of supervised CCA [91, 174, 183, 253, 266] can only account for two sets of variables. Multiset canonical correlation analysis (MCCA) [110] extends the CCA for more than two views by finding a linear subspace which maximizes the correlations between all the views. The objective of the MCCA is to find linear relationships among several blocks of variables. While correlation-based MCCA methods [109, 110, 135] consider only between-block information, covariance-based methods [61, 100, 262, 263] take into account both the between-block and within-block information. In [100, 135, 262], several criteria have been studied to extend the CCA for three or more sets of variables. Among all these criteria, the sum of correlations (SUMCOR) is widely used to integrate more than two sets of multidimensional variables [84, 131, 205], although the maximum variance (MAXVAR) criterion has also been used in [83] to address the nonnegativity and sparsity on the canonical components. In [43], a novel graph-regularized MCCA (GMCCA) algorithm has been proposed to minimize the distance among the canonical variables based on MAXVAR criterion. The importance of the kernel in the GMCCA has been established in [43], where graph-regularized kernel MCCA (GMKCCA) has been introduced. There exist several important scientific fields [63, 69, 164, 239, 260] where the MCCA has been successfully applied to integrate the information of several multidimensional variables. In recent years, the deep learning framework has been used to learn non-linear transformations from different modalities by computing the canonical variables of the MCCA [40, 55, 165, 244].

Regularized generalized CCA (RGCCA) is a generalization of the regularized CCA for three or more sets of variables [262]. It combines the power of multiblock data analysis methods and flexibility of partial least squares path modeling. To deal with the singularity issue of the covariance matrix of each multidimensional variable, an optimal shrinkage parameter is estimated for each variable, which reduces the values of off-diagonal elements of each covariance matrix, while diagonal elements remain same [262]. Instead of reducing the off-diagonal elements of a covariance matrix, if the diagonal elements could be increased by adding a regularization parameter, the search space for finding the canonical variables will increase. This may help to improve the performance of multimodal data analysis. Moreover, all the MCCA based methods reported earlier are unsupervised in nature and do not utilize the available class label information. The Horst-Jacobi algorithm [110], which has Jacobi type recurrence structure, is the earliest iterative algorithm to solve the basis vectors of the MCCA. In [50], an improvement of the Horst-Jacobi algorithm, known as Gauss-Seidel algorithm, has been developed by adopting the Gauss-Seidel type iteration. However, both of these algorithms [50, 110] compute the basis vectors for each block in every iteration, which makes them computationally expensive.

In this regard, this chapter presents a new supervised feature extraction algorithm, termed as ReDMiCA (Regularized Discriminant Multi-View CCA), for multimodal data sets. It integrates multi-view multidimensional data sets by solving the maximal correlation problem (MCP). A new block matrix representation is introduced to compute the basis vectors of the MCCA, which reduces the computational cost for solving the canonical variables. The proposed algorithm deals with the 'large p and small n' issue of multidimensional data sets by using ridge regression optimization, where regularization parameters are optimized using the supervised information of available sample categories. A theoretical analysis is presented, which helps to compute the multiset canonical variables under ridge regression from the canonical variable of the modality having lowest dimension. The analysis also ensures that the proposed algorithm can generate sequentially the required number of relevant and significant features, without extracting all plausible features. The proposed algorithm, in turn, has significantly lesser complexity than the existing methods. The effectiveness of the proposed algorithm, along with a comparison with state-of-the-art methods, is established on several benchmark and real-life multiblock data. Some of the results of this chapter are reported in [185].

The rest of the chapter is organized as follows: Section 5.2 outlines the basic principles of MCCA. Section 5.3 presents the proposed multi-view algorithm. The effectiveness of the proposed multi-view data integration algorithm, along with a comparative performance analysis with state-of-the-art algorithms on different multi-view benchmark and omics data sets, is presented in Section 5.4. Concluding remarks are provided in Section 5.5.

5.2 Basics of Multiset CCA

The CCA, proposed by Hotelling [112], refers to the MCP, where the goal is to find the linear combination of one set of variables that correlates maximally with the linear combination of another set of variables. If the maximal correlation can be satisfactorily established, then one set of variables can be used to predict the other one. The MCCA [110] is a well-known statistical method, which is used to analyze the linear relations among more than two sets of multidimensional variables. It extracts most correlated latent features from $\mathcal{M}$ data sets, $\mathcal{X}_1 \in \Re^{m_1 \times n}$, $\mathcal{X}_2 \in \Re^{m_2 \times n}$, $\cdots$, and $\mathcal{X}_{\mathcal{M}} \in \Re^{m_{\mathcal{M}} \times n}$, by solving the MCP. Each column in $\mathcal{X}_i$ represents one of the n samples, whereas each row corresponds to one of the m_i variables, where m_i is the dimension of each sample of the i-th variable $\mathcal{X}_i$. Without loss of generality, it is assumed that each multidimensional variable is centered to have zero mean across the samples, that is, $\mathcal{E}[\mathcal{X}_i] = 0, \forall i \in \{1, 2, \cdots, \mathcal{M}\}$.

The main objective of the MCCA is to find the optimal basis vectors $w_1 \in \Re^{m_1}$, $w_2 \in \Re^{m_2}$, $\cdots$, and $w_{\mathcal{M}} \in \Re^{m_{\mathcal{M}}}$ that maximize some merit functions under certain constraints. Some of the constraints are as follows:

i) the basis vectors are unit vectors within each set, that is, ${w_i}^T w_i = 1$;

ii) the sum of the variances of the canonical variables is unity, that is, $\sum\limits_{i=1}^{\mathcal{M}} \mathcal{U}_i {\mathcal{U}_i}^T = \sum\limits_{i=1}^{\mathcal{M}} {w_i}^T \mathcal{C}_{ii} w_i = 1$;

where A^T denotes the transpose of the matrix A, $\mathcal{U}_i = {w_i}^T \mathcal{X}_i$ is the i-th canonical variable and $\mathcal{C}_{ii} = \mathcal{X}_i {\mathcal{X}_i}^T \in \Re^{m_i \times m_i}$ denotes the covariance matrix of $\mathcal{X}_i$. Generally, the following constrained optimization problem is considered to maximize the sum of correlations across all pairs of modalities [110, 205]:

$$\max_{\{w_i\}_{i=1}^{\mathcal{M}}} \quad \sum_{i=1}^{\mathcal{M}} \sum_{\substack{j=1 \\ j \neq i}}^{\mathcal{M}} {w_i}^T \mathcal{C}_{ij} w_j; \tag{5.1}$$

$$\text{subject to} \quad \sum_{i=1}^{\mathcal{M}} {w_i}^T \mathcal{C}_{ii} w_i = 1; \tag{5.2}$$

where $\mathcal{C}_{ij} = \mathcal{X}_i {\mathcal{X}_j}^T \in \Re^{m_i \times m_j}$ is the cross-covariance matrix of $\mathcal{X}_i$ and $\mathcal{X}_j$. The above approach is known as the SUMCOR. The Lagrangian of this problem is given by [205]

$$\mathcal{L} = \sum_{i=1}^{\mathcal{M}} \sum_{\substack{j=1 \\ j \neq i}}^{\mathcal{M}} {w_i}^T \mathcal{C}_{ij} w_j - \rho \left(\sum_{i=1}^{\mathcal{M}} {w_i}^T \mathcal{C}_{ii} w_i - 1 \right); \tag{5.3}$$

where ρ is the Lagrange multiplier. Differentiating $\mathcal{L}$ with respect to w_i, and setting the vectors of derivative to zero, we obtain the following equation [50, 314]:

$$\sum_{\substack{j=1\\ j\neq i}}^{\mathcal{M}} \mathcal{C}_{ii}{}^{-1}\mathcal{C}_{ij}w_j = \rho w_i, \qquad \forall i \in \{1, 2, \cdots, \mathcal{M}\};$$

$$\Rightarrow \begin{bmatrix} \mathbf{0}^{[m_1]} & \mathcal{C}_{11}^{-1}\mathcal{C}_{12} & \cdots \mathcal{C}_{11}^{-1}\mathcal{C}_{1\mathcal{M}} \\ \mathcal{C}_{22}^{-1}\mathcal{C}_{21} & \mathbf{0}^{[m_2]} & \cdots \mathcal{C}_{22}^{-1}\mathcal{C}_{2\mathcal{M}} \\ \vdots & \vdots & \ddots \quad \vdots \\ \mathcal{C}_{\mathcal{M}\mathcal{M}}{}^{-1}\mathcal{C}_{\mathcal{M}1} & \mathcal{C}_{\mathcal{M}\mathcal{M}}{}^{-1}\mathcal{C}_{\mathcal{M}2} & \cdots \quad \mathbf{0}^{[m_{\mathcal{M}}]} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_{\mathcal{M}} \end{bmatrix} = \rho \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_{\mathcal{M}} \end{bmatrix} \Rightarrow \mathcal{A}_{\mathcal{M}}\mathcal{W}_{\mathcal{M}} = \Gamma\mathcal{W}_{\mathcal{M}}; \qquad (5.4)$$

where $\mathcal{A}_{\mathcal{M}} \in \Re^{\sum_{i=1}^{\mathcal{M}} m_i \times \sum_{i=1}^{\mathcal{M}} m_i}$ and $\mathcal{W}_{\mathcal{M}} \in \Re^{\sum_{i=1}^{\mathcal{M}} m_i}$, and Γ is a diagonal matrix with same ($= \rho$) diagonal elements.

Either Horst-Jacobi [110] or Gauss-Seidel algorithm [50] can be used to solve the basis vectors of the MCCA. Both the algorithms compute $\mathcal{M}$ basis vectors $\{w_1, w_2, \cdots, w_{\mathcal{M}}\}$ for $\mathcal{M}$ blocks in every iteration, which makes these algorithms computationally very expensive. The computational complexity to calculate each basis vector is $\mathcal{O}(m_{\mathcal{M}}^2 n + m_{\mathcal{M}} m_{\mathcal{M}} - 1n + m_{\mathcal{M}}^3 + m_{\mathcal{M}}^2 m_{\mathcal{M}} - 1 + m_{\mathcal{M}} m_{\mathcal{M}} - 1) \approx \mathcal{O}(m_{\mathcal{M}}^3)$, where $m_{\mathcal{M}} \geqslant m_{\mathcal{M}} - 1 \geqslant \cdots \geqslant m_1$. Hence, each iteration has the time complexity $\mathcal{O}(\mathcal{M}m_{\mathcal{M}}^3)$ to compute all the basis vectors. If η denotes the maximum number of iterations required to converge, each algorithm has $\mathcal{O}(\eta\mathcal{M}m_{\mathcal{M}}^3)$ time complexity.

5.3 ReDMiCA: Proposed Algorithm

This section presents a new sequential feature extraction algorithm, termed as ReDMiCA, which integrates the information of multidimensional multimodal data sets. To establish the importance of the proposed algorithm, some important analytical formulations are presented in this section.

5.3.1 Block Matrix to Solve Basis Vectors of MCCA

To find the basis vectors of the MCCA, either Horst-Jacobi or Gauss-Seidel algorithm can be used, which are iterative in nature, and hence time consuming. A new approach, using the properties of block matrix, is proposed in this regard to evaluate the basis vectors. Let us consider from (5.4) that

$$\mathcal{A}_{\mathcal{M}} = \begin{bmatrix} \mathbf{0}^{[m_1]} & a_{12} & \cdots & a_{1\mathcal{M}} \\ a_{21} & \mathbf{0}^{[m_2]} & \cdots & a_{2\mathcal{M}} \\ \vdots & \vdots & \ddots & \vdots \\ a_{\mathcal{M}1} & a_{\mathcal{M}2} & \cdots & \mathbf{0}^{[m_{\mathcal{M}}]} \end{bmatrix};$$

where $a_{ij} = {\mathcal{C}_{ii}}^{-1}\mathcal{C}_{ij}$. Now,

$$\mathcal{A}_{\mathcal{M}} = \begin{bmatrix} \begin{bmatrix} \mathbf{0}^{[m_1]} & a_{12} & \cdots & a_{1(\mathcal{M}-1)} \\ a_{21} & \mathbf{0}^{[m_2]} & \cdots & a_{2(\mathcal{M}-1)} \\ \vdots & \vdots & \ddots & \vdots \\ a_{(\mathcal{M}-1)1} & a_{(\mathcal{M}-1)2} & \cdots & \mathbf{0}^{[m_{\mathcal{M}-1}]} \end{bmatrix} & \begin{bmatrix} a_{1\mathcal{M}} \\ a_{2\mathcal{M}} \\ \vdots \\ a_{(\mathcal{M}-1)\mathcal{M}} \end{bmatrix} \\ \begin{bmatrix} a_{\mathcal{M}1} & a_{\mathcal{M}2} & \cdots & a_{\mathcal{M}(\mathcal{M}-1)} \end{bmatrix} & \mathbf{0}^{[m_{\mathcal{M}}]} \end{bmatrix}$$

$$= \begin{bmatrix} \mathcal{A}_{(\mathcal{M}-1)} & \mathbf{0}^{[\sum_{i=1}^{\mathcal{M}-1} m_i, m_{\mathcal{M}}]} \\ \mathbf{0}^{[m_{\mathcal{M}}, \sum_{i=1}^{\mathcal{M}-1} m_i]} & \mathbf{0}^{[m_{\mathcal{M}}]} \end{bmatrix} + \begin{bmatrix} \mathbf{0}^{[\sum_{i=1}^{\mathcal{M}-1} m_i]} & \Theta_{(\mathcal{M}-1)} \\ \Phi_{(\mathcal{M}-1)} & \mathbf{0}^{[m_{\mathcal{M}}]} \end{bmatrix}; \tag{5.5}$$

$$\text{where} \quad \Theta_{(\mathcal{M}-1)} = \begin{bmatrix} a_{1\mathcal{M}} \\ a_{2\mathcal{M}} \\ \cdots \\ a_{(\mathcal{M}-1)\mathcal{M}} \end{bmatrix}; \tag{5.6}$$

$$\text{and} \quad \Phi_{(\mathcal{M}-1)} = \begin{bmatrix} a_{\mathcal{M}1} & a_{\mathcal{M}2} & \cdots & a_{\mathcal{M}(\mathcal{M}-1)} \end{bmatrix}; \tag{5.7}$$

$\mathbf{0}^{[k]}$ and $\mathbf{0}^{[k,\ell]}$ denote the square null matrix with dimension k and rectangular null matrix with dimension $k \times \ell$, respectively. Let us assume that

$$\widehat{\mathcal{B}}_{(\mathcal{M}-1)} = \begin{bmatrix} \mathbf{0}^{[\sum_{i=1}^{\mathcal{M}-1} m_i]} & \Theta_{(\mathcal{M}-1)} \\ \Phi_{(\mathcal{M}-1)} & \mathbf{0}^{[m_{\mathcal{M}}]} \end{bmatrix}$$

has eigenvectors $\Psi = [\mathit{\Psi}_1, \mathit{\Psi}_2, \cdots, \mathit{\Psi}_p]$ with corresponding eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_p$, where

$$\mathit{\Psi}_t = \begin{bmatrix} \psi_{1t} \\ \psi_{2t} \end{bmatrix} \quad \text{and} \quad \begin{matrix} \psi_{1t} \in \Re^{m_1+m_2+\cdots+m_{(\mathcal{M}-1)}} \\ \psi_{2t} \in \Re^{m_{\mathcal{M}}} \end{matrix}$$

$\forall t \in \{1, 2, \cdots, p\}$ and $p = \min(m_1, m_2, \cdots, m_{\mathcal{M}})$.

$$\text{Hence,} \qquad \widehat{\mathcal{B}}_{(\mathcal{M}-1)}\Psi = \Psi\Lambda; \tag{5.8}$$

where Λ is a diagonal matrix with diagonal elements $\lambda_1, \lambda_2, \cdots, \lambda_p$. So,

$$\widehat{\mathcal{B}}_{(\mathcal{M}-1)}\mathit{\Psi}_t = \lambda_t\mathit{\Psi}_t; \Rightarrow \begin{bmatrix} \mathbf{0}^{[\sum_{i=1}^{\mathcal{M}-1} m_i]} & \Theta_{(\mathcal{M}-1)} \\ \Phi_{(\mathcal{M}-1)} & \mathbf{0}^{[m_{\mathcal{M}}]} \end{bmatrix} \begin{bmatrix} \psi_{1t} \\ \psi_{2t} \end{bmatrix} = \lambda_t \begin{bmatrix} \psi_{1t} \\ \psi_{2t} \end{bmatrix}. \tag{5.9}$$

From (5.9), we get

$$\Theta_{(\mathcal{M}-1)}\psi_{2t} = \lambda_t\psi_{1t} \Rightarrow \psi_{1t} = \frac{1}{\lambda_t}\Theta_{(\mathcal{M}-1)}\psi_{2t}; \tag{5.10}$$

$$\text{and} \qquad \Phi_{(\mathcal{M}-1)}\psi_{1t} = \lambda_t \psi_{2t} \Rightarrow \psi_{2t} = \frac{1}{\lambda_t}\Phi_{(\mathcal{M}-1)}\psi_{1t}. \tag{5.11}$$

Using (5.10) and (5.11), we get

$$\Theta_{(\mathcal{M}-1)}\Phi_{(\mathcal{M}-1)}\psi_{1t} = \lambda_t^2 \psi_{1t}; \tag{5.12}$$

$$\text{and} \qquad \Phi_{(\mathcal{M}-1)}\Theta_{(\mathcal{M}-1)}\psi_{2t} = \lambda_t^2 \psi_{2t}. \tag{5.13}$$

Combining (5.12) and (5.13), we get

$$\left[\begin{array}{cc} \Theta_{(\mathcal{M}-1)}\Phi_{(\mathcal{M}-1)} & \mathbf{0}^{[\sum_{i=1}^{\mathcal{M}-1} m_i, m_{\mathcal{M}}]} \\ \mathbf{0}^{[m_{\mathcal{M}}, \sum_{i=1}^{\mathcal{M}-1} m_i]} & \Phi_{(\mathcal{M}-1)}\Theta_{(\mathcal{M}-1)} \end{array}\right] \left[\begin{array}{c} \psi_{1t} \\ \psi_{2t} \end{array}\right] = \lambda_t^2 \left[\begin{array}{c} \psi_{1t} \\ \psi_{2t} \end{array}\right]$$

$$\Rightarrow \widetilde{\mathcal{B}}_{(\mathcal{M}-1)}\Psi_t = \lambda_t^2 \Psi_t;$$

$$\Rightarrow \widetilde{\mathcal{B}}_{(\mathcal{M}-1)}\Psi = \Psi\Lambda^2; \tag{5.14}$$

$$\text{where} \qquad \widetilde{\mathcal{B}}_{(\mathcal{M}-1)} = \left[\begin{array}{cc} \Theta_{(\mathcal{M}-1)}\Phi_{(\mathcal{M}-1)} & \mathbf{0}^{[\sum_{i=1}^{\mathcal{M}-1} m_i, m_{\mathcal{M}}]} \\ \mathbf{0}^{[m_{\mathcal{M}}, \sum_{i=1}^{\mathcal{M}-1} m_i]} & \Phi_{(\mathcal{M}-1)}\Theta_{(\mathcal{M}-1)} \end{array}\right];$$

and Λ^2 is a diagonal matrix with diagonal elements $\lambda_1^2, \lambda_2^2, \cdots, \lambda_p^2$. Hence, (5.8) and (5.3.1) lead to the conclusions that both the matrices $\widehat{\mathcal{B}}_{(\mathcal{M}-1)}$ and $\widetilde{\mathcal{B}}_{(\mathcal{M}-1)}$ have same eigenvectors $\Psi = [\Psi_1, \Psi_2, \cdots, \Psi_p]$. However, the eigenvalues of the matrix $\widetilde{\mathcal{B}}_{(\mathcal{M}-1)}$ are the square of the corresponding eigenvalues of the matrix $\widehat{\mathcal{B}}_{(\mathcal{M}-1)}$. So,

$$\widehat{\mathcal{B}}_{(\mathcal{M}-1)} = \Psi\Lambda\Psi^T \quad \text{and} \quad \widetilde{\mathcal{B}}_{(\mathcal{M}-1)} = \Psi\Lambda^2\Psi^T. \tag{5.15}$$

Now, combining (5.8) and (5.15), we get

$$\widetilde{\mathcal{B}}_{(\mathcal{M}-1)} = \Psi\Lambda\Lambda\Psi^T = \widehat{\mathcal{B}}_{(\mathcal{M}-1)}\Psi\Lambda\Psi^T = \widehat{\mathcal{B}}_{(\mathcal{M}-1)}\widehat{\mathcal{B}}_{(\mathcal{M}-1)}$$

$$\Rightarrow \widehat{\mathcal{B}}_{(\mathcal{M}-1)} = \widetilde{\mathcal{B}}_{(\mathcal{M}-1)}^{1/2}. \tag{5.16}$$

Using (5.3.1), the matrix $\mathcal{A}_{\mathcal{M}}$ of (5.3.1) becomes

$$\mathcal{A}_{\mathcal{M}} = \left[\begin{array}{cc} \mathcal{A}_{(\mathcal{M}-1)} & \mathbf{0}^{[\sum_{i=1}^{\mathcal{M}-1} m_i, m_{\mathcal{M}}]} \\ \mathbf{0}^{[m_{\mathcal{M}}, \sum_{i=1}^{\mathcal{M}-1} m_i]} & \mathbf{0}^{[m_{\mathcal{M}}]} \end{array}\right] + \left[\begin{array}{cc} [\Theta_{(\mathcal{M}-1)}\Phi_{(\mathcal{M}-1)}]^{1/2} & \mathbf{0}^{[\sum_{i=1}^{\mathcal{M}-1} m_i, m_{\mathcal{M}}]} \\ \mathbf{0}^{[m_{\mathcal{M}}, \sum_{i=1}^{\mathcal{M}-1} m_i]} & [\Phi_{(\mathcal{M}-1)}\Theta_{(\mathcal{M}-1)}]^{1/2} \end{array}\right]$$

$$= \begin{bmatrix} \mathcal{A}_{(\mathcal{M}-1)} + [\Theta_{(\mathcal{M}-1)}\Phi_{(\mathcal{M}-1)}]^{1/2} & \mathbf{0}^{[\sum_{i=1}^{\mathcal{M}-1} m_i, m_{\mathcal{M}}]} \\ \mathbf{0}^{[m_{\mathcal{M}}, \sum_{i=1}^{\mathcal{M}-1} m_i]} & [\Phi_{(\mathcal{M}-1)}\Theta_{(\mathcal{M}-1)}]^{1/2} \end{bmatrix}. \tag{5.17}$$

According to (5.4), the eigenvectors of $\mathcal{A}_{\mathcal{M}}$ are the basis vectors $w_1, w_2, \cdots, w_{\mathcal{M}}$ of $\mathcal{M}$ multidimensional variables, that is,

$$\begin{bmatrix} \mathcal{A}_{(\mathcal{M}-1)} + [\Theta_{(\mathcal{M}-1)}\Phi_{(\mathcal{M}-1)}]^{1/2} & \mathbf{0}^{[\sum_{i=1}^{\mathcal{M}-1} m_i, m_{\mathcal{M}}]} \\ \mathbf{0}^{[m_{\mathcal{M}}, \sum_{i=1}^{\mathcal{M}-1} m_i]} & [\Phi_{(\mathcal{M}-1)}\Theta_{(\mathcal{M}-1)}]^{1/2} \end{bmatrix} \begin{bmatrix} \mathcal{W}_{(\mathcal{M}-1)} \\ w_{\mathcal{M}} \end{bmatrix} = \rho \begin{bmatrix} \mathcal{W}_{(\mathcal{M}-1)} \\ w_{\mathcal{M}} \end{bmatrix}; \tag{5.18}$$

$$\text{where} \qquad \mathcal{W}_{(\mathcal{M}-1)} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_{(\mathcal{M}-1)} \end{bmatrix}. \tag{5.19}$$

From (5.18), it is evident that the eigenvectors of the matrices $[\mathcal{A}_{(\mathcal{M}-1)} + [\Theta_{(\mathcal{M}-1)}\Phi_{(\mathcal{M}-1)}]^{1/2}]$ and $[\Phi_{(\mathcal{M}-1)}\Theta_{(\mathcal{M}-1)}]^{1/2}$ are $\mathcal{W}_{(\mathcal{M}-1)}$ and $w_{\mathcal{M}}$, respectively, with corresponding eigenvalue ρ, that is,

$$\left[\mathcal{A}_{(\mathcal{M}-1)} + [\Theta_{(\mathcal{M}-1)}\Phi_{(\mathcal{M}-1)}]^{1/2}\right] \mathcal{W}_{(\mathcal{M}-1)} = \rho \mathcal{W}_{(\mathcal{M}-1)}; \tag{5.20}$$

$$\text{and} \qquad [\Phi_{(\mathcal{M}-1)}\Theta_{(\mathcal{M}-1)}]^{1/2} w_{\mathcal{M}} = \rho w_{\mathcal{M}}. \tag{5.21}$$

Using (5.21), it can be proved that the basis vector of the $\mathcal{M}$-th multidimensional variable is the eigenvector of the matrix $[\Phi_{(\mathcal{M}-1)}\Theta_{(\mathcal{M}-1)}]$ with corresponding eigenvalue ρ^2, as

$$[\Phi_{(\mathcal{M}-1)}\Theta_{(\mathcal{M}-1)}] w_{\mathcal{M}} = [\Phi_{(\mathcal{M}-1)}\Theta_{(\mathcal{M}-1)}]^{1/2} \rho w_{\mathcal{M}} = \rho^2 w_{\mathcal{M}}. \tag{5.22}$$

Now,

$$\Phi_{(\mathcal{M}-1)}\Theta_{(\mathcal{M}-1)} = \sum_{j=1}^{\mathcal{M}-1} a_{\mathcal{M}j} a_{j\mathcal{M}} = \sum_{j=1}^{\mathcal{M}-1} C_{\mathcal{M}\mathcal{M}}{}^{-1} C_{\mathcal{M}j} C_{jj}{}^{-1} C_{j\mathcal{M}}; \tag{5.23}$$

and $\mathcal{M} \geqslant 2$. Hence, the basis vector of the i-th multidimensional variable is the eigenvector of the following matrix

$$\mathcal{H}_i = \sum_{\substack{j=1 \\ j \neq i}}^{\mathcal{M}} C_{ii}{}^{-1} C_{ij} C_{jj}{}^{-1} C_{ji}; \qquad \forall i \in \{1, 2, \cdots, \mathcal{M}\}. \tag{5.24}$$

Using (5.24), it is now possible to find a basis vector by maximizing the correlations among all the modalities. This is not the case for the conventional implementation of the MCCA as mentioned in (5.4), where the pairwise correlations are considered to find the basis vectors. The relation (5.24) also implies that the basis vector of the i-th multidimensional variable is independent of the basis vectors of other canonical variables. It makes the proposed algorithm scalable. Unlike (5.4), as the relation (5.24) is not iterative in nature, there is no need to calculate the eigenvectors of the matrix repeatedly to compute each basis vector. Without loss of generality, if we assume that $m_{\mathcal{M}} \geqslant m_{\mathcal{M}} - 1 \geqslant \cdots \geqslant m_1$, the computational complexity to calculate the matrix $\mathcal{H}_i$ becomes $\mathcal{O}(m_{\mathcal{M}}^2 n + m_{\mathcal{M}}^3 + m_{\mathcal{M}-1}^2 n + m_{\mathcal{M}-1}^3 + m_{\mathcal{M}} m_{\mathcal{M}-1} n + m_{\mathcal{M}}^2 m_{\mathcal{M}-1} + m_{\mathcal{M}} m_{\mathcal{M}-1}^2) \approx \mathcal{O}(m_{\mathcal{M}}^2 (m_{\mathcal{M}} + m_{\mathcal{M}-1}))$. On the other hand, the time complexity to calculate each basis vector is $\mathcal{O}(m_{\mathcal{M}}^3 + m_{\mathcal{M}}^2 (m_{\mathcal{M}} + m_{\mathcal{M}-1})) \approx \mathcal{O}(m_{\mathcal{M}}^3)$. Hence, all $\mathcal{M}$ basis vectors can be computed with complexity $\mathcal{O}(\mathcal{M} m_{\mathcal{M}}^3)$, which is lesser than that of both Horst-Jacobi and Gauss-Seidel algorithms.

5.3.2 Multiset Ridge Regression Model

From (5.24), it is seen that the inverse of the covariance matrix $\mathcal{C}_{ii}$ is needed to compute the basis vector $w_i, \forall i \in \{1, 2, \cdots, \mathcal{M}\}$. If $n \ll m_i$, the covariance matrix $\mathcal{C}_{ii}$ becomes non-invertible. The singularity problem of the covariance matrix may also arise due to the presence of noise in multimodal data [68]. In effect, it leads to the invalid computation of the canonical variables. To overcome this problem, either the values of off-diagonal elements of each covariance matrix could be reduced by a shrinkage parameter [262], or the diagonal elements could be increased by adding a regularization parameter [278]. In the proposed algorithm, a ridge regression optimization scheme is used by adding a small positive quantity $\mathfrak{r}_i$, known as regularization parameter, to the diagonals of the covariance matrix $\mathcal{C}_{ii}$. It facilitates increased search space of finding the appropriate canonical variables, which may help to improve the performance of multimodal data analysis.

Let us assume that the ℓ-th dimension of the i-th multidimensional variable $\mathcal{X}_i[\ell]$ is contaminated with noise $\varepsilon_i[\ell]$, $\forall \ell \in \{1, 2, \cdots, m_i\}$ and $\forall i \in \{1, 2, \cdots, \mathcal{M}\}$, such that $\mathcal{E}[\varepsilon_i[\ell]] = 0$, $\mathcal{E}[\varepsilon_i[\ell]\varepsilon_i[k]^T] = 0$ for $\ell \neq k$, $\mathcal{E}[\varepsilon_i[\ell]\mathcal{X}_i[\ell]^T] = 0$ and $\mathcal{E}[\varepsilon_i[\ell]\varepsilon_i[\ell]^T] = \mathfrak{r}_i \geqslant 0$. Under these assumptions, the cross-covariance matrix of $\mathcal{X}_i$ and $\mathcal{X}_j$ is $\mathcal{C}_{ij}$, while the covariance matrix of $\mathcal{X}_i$ becomes $[\mathcal{C}_{ii} + \mathfrak{r}_i I]$. To estimate the basis vector w_i, the covariance matrix $\mathcal{C}_{ii}$ needs to be replaced by $[\mathcal{C}_{ii} + \mathfrak{r}_i I]$. This modification is similar to the ridge regression modification [278]. The optimal set of regularization parameters can be estimated in such a way that the correlation between multiset canonical variables becomes maximum. To estimate the optimal set of regularization parameters, a grid search optimization is performed, where each regularization parameter $\mathfrak{r}_i$ follows an arithmetic progression and is varied within a specified range $[\mathfrak{r}_{min}, \mathfrak{r}_{max}]$. Let d_i be the common difference for regularization parameter $\mathfrak{r}_i$, while the parameter $\mathfrak{t}_i$ indicates the number of possible values of $\mathfrak{r}_i$. Hence, to compute the matrix $\mathcal{H}_i$ of (5.24), the inverse of the covariance matrix of each multidimensional variable has to be computed $\mathfrak{t}_i$ times. According to [174], as the diagonal elements of $\mathcal{C}_{ii}$ are only changed by adding $\mathfrak{r}_i$, the eigenvalues of $[\mathcal{C}_{ii} + (\mathfrak{r}_i + k_i d_i) I]$ are changed, but the corresponding eigenvectors remain same $\forall k_i \in \{0, 1, \cdots, (\mathfrak{t}_i - 1)\}$. Also, there exists a relation between the eigenvalues of $[\mathcal{C}_{ii} + \mathfrak{r}_i I]$ and that of $[\mathcal{C}_{ii} + (\mathfrak{r}_i + k_i d_i) I]$, which is given by [248]

$$\Delta_{i k_i} = \Delta_i + k_i d_i I; \tag{5.25}$$

where $\Delta_{i k_i}$ is the diagonal matrix, whose diagonal elements are the eigenvalues of $[\mathcal{C}_{ii} + (\mathfrak{r}_i +$

$\mathscr{k}_i\mathscr{d}_i)\mathcal{I}]$, $\Delta_i = \Delta_{i0}$ and $\mathcal{I}$ is the identity matrix of appropriate order. Let the corresponding eigenvectors of the matrix $[\mathcal{C}_{ii} + (\mathfrak{r}_i + \mathscr{k}_i\mathscr{d}_i)\mathcal{I}]$ be the columns of Ω_i. Based on spectral decomposition, the covariance matrix $[\mathcal{C}_{ii} + (\mathfrak{r}_i + \mathscr{k}_i\mathscr{d}_i)\mathcal{I}]$ can be expressed as follows [269]:

$$[\mathcal{C}_{ii} + (\mathfrak{r}_i + \mathscr{k}_i\mathscr{d}_i)\mathcal{I}] = \Omega_i\Delta_{i\mathscr{k}_i}\Omega_i^T = \Omega_i[\Delta_i + \mathscr{k}_i\mathscr{d}_i\mathcal{I}]\Omega_i^T = \sum_{\ell=1}^{m_i}(\delta_{i\ell} + \mathscr{k}_i\mathscr{d}_i)\omega_{i\ell}\omega_{i\ell}^T; \tag{5.26}$$

and the inverse covariance matrix $[\mathcal{C}_{ii} + (\mathfrak{r}_i + \mathscr{k}_i\mathscr{d}_i)\mathcal{I}]^{-1}$ can be computed as follows:

$$[\mathcal{C}_{ii} + (\mathfrak{r}_i + \mathscr{k}_i\mathscr{d}_i)\mathcal{I}]^{-1} = \sum_{\ell=1}^{m_i}\frac{1}{(\delta_{i\ell} + \mathscr{k}_i\mathscr{d}_i)}\omega_{i\ell}\omega_{i\ell}^T; \tag{5.27}$$

where the ℓ-th element $\delta_{i\ell}$ of the diagonal matrix Δ_i denotes the ℓ-th eigenvalue of the matrix $[\mathcal{C}_{ii} + \mathfrak{r}_i\mathcal{I}]$. The ℓ-th column of the matrix Ω_i represents the orthogonalized eigenvector $\omega_{i\ell}$ corresponding to the eigenvalue $\delta_{i\ell}$, $\forall \ell \in \{1, 2, \cdots, m_i\}$. From (5.27), it can be observed that there is no need to compute the eigenvalue for every regularization parameter $\mathfrak{r}_i$ of each multidimensional variable $\mathcal{X}_i$. It is sufficient to calculate the eigenvalues $\delta_{i\ell}$ and eigenvectors $\omega_{i\ell}$ of the covariance matrix corresponding to the initial value of $\mathfrak{r}_i$. Moreover, as the regularization parameters follow an arithmetic progression, the ℓ-th element of each diagonal matrix $[\Delta_i + \mathscr{k}_i\mathscr{d}_i\mathcal{I}]$ is in arithmetic progression. Hence, the ℓ-th element of each diagonal matrix $[\Delta_i + \mathscr{k}_i\mathscr{d}_i\mathcal{I}]^{-1}$ follows harmonic progression, that is, the ℓ-th element of all diagonal matrices $[\Delta_i + \mathscr{k}_i\mathscr{d}_i\mathcal{I}]^{-1}$ be $\frac{1}{\delta_{i\ell}}, \frac{1}{\delta_{i\ell}+\mathscr{d}_i}, \frac{1}{\delta_{i\ell}+2\mathscr{d}_i}, \cdots, \frac{1}{\delta_{i\ell}+(\mathfrak{t}_i-1)\mathscr{d}_i}$. Now,

$$\frac{1}{\delta_{i\ell} + \mathscr{k}_i\mathscr{d}_i} = \frac{1}{\delta_{i\ell}} - \sum_{j=1}^{\mathscr{k}_i}\frac{\mathscr{d}_i}{(\delta_{i\ell} + (j-1)\mathscr{d}_i)(\delta_{i\ell} + j\mathscr{d}_i)}. \tag{5.28}$$

Let us assume that $\mathcal{V}_{i\mathscr{k}_i} \in \Re^{m_i}$ be a row vector, where

$$\mathcal{V}_{i\mathscr{k}_i} = \begin{bmatrix} \frac{\mathscr{d}_i}{(\delta_{i1}+(\mathscr{k}_i-1)\mathscr{d}_i)(\delta_{i1}+\mathscr{k}_i\mathscr{d}_i)} \\ \frac{\mathscr{d}_i}{(\delta_{i2}+(\mathscr{k}_i-1)\mathscr{d}_i)(\delta_{i2}+\mathscr{k}_i\mathscr{d}_i)} \\ \vdots \\ \frac{\mathscr{d}_i}{(\delta_{im_i}+(\mathscr{k}_i-1)\mathscr{d}_i)(\delta_{im_i}+\mathscr{k}_i\mathscr{d}_i)} \end{bmatrix}^T;$$

$\forall \mathscr{k}_i \in \{1, 2, \cdots, (\mathfrak{t}_i - 1)\}$. Let $\widehat{\mathcal{C}_\ell}$ and $\widehat{\mathcal{D}_\ell}$ be a column vector and a square matrix of dimension m_i, respectively, where

$$\widehat{\mathcal{C}_\ell}[j] = \begin{cases} 1 & \text{if} \quad \ell = j, \\ 0 & \text{otherwise;} \end{cases}$$

$$\text{and} \quad \widehat{\mathcal{D}_\ell}[j, t] = \begin{cases} 1 & \text{if} \quad \ell = j = t, \\ 0 & \text{otherwise;} \end{cases}$$

$\forall \ell, j, t \in \{1, 2, \cdots, m_i\}$. Hence, the diagonal matrix $[\Delta_i + k_i d_i I]^{-1}$ can be expressed as

$$[\Delta_i + k_i d_i I]^{-1} = \Delta_i^{-1} - \sum_{\ell=1}^{m_i} \sum_{j=1}^{k_i} \widehat{\mathcal{C}_\ell} \mathcal{V}_{ij} \widehat{\mathcal{D}_\ell}. \tag{5.29}$$

Using (5.29), the inverse covariance matrix of (5.27) can be expressed as

$$\begin{aligned}
[\mathcal{C}_{ii} + (\mathfrak{r}_i + k_i d_i) I]^{-1} &= \Omega_i \left[\Delta_i^{-1} - \sum_{\ell=1}^{m_i} \sum_{j=1}^{k_i} \widehat{\mathcal{C}_\ell} \mathcal{V}_{ij} \widehat{\mathcal{D}_\ell} \right] \Omega_i^T \\
&= \Omega_i \left[\Delta_i^{-1} - \sum_{\ell=1}^{m_i} \sum_{j=1}^{k_i-1} \widehat{\mathcal{C}_\ell} \mathcal{V}_{ij} \widehat{\mathcal{D}_\ell} \right] \Omega_i^T - \Omega_i \left[\sum_{\ell=1}^{m_i} \widehat{\mathcal{C}_\ell} \mathcal{V}_{i k_i} \widehat{\mathcal{D}_\ell} \right] \Omega_i^T \\
&= [\mathcal{C}_{ii} + (\mathfrak{r}_i + (k_i - 1) d_i) I]^{-1} - \Upsilon_{i k_i} \\
&= [\mathcal{C}_{ii} + \mathfrak{r}_i I]^{-1} - \sum_{s=1}^{k_i} \Upsilon_{is} \\
&= \Omega_i \Delta_i^{-1} \Omega_i^T - \sum_{s=1}^{k_i} \Upsilon_{is} \qquad (5.30) \\
\text{where} \quad \Upsilon_{i k_i} &= \Omega_i \left[\sum_{\ell=1}^{m_i} \widehat{\mathcal{C}_\ell} \mathcal{V}_{i k_i} \widehat{\mathcal{D}_\ell} \right] \Omega_i^T.
\end{aligned}$$

From (5.3.2), it is observed that the covariance matrix of each multidimensional variable $\mathcal{X}_i$, corresponding to every regularization parameter $\mathfrak{r}_i$, can be computed from the covariance matrix corresponding to initial value of $\mathfrak{r}_i$. Hence, the matrix $\mathcal{H}_i$ of (5.24) can be expressed as

$$\begin{aligned}
\mathcal{H}_{ir} &= \sum_{\substack{j=1 \\ j \neq i}}^{\mathcal{M}} [\mathcal{C}_{ii} + (\mathfrak{r}_i + k_i d_i) I]^{-1} \mathcal{C}_{ij} [\mathcal{C}_{jj} + (\mathfrak{r}_j + k_j d_j) I]^{-1} \mathcal{C}_{ji} \\
&= \sum_{\substack{j=1 \\ j \neq i}}^{\mathcal{M}} ([\mathcal{C}_{ii} + (\mathfrak{r}_i + (k_i - 1) d_i) I]^{-1} - \Upsilon_{i k_i}) \mathcal{C}_{ij} ([\mathcal{C}_{jj} + (\mathfrak{r}_j + (k_j - 1) d_j) I]^{-1} - \Upsilon_{j k_j}) \mathcal{C}_{ji} \\
&= \mathcal{H}_{i(r-1)} - \tilde{\mathcal{G}}_{ir} - \hat{\mathcal{G}}_{ir} + \bar{\mathcal{G}}_{ir} \\
&= \mathcal{H}_{i1} + \sum_{s=1}^{r} \bar{\mathcal{G}}_{is} - \tilde{\mathcal{G}}_{is} - \hat{\mathcal{G}}_{is}
\end{aligned}$$

$$= \sum_{\substack{j=1\\j\neq i}}^{\mathcal{M}} \Omega_i \Delta_i^{-1} \Omega_i^T \mathcal{C}_{ij} \Omega_j \Delta_j^{-1} \Omega_j^T \mathcal{C}_{ji} + \sum_{s=1}^{r} \bar{\mathcal{G}}_{i_s} - \tilde{\mathcal{G}}_{i_s} - \hat{\mathcal{G}}_{i_s} \tag{5.31}$$

where $$\tilde{\mathcal{G}}_{i_r} = \Upsilon_{i_{\mathcal{k}_i}} \sum_{\substack{j=1\\j\neq i}}^{\mathcal{M}} \mathcal{C}_{ij} [\mathcal{C}_{jj} + (\mathfrak{r}_j + (\mathcal{k}_j - 1)\mathcal{d}_j) I]^{-1} \mathcal{C}_{ji} = \Upsilon_{i_{\mathcal{k}_i}} \sum_{\substack{j=1\\j\neq i}}^{\mathcal{M}} \mathcal{C}_{ij} \Omega_j [\Delta_j + (\mathcal{k}_j - 1)\mathcal{d}_j I]^{-1} \Omega_j^T \mathcal{C}_{ji}; \tag{5.32}$$

$$\hat{\mathcal{G}}_{i_r} = [\mathcal{C}_{ii} + (\mathfrak{r}_i + (\mathcal{k}_i - 1)\mathcal{d}_i) I]^{-1} \sum_{\substack{j=1\\j\neq i}}^{\mathcal{M}} \mathcal{C}_{ij} \Upsilon_{j_{\mathcal{k}_j}} \mathcal{C}_{ji} = \Omega_i [\Delta_i + (\mathcal{k}_i - 1)\mathcal{d}_i I]^{-1} \Omega_i^T \sum_{\substack{j=1\\j\neq i}}^{\mathcal{M}} \mathcal{C}_{ij} \Upsilon_{j_{\mathcal{k}_j}} \mathcal{C}_{ji}; \tag{5.33}$$

$$\text{and} \quad \bar{\mathcal{G}}_{i_r} = \Upsilon_{i_{\mathcal{k}_i}} \sum_{\substack{j=1\\j\neq i}}^{\mathcal{M}} \mathcal{C}_{ij} \Upsilon_{j_{\mathcal{k}_j}} \mathcal{C}_{ji}; \tag{5.34}$$

$\forall \mathcal{k}_i \in \{1, 2, \cdots, (\mathfrak{t}_i - 1)\}$ and $\forall r \in \{2, 3, \cdots, T\}$, where $\mathcal{T} = \prod_{\ell=1}^{\mathcal{M}} \mathfrak{t}_\ell$ represents the number of all possible combinations of regularization parameters. From (5.3.2), it is clear that if the eigenvalues and eigenvectors of $[\mathcal{C}_{ii} + \mathfrak{r}_i I]$ are calculated to compute $\mathcal{H}_{i1}$ for the initial value of $\mathfrak{r}_i$, there is no need to compute the eigenvalues and eigenvectors at other values of $\mathfrak{r}_i$ for computing $\mathcal{H}_{ir}$, as the eigenvalues and eigenvectors corresponding to the initial regularization parameter can be used to compute different $\mathcal{H}_{ir}$. Also, if the minimum value $\mathfrak{r}_{min}$ of $\mathfrak{r}_i$ is set to 0, then the eigenvalues and eigenvectors of the MCCA can be used to compute the $\mathcal{H}_{ir}$.

According to (5.22), (5.23) and (5.24), the basis vector w_{ir} is the eigenvector of the matrix $\mathcal{H}_{ir}$, $\forall i \in \{1, 2, \cdots, \mathcal{M}\}$ and $\forall r \in \{1, 2, \cdots, \mathcal{T}\}$. So, the eigenvectors of the matrices $\mathcal{H}_{ir}$ have to be computed for each regularization combination of each multidimensional variable. Let say, $\mathcal{H}_{ir}$ has t-th eigenvalue $\rho_r^2(t)$ and corresponding eigenvector is $w_{ir}(t)$, $\forall t \in \{1, 2, \cdots, p\}$, where $p = \min(m_1, m_2, \cdots, m_{\mathcal{M}})$. Hence,

$$\mathcal{H}_{ir} w_{ir}(t) = \rho_r^2(t) w_{ir}(t). \tag{5.35}$$

The cross-covariance matrix of $\mathcal{X}_i$ and $\mathcal{X}_j$ is $\mathcal{C}_{ij} = \mathcal{X}_i {\mathcal{X}_j}^T$; $\forall i, j \in \{1, 2, \cdots, \mathcal{M}\}$. Without loss of generality, it is assumed that $n \ll m_i$ and the matrix $\mathcal{X}_i$ is a full rank matrix, which implies that $\mathcal{X}_i$ has linearly independent columns. Thus, the pseudoinverse of $\mathcal{X}_i$ is ${\mathcal{X}_i}^\dagger = \left({\mathcal{X}_i}^T \mathcal{X}_i\right)^{-1} {\mathcal{X}_i}^T$ and ${\mathcal{X}_i}^\dagger \mathcal{X}_i = {\mathcal{X}_i}^T ({\mathcal{X}_i}^T)^\dagger = I$. Now the inverse of the covariance matrix of $\mathcal{X}_i$ is given by ${\mathcal{C}_{ii}}^{-1} = \left(\mathcal{X}_i {\mathcal{X}_i}^T\right)^{-1} = \left({\mathcal{X}_i}^T\right)^\dagger {\mathcal{X}_i}^\dagger$. Hence, the matrix $\mathcal{C}_{(i+1)(i+1)}^{-1} \mathcal{C}_{(i+1)i} \mathcal{C}_{ii}^{-1} \mathcal{C}_{ij} \mathcal{C}_{jj}^{-1} \mathcal{C}_{ji}$ can be written as

$$\mathcal{C}_{(i+1)(i+1)}^{-1} \mathcal{C}_{(i+1)i} \mathcal{C}_{ii}^{-1} \mathcal{C}_{ij} \mathcal{C}_{jj}^{-1} \mathcal{C}_{ji} = \mathcal{C}_{(i+1)(i+1)}^{-1} \left[\mathcal{X}_{i+1} {\mathcal{X}_i}^T\right] \left[({\mathcal{X}_i}^T)^\dagger {\mathcal{X}_i}^\dagger\right] \left[\mathcal{X}_i {\mathcal{X}_j}^T\right] \left[({\mathcal{X}_j}^T)^\dagger {\mathcal{X}_j}^\dagger\right] \left[\mathcal{X}_j {\mathcal{X}_i}^T\right]$$

$$= {\mathcal{C}_{(i+1)(i+1)}}^{-1} \mathcal{X}_{i+1} \mathcal{X}_i^T$$

$$= \mathcal{C}_{(i+1)(i+1)}{}^{-1}\mathcal{X}_{i+1}\big[\mathcal{X}_j{}^T(\mathcal{X}_j{}^T)^\dagger\big]\big[\mathcal{X}_j^\dagger\mathcal{X}_j\big]\big[\mathcal{X}_{i+1}^T(\mathcal{X}_{i+1}^T)^\dagger\big]\big[\mathcal{X}_{i+1}{}^\dagger\mathcal{X}_{i+1}\big]\mathcal{X}_i{}^T$$

$$= \mathcal{C}_{(i+1)(i+1)}^{-1}\big[\mathcal{X}_{i+1}\mathcal{X}_j{}^T\big]\big[(\mathcal{X}_j{}^T)^\dagger\mathcal{X}_j{}^\dagger\big]\big[\mathcal{X}_j\mathcal{X}_{i+1}^T\big]\big[(\mathcal{X}_{i+1}^T)^\dagger\mathcal{X}_{i+1}^\dagger\big]\big[\mathcal{X}_{i+1}\mathcal{X}_i{}^T\big]$$

$$= \mathcal{C}_{(i+1)(i+1)}^{-1}\mathcal{C}_{(i+1)j}\mathcal{C}_{jj}{}^{-1}\mathcal{C}_{j(i+1)}\mathcal{C}_{(i+1)(i+1)}^{-1}\mathcal{C}_{(i+1)i}; \tag{5.36}$$

$\forall i \in \{1, 2, \cdots, (\mathcal{M}-1)\}$ and $\forall j \in \{1, 2, \cdots, \mathcal{M}\}$ and $i \neq j$. Hence, using (5.35) and (5.3.2), we get

$$\sum_{\substack{j=1\\ j\neq i}}^{\mathcal{M}}[\mathcal{C}_{(i+1)(i+1)}+(\mathfrak{r}_{(i+1)}+k_{(i+1)}d_{(i+1)})I]^{-1}\mathcal{C}_{(i+1)i}[\mathcal{C}_{ii}+(\mathfrak{r}_i+k_i d_i)I]^{-1}\mathcal{C}_{ij}[\mathcal{C}_{jj}+(\mathfrak{r}_j+k_j d_j)I]^{-1}\mathcal{C}_{ji}w_{ir}(t)$$

$$= \rho_r^2(t)[\mathcal{C}_{(i+1)(i+1)} + (\mathfrak{r}_{(i+1)} + k_{(i+1)}d_{(i+1)})I]^{-1}\mathcal{C}_{(i+1)i}w_{ir}(t)$$

$$\Rightarrow \sum_{\substack{j=1\\ j\neq i}}^{\mathcal{M}}[\mathcal{C}_{(i+1)(i+1)} + (\mathfrak{r}_{(i+1)} + k_{(i+1)}d_{(i+1)})I]^{-1}\mathcal{C}_{(i+1)j}[\mathcal{C}_{jj} + (\mathfrak{r}_j + k_j d_j)I]^{-1}\mathcal{C}_{j(i+1)}[\mathcal{C}_{(i+1)(i+1)}$$

$$+(\mathfrak{r}_{(i+1)} + k_{(i+1)}d_{(i+1)})I]^{-1}\mathcal{C}_{(i+1)i}w_{ir}(t)$$

$$= \rho_r^2(t)[\mathcal{C}_{(i+1)(i+1)} + (\mathfrak{r}_{(i+1)} + k_{(i+1)}d_{(i+1)})I]^{-1}\mathcal{C}_{(i+1)i}w_{ir}(t)$$

$$\Rightarrow \mathcal{H}_{(i+1)_r}w_{(i+1)_r}(t) = \rho_r^2(t)w_{(i+1)_r}(t); \tag{5.37}$$

$$\text{where} \quad w_{(i+1)_r}(t) = [\mathcal{C}_{(i+1)(i+1)} + (\mathfrak{r}_{(i+1)} + k_{(i+1)}d_{(i+1)})I]^{-1}\mathcal{C}_{(i+1)i}w_{ir}(t); \tag{5.38}$$

$$\Rightarrow w_{(i+1)_r}(t) = \prod_{j=0}^{i-1}[\mathcal{C}_{(i+1-j)(i+1-j)} + (\mathfrak{r}_{(i+1-j)} + k_{(i+1-j)}d_{(i+1-j)})I]^{-1}\mathcal{C}_{(i+1-j)(i-j)}w_{1r}(t); \tag{5.39}$$

$\forall i \in \{1, 2, \cdots, (\mathcal{M}-1)\}$. So, the t-th eigenvector $w_{(i+1)_r}(t)$ of $\mathcal{H}_{(i+1)_r}$ is proportional to $[\mathcal{C}_{(i+1)(i+1)} + (\mathfrak{r}_{(i+1)} + k_{(i+1)}d_{(i+1)})I]^{-1}\mathcal{C}_{(i+1)i}$ and can be derived from the t-th eigenvector $w_{ir}(t)$ of $\mathcal{H}_{ir}$ using (5.38). From (5.39), it is evident that $\mathcal{H}_{1r}$ matrix is enough to calculate the eigenvectors of all $\mathcal{H}_{ir}$ matrices, $\forall i \in \{2, 3, \cdots, \mathcal{M}\}$ and $\forall r \in \{1, 2, \cdots, \mathcal{T}\}$. On the other hand, according to (5.3.2), $\mathcal{H}_{1r}$ matrix can be computed using $\mathcal{H}_{11}$ matrix. Hence, $\mathcal{H}_{11}$ matrix is enough to compute the basis vectors of all the modalities corresponding to all possible combinations of the regularization parameters.

5.3.3 Sequential Generation of Canonical Variables

For each $\mathcal{H}_{ir}$, p eigenvectors can be computed simultaneously using the Jacobi method [248], with a computational complexity $\mathcal{O}(p^3)$. These eigenvectors are the basis vectors of $\mathcal{H}_{ir}$.

The corresponding canonical variables can be computed from these basis vectors. Finally, p features can be extracted simultaneously for each combination of regularization parameters. One of the major goals in data science is how to extract a compact set of most relevant features. This is an important problem in machine learning and termed as feature selection. Instead of producing all p eigenvectors simultaneously using the Jacobi method, if each eigenvector of $\mathcal{H}_{ir}$ matrix is computed sequentially, the quality of each generated feature can be evaluated independently, and eventually, $\mathcal{D}$ features can be selected based on their quality, where $\mathcal{D} \leqslant p$. Moreover, for real-world multimodal high-dimensional data analysis, the value of p is large, while a small fraction $\mathcal{D} < p$ is typically enough to deal with a problem.

In order to address the above problems, the proposed algorithm uses the Power method [248] to compute the eigenvectors of $\mathcal{H}_{ir}$ matrix sequentially. The first eigenvalue-eigenvector pair is enough to compute any t-th eigenvalue-eigenvector pair as described below. The analytical formulations reported next establish the correlation between t-th and $(t+1)$-th eigenvalues and corresponding eigenvectors, which help to generate correlated features sequentially. Also, it is clear that the t-th eigenvalue of the matrix $\mathcal{H}_{1r}$ is $\rho_r^2(t)$ and corresponding eigenvector is $w_{1r}(t)$. Using the Deflation method [293], we get

$$\mathcal{H}_{1r} w_{1r}(t) = \rho_r^2(t) w_{1r}(t)$$

$$\Rightarrow \mathcal{H}_{1r} w_{1r}(t) w_{1r}(t)^T = \rho_r^2(t) w_{1r}(t) w_{1r}(t)^T$$

$$\Rightarrow \mathcal{H}_{1r} - \mathcal{H}_{1r} w_{1r}(t) w_{1r}(t)^T = \mathcal{H}_{1r} - \rho_r^2(t) w_{1r}(t) w_{1r}(t)^T$$

$$\Rightarrow [\mathcal{H}_{1r} - \mathcal{H}_{1r} w_{1r}(t) w_{1r}^T(t)] w_{1r}(t+1) = [\mathcal{H}_{1r} - \rho_r^2(t) w_{1r}(t) w_{1r}^T(t)] w_{1r}(t+1)$$

$$\Rightarrow \mathcal{H}_{1r} w_{1r}(t+1) - \mathcal{H}_{1r} w_{1r}(t) w_{1r}^T(t) w_{1r}(t+1) = [\mathcal{H}_{1r} - \rho_r^2(t) w_{1r}(t) w_{1r}^T(t)] w_{1r}(t+1)$$

$$\Rightarrow \mathcal{H}_{1r} w_{1r}(t+1) = [\mathcal{H}_{1r} - \rho_r^2(t) w_{1r}(t) w_{1r}^T(t)] w_{1r}(t+1); \tag{5.40}$$

$\forall t \in \{1, 2, \cdots, (\mathcal{D}-1)\}$, where $\mathcal{D} \leqslant p$ and $\forall r \in \{1, 2, \cdots, \mathcal{T}\}$. On the other hand, $w_{1r}(t+1)$ is the $(t+1)$-th eigenvector of $\mathcal{H}_{1r}$, corresponding to the eigenvalue $\rho_r^2(t+1)$, that is,

$$\mathcal{H}_{1r} w_{1r}(t+1) = \rho_r^2(t+1) w_{1r}(t+1). \tag{5.41}$$

Comparing (5.3.3) and (5.41), we get,

$$[\mathcal{H}_{1r} - \rho_r^2(t) w_{1r}(t) w_{1r}^T(t)] w_{1r}(t+1) = \rho_r^2(t+1) w_{1r}(t+1). \tag{5.42}$$

From (5.42), it is evident that the $(t+1)$-th eigenvalue $\rho_r^2(t+1)$ and corresponding eigenvector $w_{1r}(t+1)$ of the matrix $\mathcal{H}_{1r}$ are the maximum eigenvalue and corresponding eigenvector of the matrix $[\mathcal{H}_{1r} - \rho_r^2(t) w_{1r}(t) w_{1r}^T(t)]$. To calculate the $(t+1)$-th eigenvalue and corresponding eigenvector, the matrix $\mathcal{H}_{1_r}$ can be calculated as follows:

$$\mathcal{H}_{1_r}(t+1) = \mathcal{H}_{1_r}(t) - \rho_r^2(t) w_{1_r}(t) w_{1_r}^T(t) = \mathcal{H}_{1_r} - \sum_{\ell=1}^{t} \rho_r^2(\ell) w_{1_r}(\ell) w_{1_r}^T(\ell); \tag{5.43}$$

where $\mathcal{H}_{1_r} = \mathcal{H}_{1_r}(1)$. Hence, to compute $(t+1)$-th eigenvector sequentially, the matrix $\mathcal{H}_{1_r}(t+1)$ can be computed by combining (5.3.2) and (5.43) as follows:

$$\mathcal{H}_{1_r}(t+1) = \sum_{j=2}^{\mathcal{M}} \Omega_1 \Delta_1^{-1} \Omega_1^T \mathcal{C}_{1j} \Omega_j \Delta_j^{-1} \Omega_j^T \mathcal{C}_{j1} + \sum_{s=1}^{r} \bar{\mathcal{G}}_{1_s} - \tilde{\mathcal{G}}_{1_s} - \hat{\mathcal{G}}_{1_s} - \sum_{\ell=1}^{t} \rho_r^2(l) w_{1_r}(\ell) w_{1_r}^T(\ell); \tag{5.44}$$

where $\forall r \in \{2, 3, \cdots, \mathcal{T}\}$, $\forall t \in \{1, 2, \cdots, (\mathcal{D}-1)\}$ and $\mathcal{D} \leqslant p$. Hence, using (5.39), the t-th multiset canonical variables can be computed sequentially as

$$\begin{aligned} \mathcal{U}_{(i+1)_r}(t) &= w_{(i+1)_r}^T(t) \mathcal{X}_{(i+1)} \\ &= w_{1_r}^T(t) \left[\prod_{j=2}^{i+1} \mathcal{C}_{(j-1)j} [\mathcal{C}_{jj} + (\mathfrak{r}_j + k_j d_j) I]^{-1} \right] \mathcal{X}_{(i+1)} \\ &= w_{1_r}^T(t) \left[\prod_{j=2}^{i+1} \mathcal{C}_{(j-1)j} \left(\sum_{\ell=1}^{m_j} \frac{1}{(\delta_{j_\ell} + k_j d_j)} \omega_{j_\ell} \omega_{j_\ell}^T \right) \right] \mathcal{X}_{(i+1)}; \end{aligned} \tag{5.45}$$

$\forall i \in \{1, 2, \cdots, (\mathcal{M}-1)\}$, $\forall t \in \{1, 2, \cdots, \mathcal{D}\}$, and $\forall r \in \{1, 2, \cdots, \mathcal{T}\}$

$$\text{and} \quad \mathcal{U}_{1_r}(t) = w_{1_r}^T(t) \mathcal{X}_1. \tag{5.46}$$

Finally, $\mathcal{D}$ features can be extracted sequentially as follows:

$$\mathcal{F}_r(t) = \sum_{i=1}^{\mathcal{M}} \mathcal{U}_{i_r}(t). \tag{5.47}$$

5.3.4 Selection of Optimum Regularization Parameter

Let us assume that each attribute is centered to have zero mean across the samples. Each regularization parameter $\mathfrak{r}_i$ is bounded $[\mathfrak{r}_{min}, \mathfrak{r}_{max}]$, where $\mathfrak{r}_{min} \leqslant \mathfrak{r}_i \leqslant \mathfrak{r}_{max}$. Let the number of all plausible values of $\mathfrak{r}_i$ is denoted by $\mathfrak{t}_i$, $\forall i \in \{1, 2, \cdots, \mathcal{M}\}$ within that range. Let $\mathcal{F}_r(t)$ be the t-th feature, extracted from the $\mathcal{M}$ multidimensional data sets with r-th combination of regularization parameters of $\{\mathfrak{r}_i\}$. The relevance of the feature $\mathcal{F}_r(t)$ with respect to the sample categories $\mathbb{D}$ is denoted by $\gamma_{\mathcal{F}_r(t)}(\mathbb{D})$. Let $\sigma_{\{\mathcal{F}_r(t), \mathcal{F}_\ell\}}(\mathbb{D}, \mathcal{F}_r(t))$ denote the significance of the feature $\mathcal{F}_r(t)$ with respect to the already-selected feature $\mathcal{F}_\ell \in \mathbb{S}$, where $\mathbb{S}$ denotes the set of selected features and initially $\mathbb{S} \leftarrow \emptyset$. The optimal regularization parameters for each extracted feature can be selected by using the relevance and significance of that feature. Let us assume that all the t-th extracted features which are computed by using all r-th combinations of $\{\mathfrak{r}_i\}$, are contained in the set $\mathbb{C}$, where $\forall t \in \{1, 2, \cdots, \mathcal{D} \leqslant p\}$ and $\forall r \in \{1, 2, \cdots, \mathcal{T}\}$. For $t = 1$, the most relevant feature is picked up from $\mathbb{C}$ and is put into $\mathbb{S}$, that is,

$$\mathcal{F}(t) = \arg \max_{\mathcal{F}_r(t) \in \mathbb{C}} \left\{ \gamma_{\mathcal{F}_r(t)}(\mathbb{D}) \right\}; \tag{5.48}$$

while for $t > 1$, the feature that has the highest relevance among the features of $\mathbb{C}$ and the significance with respect to the features of $\mathbb{S}$ is chosen as follows:

$$\mathcal{F}(t) = \arg \max_{\mathcal{F}_r(t) \in \mathbb{C}} \left\{ \gamma_{\mathcal{F}_r(t)}(\mathbb{D}) + \frac{1}{t-1} \sum_{\mathcal{F}_\ell \in \mathbb{S}} \sigma_{\{\mathcal{F}_r(t), \mathcal{F}_\ell\}}(\mathbb{D}, \mathcal{F}_r(t)) \right\}. \tag{5.49}$$

Thus, the problem of generating a set of most significant and relevant features $\mathbb{S}$ from a multiblock data set, based on all possible combinations of $\{\mathfrak{r}_i\}$, can be addressed by Algorithm 5.1.

In the current research work, both significance and relevance of an extracted feature are computed by using the concept of the rough hypercuboid approach [172]. It helps to optimize the regularization parameters.

5.3.5 Complexity Analysis

This section presents the time and space complexity of the proposed algorithm.

5.3.5.1 Time Complexity

Let $\mathcal{X}_1 \in \Re^{m_1 \times n}$, $\mathcal{X}_2 \in \Re^{m_2 \times n}$, $\cdots$, $\mathcal{X}_{\mathcal{M}} \in \Re^{m_{\mathcal{M}} \times n}$ be the $\mathcal{M}$ multidimensional data sets with c classes, n samples, and dimensions $m_1, m_2, \cdots, m_{\mathcal{M}}$, respectively, where $m_1 \leqslant m_2 \leqslant \cdots \leqslant m_{\mathcal{M}}$. Let us assume that the regularization parameter $\mathfrak{r}_i$ has $\mathfrak{t}_i$ possible values, $\forall i \in \{1, 2, \cdots, \mathcal{M}\}$. In Step 1, all the cross-covariance matrices $\{\mathcal{C}_{ij}\}$ are computed with complexity $\mathcal{O}(\sum\limits_{i<j} m_i m_j n) \approx \mathcal{O}(m_{\mathcal{M}} m_{\mathcal{M}-1} n)$; whereas the total time complexity to compute all the covariance matrices $\{\mathcal{C}_{ii}\}$ in Step 2 is $\mathcal{O}(\sum\limits_i m_i^2 n) \approx \mathcal{O}(m_{\mathcal{M}}^2 n)$. All the eigenvalues $\delta_{i\ell}$, along with corresponding eigenvectors $\omega_{i\ell}$, are computed with computational complexity $\mathcal{O}(\sum\limits_i m_i^3) \approx \mathcal{O}(m_{\mathcal{M}}^3)$; $\forall \ell \in \{1, 2, \cdots, m_i\}$, in Step 3. On the other hand, Step 4 and Step 5 have constant time complexity of $\mathcal{O}(1)$. Thus, the total computational complexity of these five steps is $\mathcal{O}(m_{\mathcal{M}} m_{\mathcal{M}} - 1n + m_{\mathcal{M}}^2 n + m_{\mathcal{M}}^3) \approx \mathcal{O}(m_{\mathcal{M}}^3)$ as $n << m_{\mathcal{M}}$.

In Step 6, there is a loop that has to be executed $\mathcal{D}$ times. The loop is started with constant time complexity $\mathcal{O}(1)$, followed by another loop that has to be implemented $\mathcal{T}$ times. The complexity to compute $\mathcal{H}_{1r}(t)$ is $\mathcal{O}(\sum\limits_i m_1^3 + m_i^3 + m_1^2 m_i + m_1 m_i^2 + \mathcal{T}(\sum\limits_i m_1^3 + m_i^3 + m_1^2 m_i + m_1 m_i^2) + \mathcal{D} m_1^2) \approx \mathcal{O}(\mathcal{T} m_{\mathcal{M}}^3)$. The eigenvector of the matrix $\mathcal{H}_{1r}(t)$ can be calculated with computational complexity $\mathcal{O}(m_1^2)$. In Step 6(II)(iii), the canonical variable $\mathcal{U}_{ir}(t)$ can be computed with complexity $\mathcal{O}(\sum\limits_i m_i n) \approx \mathcal{O}(m_{\mathcal{M}} n)$. Hence, a feature $\mathcal{F}_r(t)$ can be extracted with computational complexity $\mathcal{O}(n)$. Both relevance and significance of a feature have identical time complexity, which is given as $\mathcal{O}(cn)$. Thus, the total time complexity to execute the loop $\mathcal{T}$ times is $\mathcal{O}(\mathcal{T} m_{\mathcal{M}}^3 + m_1^2 + m_{\mathcal{M}} n + n + cn) \approx \mathcal{O}(\mathcal{T} m_{\mathcal{M}}^3)$. The selection of a

Algorithm 5.1 ReDMiCA: Regularized Discriminant Multiset CCA

```
Input: $\mathcal{M}$ multidimensional variables $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_{\mathcal{M}}$.
Output: A set $\mathbb{S}$ of $\mathcal{D}$ selected features.
Calculate the cross-covariance matrix $\mathcal{C}_{ij}$ of $\mathcal{X}_i$ and $\mathcal{X}_j$, $\forall i \in \{1, 2, \cdots, \mathcal{M}\}$ and $\forall j \in \{1, 2, \cdots, \mathcal{M}\}$ and $i \neq j$ and $i < j$.
Calculate the covariance matrix $\mathcal{C}_{ii}$ of $\mathcal{X}_i$, $\forall i \in \{1, 2, \cdots, \mathcal{M}\}$.
Calculate the eigenvalues $\delta_{i\ell}$ and corresponding eigenvectors $\omega_{i\ell}$ of $\mathcal{C}_{ii}$, $\forall \ell \in \{1, 2, \cdots, m_i\}$ and $\forall i \in \{1, 2, \cdots, \mathcal{M}\}$.
Compute the diagonal matrix $\Delta_i \in \Re^{m_i \times m_i}$, whose diagonal elements are $\delta_{i\ell}$, and the square matrix $\Omega_i \in \Re^{m_i \times m_i}$, whose $\ell$-th column is $\omega_{i\ell}$, $\forall \ell \in \{1, 2, \cdots, m_i\}$ and $\forall i \in \{1, 2, \cdots, \mathcal{M}\}$.
Initialize $\mathbb{S} = \emptyset$ and $t = 1$.
for each $t \leqslant \mathcal{D}$ do
    (I) Initialize $\mathbb{C} = \emptyset$.
    (II) for each $r$-th combinations of regularization parameters $\{\mathfrak{r}_i\}$, where $\forall r \in \{1, 2, \cdots, \mathcal{T}\}$ and $\forall i \in \{1, 2, \cdots, \mathcal{M}\}$. do
        (i) Calculate $\mathcal{H}_{1r}(t)$ using (5.3.2) if $t = 1$, otherwise using (5.44).
        (ii) Calculate largest eigenvalue $\rho_r^2(t)$ and eigenvector $w_{1r}(t)$ of the matrix $\mathcal{H}_{1r}(t)$, where $w_{1r}(t)$ is the $t$-th basis vector of first multidimensional variable.
        (iii) Calculate the $t$-th canonical variable $\mathcal{U}_{ir}(t)$ using (5.46) if $i = 1$, otherwise using (5.3.3).
        (iv) Extract the $t$-th feature $\mathcal{F}_r(t)$ corresponding to $r$-th combination of $\{\mathfrak{r}_i\}$ using (5.47).
        (v) Compute the relevance $\gamma_{\mathcal{F}_r(t)}(\mathbb{D})$ of the feature $\mathcal{F}_r(t)$ with respect to the class labels $\mathbb{D}$.
        (vi) Calculate the significance $\sigma_{\{\mathcal{F}_r(t), \mathcal{F}_\ell\}}(\mathbb{D}, \mathcal{F}_r(t))$ of the feature $\mathcal{F}_r(t)$ with respect to each $\mathcal{F}_\ell \in \mathbb{S}$.
        (vii) Add $\mathcal{F}_r(t)$ to $\mathbb{C}$ if its significance is non-zero with respect to each of the selected features of $\mathbb{S}$. In effect, $\mathbb{C} = \mathbb{C} \bigcup \mathcal{F}_r(t)$.
    (III) end for
    (IV) If $\mathbb{C} \neq \emptyset$, choose a feature as $t$-th feature $\mathcal{F}_r(t)$ from $\mathbb{C}$, which maximizes the condition (5.48) when $t = 1$, otherwise (5.49). In effect, $\mathbb{S} = \mathbb{S} \bigcup \mathcal{F}_r(t)$ and $t = t + 1$.
end for
Stop.
```

feature from $\mathcal{T}$ candidate features by maximizing both relevance and significance, which is carried out in Step 6(IV), has complexity $\mathcal{O}(\mathcal{T})$. Thus, the total complexity to execute the loop $\mathcal{D}$ times is $\mathcal{O}(\mathcal{D}(\mathcal{T}m_{\mathcal{M}}^3+\mathcal{T})) \approx \mathcal{O}(\mathcal{D}\mathcal{T}m_{\mathcal{M}}^3)$. So, the proposed sequential multiblock data integration algorithm has computational complexity of $\mathcal{O}(m_{\mathcal{M}}^3 + \mathcal{D}\mathcal{T}m_{\mathcal{M}}^3) \approx \mathcal{O}(\mathcal{D}\mathcal{T}m_{\mathcal{M}}^3)$.

5.3.5.2 Space Complexity

In Step 1, all cross-covariance matrices $\{\mathcal{C}_{ij}\}$ can be computed with space complexity $\mathcal{O}(\sum\limits_{i<j} m_i m_j) \approx \mathcal{O}(m_{\mathcal{M}} m_{\mathcal{M}-1})$; whereas the total space complexity to compute all the covariance matrices $\{\mathcal{C}_{ii}\}$ is $\mathcal{O}(\sum\limits_{i} m_i^2) \approx \mathcal{O}(m_{\mathcal{M}}^2)$. All the eigenvalues $\delta_{i\ell}$, along with corresponding eigenvectors $\omega_{i\ell}$, are computed with space complexity $\mathcal{O}(\sum\limits_{i} m_i + m_i^2) \approx \mathcal{O}(m_{\mathcal{M}}^2)$; $\forall \ell \in \{1, 2, \cdots, m_i\}$, in Step 3. On the other hand, Step 4 has $\mathcal{O}(m_{\mathcal{M}}^2)$ space complexity. Step 5 has constant space complexity of $\mathcal{O}(1)$. Thus, the total space complexity of these five steps is $\mathcal{O}(m_{\mathcal{M}}m_{\mathcal{M}} - 1 + m_{\mathcal{M}}^2 + m_{\mathcal{M}}^2 + m_{\mathcal{M}}^2) \approx \mathcal{O}(m_{\mathcal{M}}^2)$.

In Step 6, there is a loop that has to be executed $\mathcal{D}$ times. The loop is started with constant space complexity $\mathcal{O}(1)$, followed by another loop that has to be implemented $\mathcal{T}$ times. The space complexity to compute $\mathcal{H}_{1r}(t)$ is $\mathcal{O}(\mathcal{D}\mathcal{T}(\sum\limits_{i} m_1^2 + m_1^2 + m_1 m_i + m_1 m_i + m_i^2 + m_i^2 + m_1 m_i + m_1^2 + m_1^2 + m_1^2 + m_1^2)) \approx \mathcal{O}(\mathcal{D}\mathcal{T}m_{\mathcal{M}}^2)$. The eigenvalues and eigenvectors of the matrix $\mathcal{H}_{1r}(t)$ can be stored with complexity $\mathcal{O}(\mathcal{D} + \mathcal{D}m_1)$. In Step 6(II)(iii), the canonical variable $\mathcal{U}_{ir}(t)$ can be computed with space complexity $\mathcal{O}(\mathcal{D}n + m_{\mathcal{M}}n + m_{\mathcal{M}}m_{\mathcal{M}} - 1)$. Hence, a feature $\mathcal{F}_r(t)$ can be stored with space complexity $\mathcal{O}(n)$. Both relevance and significance of a feature have identical space complexity, which is $\mathcal{O}(cn)$. So, for $\mathcal{T}$ candidate features, this is given as $\mathcal{O}(\mathcal{T}cn)$. Thus, the total space complexity to execute the Step 6(II)(iv)-(vi) is $\mathcal{O}(\mathcal{D}(n + \mathcal{T}cn))$. The selection of a feature from $\mathcal{T}$ candidate features by maximizing both relevance and significance, which is carried out in Step 6(IV), has constant space complexity $\mathcal{O}(1)$. Thus, the total space complexity to execute the Step 6 is $\mathcal{O}(\mathcal{D}\mathcal{T}m_{\mathcal{M}}^2 + \mathcal{D} + \mathcal{D}m_1 + \mathcal{D}n + m_{\mathcal{M}}n + m_{\mathcal{M}}m_{\mathcal{M}} - 1 + \mathcal{D}n + \mathcal{D}\mathcal{T}cn) \approx \mathcal{O}(\mathcal{D}\mathcal{T}m_{\mathcal{M}}^2)$, as $n, c << m_{\mathcal{M}}$. Thus, the proposed algorithm has space complexity of $\mathcal{O}(m_{\mathcal{M}}^2 + \mathcal{D}\mathcal{T}m_{\mathcal{M}}^2) \approx \mathcal{O}(\mathcal{D}\mathcal{T}m_{\mathcal{M}}^2)$.

5.4 Performance Analysis

The performance of the proposed sequential feature extraction algorithm, termed as ReDMiCA, is extensively studied and compared with that of several existing multimodal data integration algorithms. To evaluate the performance of different algorithms, support vector machine with linear kernels is used. Each regularization parameter is varied in between 0.0 and 1.0, with a difference of 0.1. Five benchmark data sets, namely, CiteSeer, Handwritten, NUS-WIDE-OBJECT (NW-OBJECT), Reuters, and Caltech; and five cancer data sets, namely, glioblastoma multiforme (GBM), lung (LUNG), kidney (KIDNEY), lower grade glioma (LGG) and ovarian serous cystadenocarcinoma (OV), are used in the current research work. All the data sets are summarized in Table 5.1 and Table 5.2 and briefly described in Appendix A. The proposed algorithm is implemented in C language and run in Ubuntu 14.04 LTS having machine configuration Intel(R) Core(TM) i7-4790 CPU @ 3.60GHz×8 and 32 GB RAM. The source code of the proposed algorithm, written in C language, is available at https://www.isical.ac.in/~bibl/results/redmica/redmica.html.

Table 5.1: Description of Benchmark Data Sets Used

Different	Number of		Cardinality of Different Views					
Data Sets	Classes	Samples	m_1	m_2	m_3	m_4	m_5	m_6
CiteSeer	6	3309	3312	3312	3312	3703	-	-
NW-OBJECT	31	30000	64	73	128	144	225	-
Reuters	6	18758	11547	15506	21531	24892	34251	-
Handwritten	10	2000	6	47	64	76	216	240
Caltech	20	2386	40	48	254	512	928	1984

Table 5.2: Description of Omics Data Sets Used

Different	Number of		Cardinality of Different Views				
Data Sets	Classes	Samples	RNA	mDNA	miRNA	CNS	RPPA
GBM	5	213	12042	21422	534	4070	-
LUNG	2	546	20502	294668	216	49230	180
KIDNEY	2	305	20502	300451	209	9059	174
LGG	3	374	11973	293965	139	6261	181
OV	4	206	12042	20311	129	4332	195

The randomly selected 50% samples from each class are used for training and the rest are used for testing purposes for each of the data sets. The 10-fold cross-validation is also performed on each data sets to assess the performance of the proposed algorithm statistically. To analyze the statistical significance of the derived results, paired-t test (one-tailed), Wilcoxon signed rank test (one-tailed) and Friedman test (one-tailed), with a 95% confidence level, are used to compute the p-values. For each data set, 25 top-ranked correlated features are selected for the analysis.

5.4.1 Importance of Various Criteria of MCCA

Table 5.3, Table 5.4, and Table 5.5 compare the performance of the proposed ReDMiCA algorithm with that of different criteria of the MCCA, namely, SUMCOR, MAXVAR, generalized variance (GENVAR), minimum variance (MINVAR), and sum of squared correlations (SSQCOR) [135]; and several existing algorithms. These tables present the classification accuracy on each data set in case of training-testing. The mean, median, and standard deviation of 10-fold cross-validation are also reported in Table 5.3, Table 5.4, and Table 5.5 for both the benchmark and omics data sets. To perform the statistical significance analysis, the p-values computed using different tests are reported in Table 5.6, Table 5.7, and Table 5.8. Figure 5.1, Figure 5.2, Figure 5.3, and Figure 5.4 compare the performance of the proposed algorithm with that of various criteria of the MCCA.

Table 5.3: Classification Accuracy and Execution Time of Different Algorithms on Handwritten, NW-OBJECT, LUNG, and KIDNEY Data Sets

Different Algorithms		Data Sets	Accuracy (Train-Test)	Accuracy for 10-Fold CV			Time (in sec.)	Data Sets	Accuracy (Train-Test)	Accuracy for 10-Fold CV			Time (in sec.)
				Mean	Median	StdDev				Mean	Median	StdDev	
MCCA	SUMCOR	Handwritten	0.870	0.820	0.825	0.024	114.4	LUNG	0.465	0.514	0.509	0.031	37931.0
	GENVAR		0.072	0.110	0.905	0.076	10.8		0.571	0.621	0.866	0.096	1760.3
	MAXVAR		0.039	0.088	0.075	0.044	9.6		0.755	0.709	0.580	0.101	3967.4
	MINVAR		0.143	0.114	0.090	0.046	10.1		0.784	0.727	0.696	0.090	9360.2
	SSQCOR		0.085	0.094	0.135	0.040	407.0		0.791	0.725	0.723	0.070	37210.1
RGCCA			0.903	0.910	0.093	0.012	9.9		0.879	0.877	0.732	0.042	755.4
GMCCA			0.112	0.113	0.108	0.020	19.9		0.689	0.684	0.696	0.075	141.5
GMKCCA			0.066	0.106	0.110	0.045	45.7		0.861	0.861	0.875	0.083	167.1
LasCCA			0.102	0.068	0.058	0.019	13.7		0.824	0.852	0.857	0.067	21.0
DisCCA			0.056	0.127	0.128	0.049	40.9		0.476	0.507	0.482	0.088	52.2
BsMCCA			0.122	0.119	0.098	0.066	110.2		0.894	0.829	0.893	0.160	198.0
MvDA			0.924	0.946	0.953	0.022	11.1		0.923	0.948	**0.964**	0.042	21.9
MvDA-VC			0.935	0.954	0.953	0.013	10.7		0.916	0.955	0.955	0.033	20.2
ReDMiCA			**0.963**	**0.969**	**0.970**	0.016	1615.0		**0.949**	**0.957**	0.955	0.032	6162.4
MCCA	SUMCOR	NW-OBJECT	0.303	0.322	0.321	0.005	90.8	KIDNEY	0.559	0.610	0.613	0.062	5940.9
	GENVAR		0.093	0.079	0.136	0.013	37.4		0.691	0.655	0.919	0.082	639.7
	MAXVAR		0.051	0.054	0.080	0.008	34.4		0.757	0.768	0.661	0.060	2528.5
	MINVAR		0.080	0.071	0.053	0.010	33.4		0.711	0.777	0.774	0.056	4572.6
	SSQCOR		0.095	0.073	0.071	0.008	87.9		0.757	0.839	0.774	0.091	6934.0
RGCCA			0.189	0.135	0.074	0.003	392.8		0.914	0.929	0.855	0.048	674.3
GMCCA			0.046	0.055	0.055	0.005	16592.0		0.855	0.758	0.774	0.063	176.6
GMKCCA			0.064	0.066	0.067	0.019	21403.7		0.829	0.816	0.839	0.079	195.0
LasCCA			0.074	0.084	0.083	0.006	76.1		0.743	0.790	0.774	0.110	30.6
DisCCA			0.109	0.106	0.105	0.016	82.0		0.586	0.610	0.645	0.089	44.9
BsMCCA			0.079	0.091	0.093	0.005	38.1		0.855	0.906	0.903	0.049	101.3
MvDA			0.290	0.280	0.279	0.008	53.2		0.928	0.932	0.935	0.024	15.2
MvDA-VC			0.286	0.279	0.281	0.008	73.8		0.947	0.942	0.935	0.025	15.4
ReDMiCA			**0.377**	**0.382**	**0.382**	0.009	1344.7		**0.961**	**0.971**	**0.968**	0.028	6774.5

Table 5.4: Classification Accuracy and Execution Time of Different Algorithms on Reuters, Caltech, LGG, and OV Data Sets

Different Algorithms		Data Sets	Accuracy (Train-Test)	Accuracy for 10-Fold CV			Time (in sec.)	Data Sets	Accuracy (Train-Test)	Accuracy for 10-Fold CV			Time (in sec.)
				Mean	Median	StdDev				Mean	Median	StdDev	
MCCA	SUMCOR	Reuters	0.575	0.657	0.661	0.013	13922.7	LGG	0.398	0.355	0.342	0.077	12578.5
	GENVAR		0.215	0.243	0.366	0.022	3511.3		0.484	0.439	0.474	0.074	411.9
	MAXVAR		0.298	0.278	0.238	0.019	3760.9		0.403	0.426	0.474	0.072	4812.6
	MINVAR		0.215	0.194	0.281	0.008	6395.6		0.409	0.429	0.447	0.062	9855.9
	SSQCOR		0.195	0.256	0.195	0.031	38119.2		0.376	0.447	0.408	0.081	16475.4
RGCCA			0.553	0.365	0.247	0.006	22973.0		0.414	0.450	0.461	0.110	687.0
GMCCA			0.248	0.313	0.310	0.017	69242.8		0.333	0.405	0.408	0.087	143.1
GMKCCA			0.287	0.195	0.193	0.034	91864.2		0.387	0.332	0.316	0.050	133.5
LasCCA			0.287	0.337	0.337	0.023	1306.5		0.441	0.387	0.368	0.079	18.2
DisCCA			0.233	0.257	0.222	0.082	4951.6		0.290	0.387	0.382	0.072	39.4
BsMCCA			**0.664**	0.409	0.407	0.015	1418.8		0.602	0.689	0.684	0.057	106.4
MvDA			0.560	0.578	0.579	0.012	3006.8		0.758	0.758	0.763	0.080	16.4
MvDA-VC			0.551	0.582	0.588	0.015	4215.6		0.731	0.811	0.789	0.078	16.9
ReDMiCA			0.662	**0.696**	**0.697**	0.008	11434.1		**0.946**	**0.850**	**0.842**	0.035	5958.2
MCCA	SUMCOR	Caltech	0.418	0.707	0.705	0.025	19819.1	OV	0.284	0.241	0.227	0.068	997.2
	GENVAR		0.574	0.448	0.325	0.069	501.6		0.275	0.314	0.318	0.095	612.4
	MAXVAR		0.733	0.727	0.484	0.024	223.8		0.294	0.550	0.318	0.152	2470.1
	MINVAR		0.730	0.700	0.724	0.019	259.3		0.500	0.573	0.477	0.177	4472.0
	SSQCOR		0.715	0.733	0.701	0.021	19897.2		0.490	0.582	0.545	0.128	6940.9
RGCCA			0.337	0.325	0.732	0.000	950.2		0.333	0.314	0.636	0.058	633.8
GMCCA			0.048	0.035	0.033	0.012	219.0		0.235	0.277	0.250	0.114	100.9
GMKCCA			0.075	0.092	0.096	0.011	296.7		0.324	0.309	0.318	0.143	107.7
LasCCA			0.041	0.046	0.037	0.024	55.8		0.324	0.314	0.295	0.079	15.3
DisCCA			0.032	0.115	0.100	0.062	74.9		0.245	0.255	0.273	0.081	35.9
BsMCCA			0.783	0.786	0.780	0.028	342.6		0.725	0.623	0.659	0.171	88.9
MvDA			0.763	0.788	0.789	0.015	54.6		0.500	0.655	0.682	0.114	14.9
MvDA-VC			0.753	0.785	**0.791**	0.020	54.7		0.559	0.573	0.568	0.125	14.1
ReDMiCA			**0.852**	**0.801**	**0.791**	0.026	8882.1		**0.941**	**0.709**	**0.727**	0.075	8641.2

Table 5.5: Classification Accuracy and Execution Time of Different Algorithms on CiteSeer and GBM Data Sets

Different		Data	Accuracy	Accuracy for 10-Fold CV			Time	Data	Accuracy	Accuracy for 10-Fold CV			Time
Algorithms		Sets	(Train-Test)	Mean	Median	StdDev	(in sec.)	Sets	(Train-Test)	Mean	Median	StdDev	(in sec.)
MCCA	SUMCOR	CiteSeer	0.581	0.592	0.599	0.028	6932.4	GBM	0.219	0.179	0.200	0.065	2562.5
	GENVAR		0.427	0.426	0.359	0.091	255.0		0.314	0.342	0.342	0.081	1863.1
	MAXVAR		0.567	0.560	0.453	0.010	71.2		0.371	0.588	0.463	0.101	1567.7
	MINVAR		0.501	0.533	0.562	0.026	113.7		0.362	0.558	0.500	0.125	1599.6
	SSQCOR		0.525	0.548	0.524	0.031	6789.3		0.457	0.513	0.433	0.104	2587.5
RGCCA			0.276	0.363	0.554	0.016	155.0		0.229	0.350	0.288	0.115	660.7
GMCCA			0.234	0.462	0.458	0.040	85.1		0.457	0.350	0.313	0.102	44.5
GMKCCA			0.250	0.157	0.159	0.027	247.8		0.181	0.238	0.233	0.074	46.2
LasCCA			0.223	0.484	0.480	0.025	256.6		0.571	0.617	0.463	0.094	342.7
DisCCA			0.207	0.266	0.267	0.017	35.7		0.390	0.350	0.321	0.115	123.1
BsMCCA			0.234	0.177	0.227	0.000	14.1		0.629	0.525	0.575	0.217	14.3
MvDA			0.377	0.414	0.416	0.020	30.5		0.629	0.692	0.588	0.077	38.2
MvDA-VC			0.435	0.477	0.473	0.032	27.2		**0.733**	0.671	0.600	0.099	36.3
ReDMiCA			**0.646**	**0.641**	**0.643**	0.027	447.6		0.714	**0.717**	**0.729**	0.047	2150.7

Table 5.6: Statistical Significance Analysis of Different Algorithms on CiteSeer and GBM Data Sets

Different		Data	p-values for 10-Fold CV			Data	p-values for 10-Fold CV		
Algorithms		Sets	Paired-t	Wilcoxon	Friedman	Sets	Paired-t	Wilcoxon	Friedman
MCCA	SUMCOR	CiteSeer	**2.57E-05**	**2.53E-03**	**1.57E-03**	GBM	**5.37E-09**	**2.50E-03**	**1.57E-03**
	GENVAR		**2.68E-05**	**2.53E-03**	**1.57E-03**		**4.04E-08**	**2.50E-03**	**1.57E-03**
	MAXVAR		**8.43E-06**	**2.47E-03**	**1.57E-03**		**2.66E-03**	**8.20E-03**	**1.14E-02**
	MINVAR		**2.75E-06**	**2.52E-03**	**1.57E-03**		**1.70E-03**	**5.81E-03**	**4.68E-03**
	SSQCOR		**5.67E-05**	**3.42E-03**	**1.14E-02**		**9.98E-06**	**2.39E-03**	**1.57E-03**
RGCCA			**3.93E-10**	**2.53E-03**	**1.57E-03**		**1.58E-06**	**2.46E-03**	**1.57E-03**
GMCCA			**4.64E-09**	**2.50E-03**	**1.57E-03**		**2.51E-06**	**2.49E-03**	**1.57E-03**
GMKCCA			**8.58E-11**	**2.53E-03**	**1.57E-03**		**1.41E-08**	**2.38E-03**	**1.57E-03**
LasCCA			**2.95E-10**	**2.52E-03**	**1.57E-03**		**2.20E-03**	**5.61E-03**	**4.68E-03**
DisCCA			**2.01E-11**	**2.53E-03**	**1.57E-03**		**3.59E-07**	**2.42E-03**	**1.57E-03**
BsMCCA			**1.23E-10**	**2.53E-03**	**1.57E-03**		**1.69E-02**	**1.88E-02**	*9.56E-02*
MvDA			**9.74E-09**	**2.53E-03**	**1.57E-03**		*1.91E-01*	*3.67E-01*	7.06E-01
MvDA-VC			**3.19E-07**	**2.52E-03**	**1.57E-03**		*6.00E-02*	*6.13E-02*	*1.57E-01*

Table 5.7: Statistical Significance Analysis of Different Algorithms on Handwritten, NW-OBJECT, LUNG, and KIDNEY Data Sets

Different Algorithms		Data Sets	p-values for 10-Fold CV			Data Sets	p-values for 10-Fold CV		
			Paired-t	Wilcoxon	Friedman		Paired-t	Wilcoxon	Friedman
MCCA	SUMCOR	Handwritten	**1.93E-08**	**2.50E-03**	**1.57E-03**	LUNG	**4.72E-12**	**2.50E-03**	**1.57E-03**
	GENVAR		**3.91E-11**	**2.52E-03**	**1.57E-03**		**8.64E-05**	**2.46E-03**	**1.57E-03**
	MAXVAR		**2.22E-13**	**2.52E-03**	**1.57E-03**		**1.28E-06**	**2.52E-03**	**1.57E-03**
	MINVAR		**1.11E-12**	**2.50E-03**	**1.57E-03**		**1.38E-05**	**2.49E-03**	**1.57E-03**
	SSQCOR		**1.26E-14**	**2.52E-03**	**1.57E-03**		**1.09E-05**	**2.38E-03**	**1.57E-03**
RGCCA			**1.56E-05**	**3.82E-03**	**2.70E-03**		**8.42E-06**	**2.46E-03**	**1.57E-03**
GMCCA			**8.74E-15**	**2.47E-03**	**1.57E-03**		**1.78E-07**	**2.49E-03**	**1.57E-03**
GMKCCA			**9.23E-13**	**2.53E-03**	**1.57E-03**		**2.20E-03**	**3.82E-03**	**2.70E-03**
LasCCA			**1.12E-15**	**2.46E-03**	**1.57E-03**		**9.24E-05**	**2.50E-03**	**1.57E-03**
DisCCA			**3.03E-12**	**2.50E-03**	**1.57E-03**		**3.92E-08**	**2.52E-03**	**1.57E-03**
BsMCCA			**7.94E-12**	**2.52E-03**	**1.57E-03**		**1.52E-02**	**1.39E-02**	*9.56E-02*
MvDA			**1.34E-03**	**2.42E-03**	**1.57E-03**		*2.26E-01*	*1.30E-01*	*1.57E-01*
MvDA-VC			**2.47E-02**	**2.30E-02**	*5.78E-02*		*4.16E-01*	*3.98E-01*	7.06E-01
MCCA	SUMCOR	NW-OBJECT	**3.44E-10**	**2.53E-03**	**1.57E-03**	KIDNEY	**5.60E-09**	**2.38E-03**	**1.57E-03**
	GENVAR		**8.17E-14**	**2.53E-03**	**1.57E-03**		**6.64E-03**	**1.28E-02**	**1.43E-02**
	MAXVAR		**4.85E-16**	**2.53E-03**	**1.57E-03**		**1.41E-06**	**2.50E-03**	**1.57E-03**
	MINVAR		**9.57E-14**	**2.53E-03**	**1.57E-03**		**2.34E-06**	**2.52E-03**	**1.57E-03**
	SSQCOR		**1.80E-15**	**2.53E-03**	**1.57E-03**		**2.77E-06**	**2.50E-03**	**1.57E-03**
RGCCA			**7.25E-15**	**2.53E-03**	**1.57E-03**		**1.89E-03**	**4.49E-03**	**1.14E-02**
GMCCA			**4.80E-15**	**2.53E-03**	**1.57E-03**		**1.64E-06**	**2.52E-03**	**1.57E-03**
GMKCCA			**1.93E-12**	**2.53E-03**	**1.57E-03**		**4.21E-05**	**2.50E-03**	**1.57E-03**
LasCCA			**4.84E-15**	**2.53E-03**	**1.57E-03**		**7.98E-04**	**2.49E-03**	**1.57E-03**
DisCCA			**4.41E-13**	**2.52E-03**	**1.57E-03**		**1.19E-07**	**2.40E-03**	**1.57E-03**
BsMCCA			**1.81E-14**	**2.52E-03**	**1.57E-03**		**1.45E-03**	**2.20E-03**	**1.57E-03**
MvDA			**1.51E-10**	**2.52E-03**	**1.57E-03**		**2.56E-03**	**8.18E-03**	**8.15E-03**
MvDA-VC			**7.07E-10**	**2.53E-03**	**1.57E-03**		**9.36E-03**	**1.05E-02**	**1.96E-02**

5.4.1.1 Performance on Benchmark Data

The results presented in Figure 5.1 and Figure 5.2 convey that the SUMCOR provides the highest accuracy irrespective of the features among different criteria of the MCCA on both the Handwritten and Caltech data sets. However, the performance of the proposed algorithm is significantly higher as compared to that of various criteria of the MCCA, irrespective of the generated features and data sets used. The results reported in Table 5.3, Table 5.4, and Table 5.5 demonstrate that the SUMCOR attains highest accuracy of 0.581, 0.870, 0.303, and 0.575 on CiteSeer, Handwritten, NW-OBJECT, and Reuters data sets, respectively; and MAXVAR provides the highest classification accuracy of 0.732573 on Caltech data set, among five criteria of the MCCA. The results reported in Table 5.6, Table 5.7, and Table 5.8 establish that the proposed ReDMiCA algorithm attains significantly better p-values than the different five criteria of the MCCA, irrespective of the significance analysis and data sets used in all 75 cases. However, the proposed algorithm achieves the highest classification accuracy on both data sets. The scatter plots of the first two extracted features using different criteria of the MCCA and proposed algorithm on

Table 5.8: Statistical Significance Analysis of Different Algorithms on Reuters, Caltech, LGG, and OV Data Sets

Different Algorithms		Data Sets	*p*-values for 10-Fold CV			Data Sets	*p*-values for 10-Fold CV		
			Paired-*t*	Wilcoxon	Friedman		Paired-*t*	Wilcoxon	Friedman
MCCA	SUMCOR	Reuters	**7.01E-08**	**2.53E-03**	**1.57E-03**	LGG	**2.17E-08**	**2.49E-03**	**1.57E-03**
	GENVAR		**4.06E-13**	**2.53E-03**	**1.57E-03**		**1.39E-06**	**2.25E-03**	**1.57E-03**
	MAXVAR		**1.94E-13**	**2.53E-03**	**1.57E-03**		**2.45E-08**	**2.46E-03**	**1.57E-03**
	MINVAR		**2.20E-16**	**2.52E-03**	**1.57E-03**		**5.49E-09**	**2.50E-03**	**1.57E-03**
	SSQCOR		**5.57E-12**	**2.53E-03**	**1.57E-03**		**1.21E-08**	**2.50E-03**	**1.57E-03**
RGCCA			**5.22E-15**	**2.53E-03**	**1.57E-03**		**1.90E-07**	**2.52E-03**	**1.57E-03**
GMCCA			**2.61E-13**	**2.53E-03**	**1.57E-03**		**1.09E-07**	**2.53E-03**	**1.57E-03**
GMKCCA			**4.92E-13**	**2.53E-03**	**1.57E-03**		**2.43E-12**	**2.34E-03**	**1.57E-03**
LasCCA			**7.45E-12**	**2.53E-03**	**1.57E-03**		**1.89E-08**	**2.46E-03**	**1.57E-03**
DisCCA			**2.79E-08**	**2.53E-03**	**1.57E-03**		**3.41E-08**	**2.50E-03**	**1.57E-03**
BsMCCA			**1.35E-12**	**2.53E-03**	**1.57E-03**		**2.91E-05**	**2.50E-03**	**1.57E-03**
MvDA			**4.52E-11**	**2.53E-03**	**1.57E-03**		**1.81E-03**	**5.66E-03**	**4.68E-03**
MvDA-VC			**3.85E-10**	**2.52E-03**	**1.57E-03**		*9.27E-02*	*1.20E-01*	*2.06E-01*
MCCA	SUMCOR	Caltech	**1.34E-05**	**2.50E-03**	**1.57E-03**	OV	**2.18E-08**	**2.49E-03**	**1.57E-03**
	GENVAR		**3.35E-08**	**2.53E-03**	**1.57E-03**		**5.80E-07**	**2.52E-03**	**1.57E-03**
	MAXVAR		**7.85E-06**	**2.53E-03**	**1.57E-03**		**5.16E-06**	**2.52E-03**	**1.57E-03**
	MINVAR		**1.44E-06**	**2.53E-03**	**1.57E-03**		**1.03E-02**	**1.42E-02**	*9.56E-02*
	SSQCOR		**6.63E-07**	**2.53E-03**	**1.57E-03**		**3.63E-02**	**4.61E-02**	*2.06E-01*
RGCCA			**3.38E-13**	**2.49E-03**	**1.57E-03**		**3.69E-03**	**5.66E-03**	**4.68E-03**
GMCCA			**1.68E-14**	**2.52E-03**	**1.57E-03**		**6.82E-06**	**2.52E-03**	**1.57E-03**
GMKCCA			**1.40E-14**	**2.53E-03**	**1.57E-03**		**1.83E-05**	**2.52E-03**	**1.57E-03**
LasCCA			**8.22E-15**	**2.53E-03**	**1.57E-03**		**2.09E-08**	**2.47E-03**	**1.57E-03**
DisCCA			**2.55E-10**	**2.53E-03**	**1.57E-03**		**4.85E-07**	**2.52E-03**	**1.57E-03**
BsMCCA			**1.27E-02**	**1.20E-02**	**1.96E-02**		**3.65E-02**	**4.52E-02**	*2.57E-01*
MvDA			*8.70E-02*	*8.64E-02*	*3.17E-01*		*1.35E-01*	*1.99E-01*	*4.80E-01*
MvDA-VC			*7.41E-02*	*1.31E-01*	1.00E+00		**8.30E-03**	**1.08E-02**	*5.78E-02*

benchmark data sets are reported in Figure 5.5, which also establish the superiority of the proposed ReDMiCA algorithm over different criteria of the MCCA. The value of the class separability index (CSI) is also reported at the top of each figure. From the results reported in Figure 5.5, it is evident that the CSI of the extracted features using different criteria of the MCCA is higher compared to that of the proposed ReDMiCA algorithm.

5.4.1.2 Performance on Omics Data

All the results reported in Figure 5.3, and Figure 5.4 demonstrate that the classification accuracy of the proposed algorithm is significantly higher compared to that of various criteria of the MCCA, irrespective of the generated features, data sets and experimental setup used. All the results reported in Table 5.3, Table 5.4, and Table 5.5 confirm that the proposed algorithm attains highest mean and median accuracy, irrespective of the data sets. Out of total 75 cases, ReDMiCA achieves significantly better (marked in bold) *p*-values than different criteria of the MCCA in 73 cases, with 0.05 significance level. On the other hand, the proposed algorithm provides better but not significant (marked in italics) *p*-values in

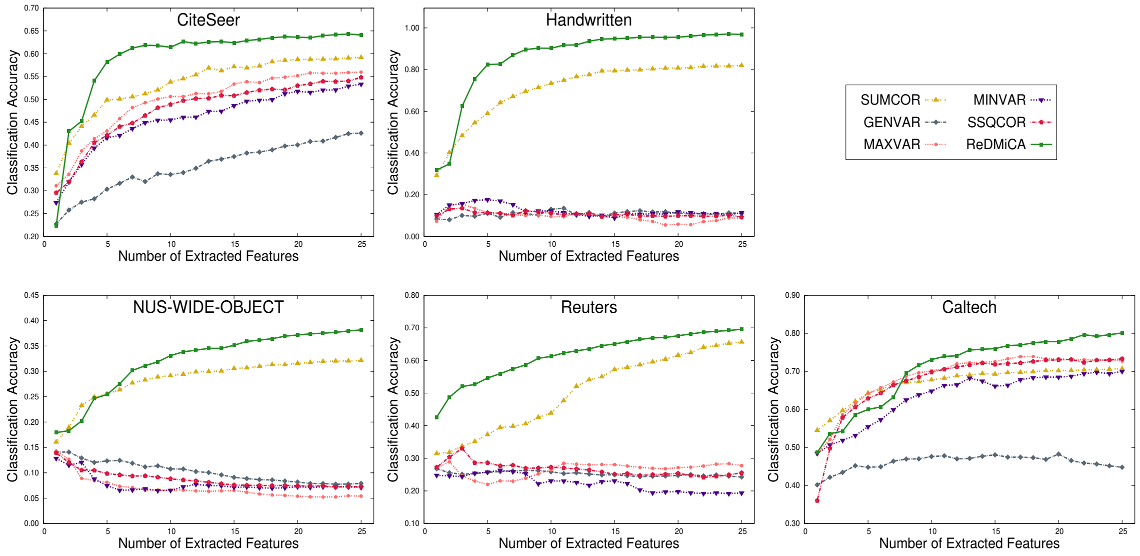


Figure 5.1: Variation of classification accuracy with respect to number of extracted features for different criteria of the MCCA and proposed (ReDMiCA) algorithm on benchmark data sets for 10-fold CV.

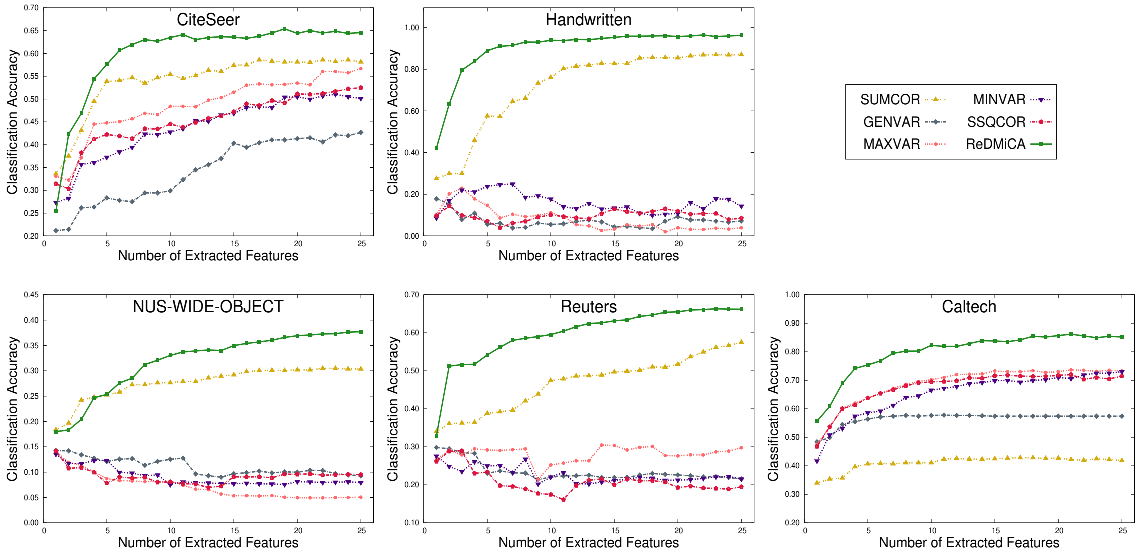


Figure 5.2: Variation of classification accuracy with respect to number of extracted features for different criteria of the MCCA and proposed (ReDMiCA) algorithm on benchmark data sets for training-testing.

only 2 cases, for MINVAR and SSQCOR using the Friedman test on OV data set. Figure 5.6 shows the scatter plots, along with the CSI, of the first two extracted features using different criteria of the MCCA and the proposed algorithm. From the results reported in Figure 5.6, it is evident that the CSI of the extracted features using different criteria of the

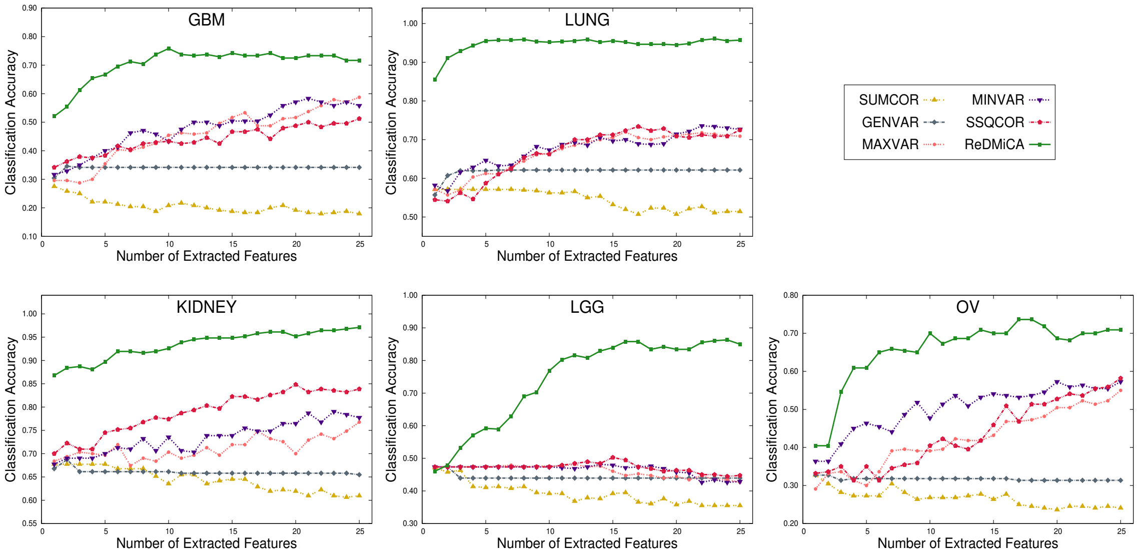


Figure 5.3: Variation of classification accuracy with respect to number of extracted features for different criteria of the MCCA and proposed (ReDMiCA) algorithm on omics data sets for 10-fold CV.

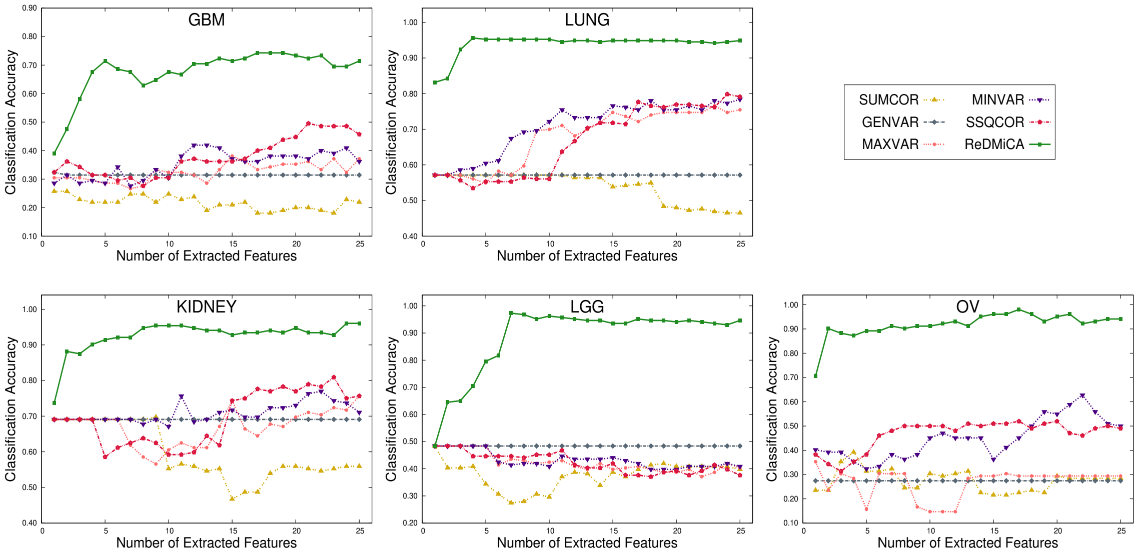
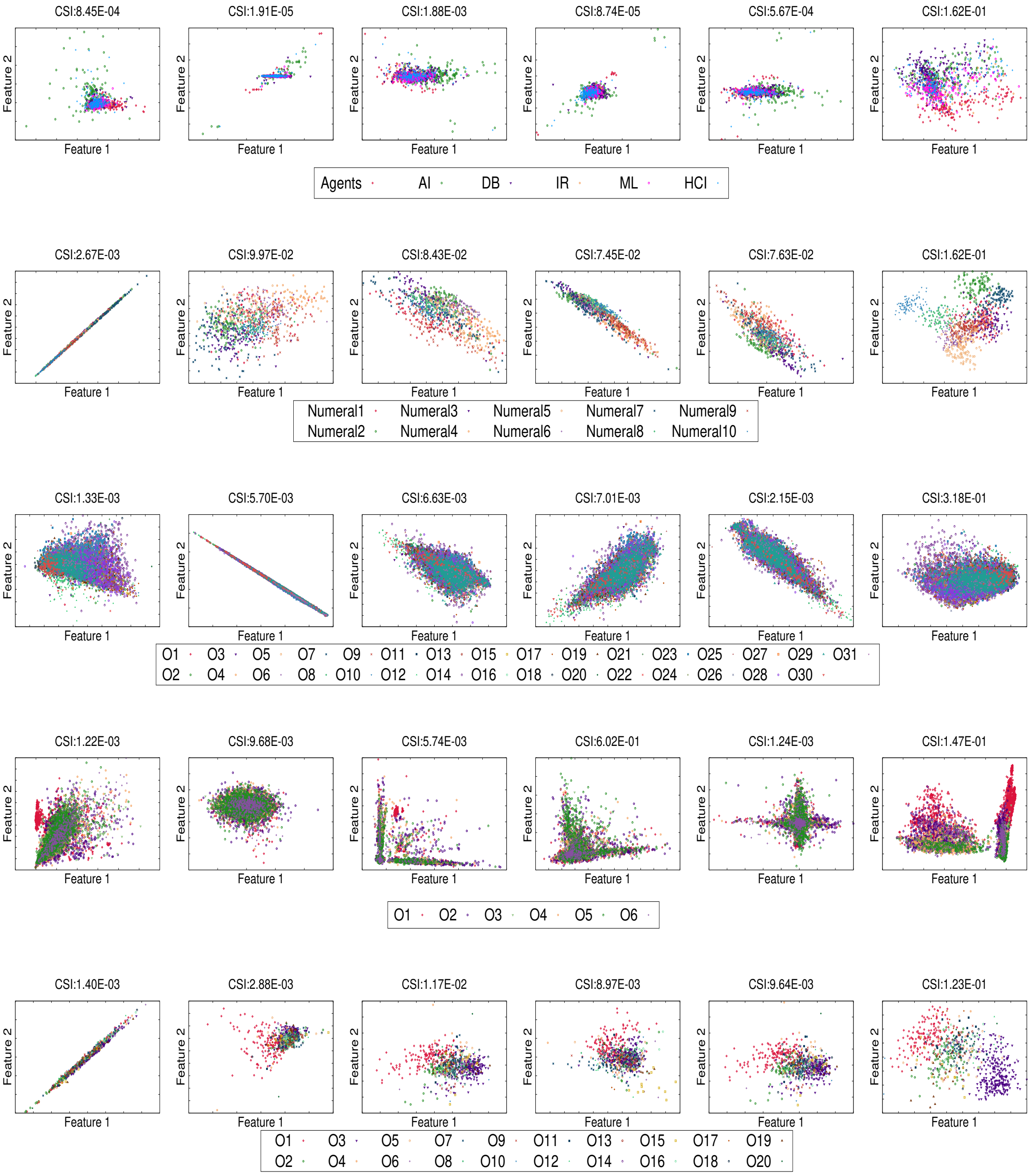


Figure 5.4: Variation of classification accuracy with respect to number of extracted features for different criteria of the MCCA and proposed (ReDMiCA) algorithm on omics data sets for training-testing.

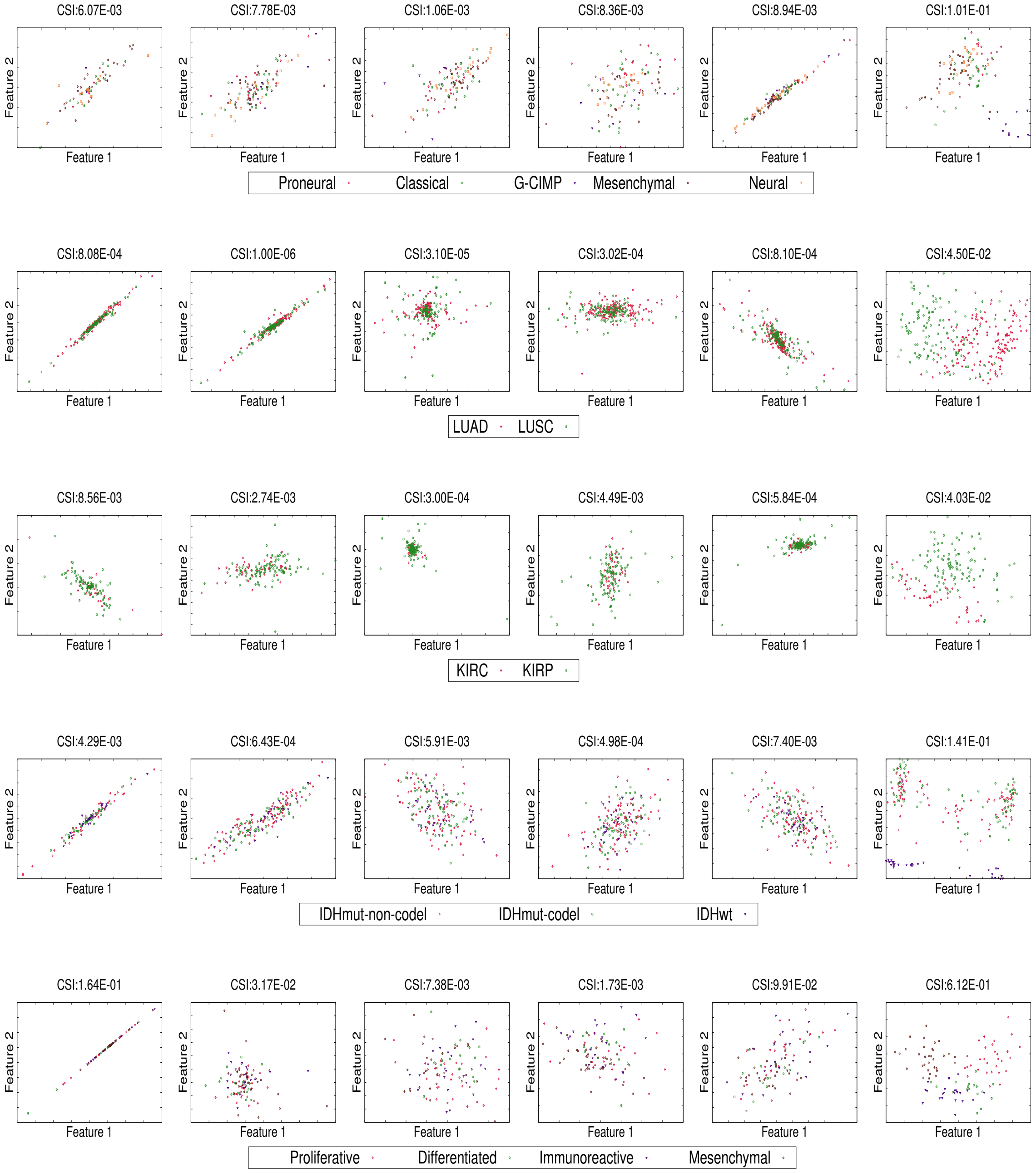
MCCA is higher compared to that of the proposed ReDMiCA algorithm. It shows that the proposed algorithm is able to separate different classes of LUNG and KIDNEY using the first two extracted features only. For the LGG data set, the proposed algorithm isolates almost all the samples of one class properly, but there is an overlap between the samples



(a) SUMCOR (b) GENVAR (c) MAXVAR (d) MINVAR (e) SSQCOR (f) ReDMiCA

Figure 5.5: Scatter plots for different criteria of the MCCA and proposed (ReDMiCA) algorithm on benchmark data sets, along with class separability index (top to bottom: CiteSeer, Handwritten, NW-OBJECT, Reuters, Caltech), each Oi denotes the i-th object class.

of the other two classes. Moreover, almost all the samples of four classes of OV data are well segregated using the proposed data integration algorithm. On the other hand, various



(a) SUMCOR (b) GENVAR (c) MAXVAR (d) MINVAR (e) SSQCOR (f) ReDMiCA

Figure 5.6: Scatter plots for different criteria of the MCCA and proposed (ReDMiCA) algorithm on omics data sets, along with class separability index (top to bottom: GBM, LUNG, KIDNEY, LGG, OV).

criteria of the MCCA cannot separate the classes properly using the first two extracted features.

Moreover, all five criteria of MCCA cannot handle the 'large p and small n' issue of mul-

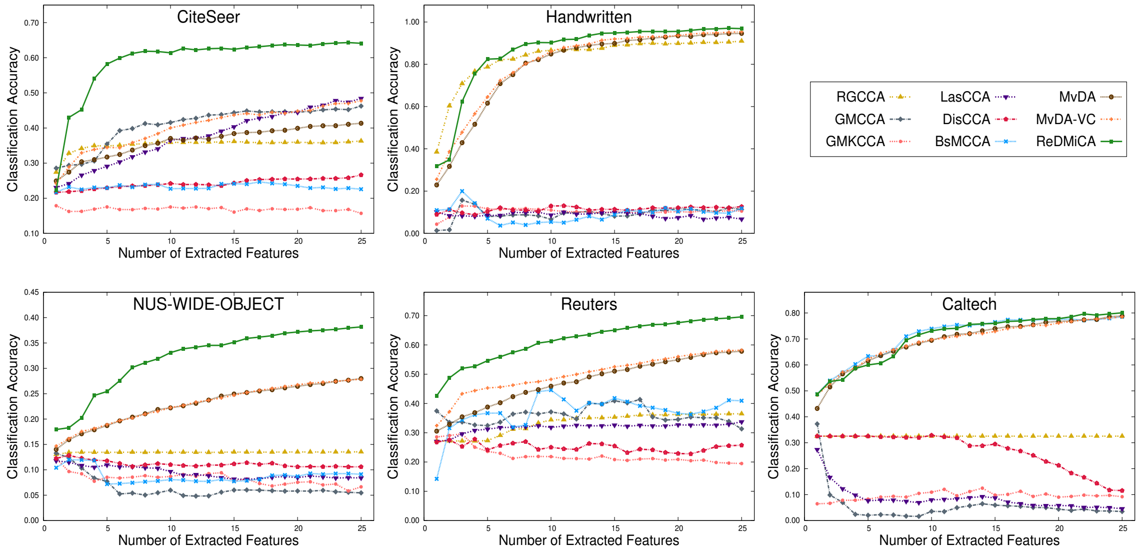


Figure 5.7: Variation of classification accuracy with respect to number of extracted features for different existing algorithms and proposed (ReDMiCA) algorithm on benchmark data sets for 10-fold CV.

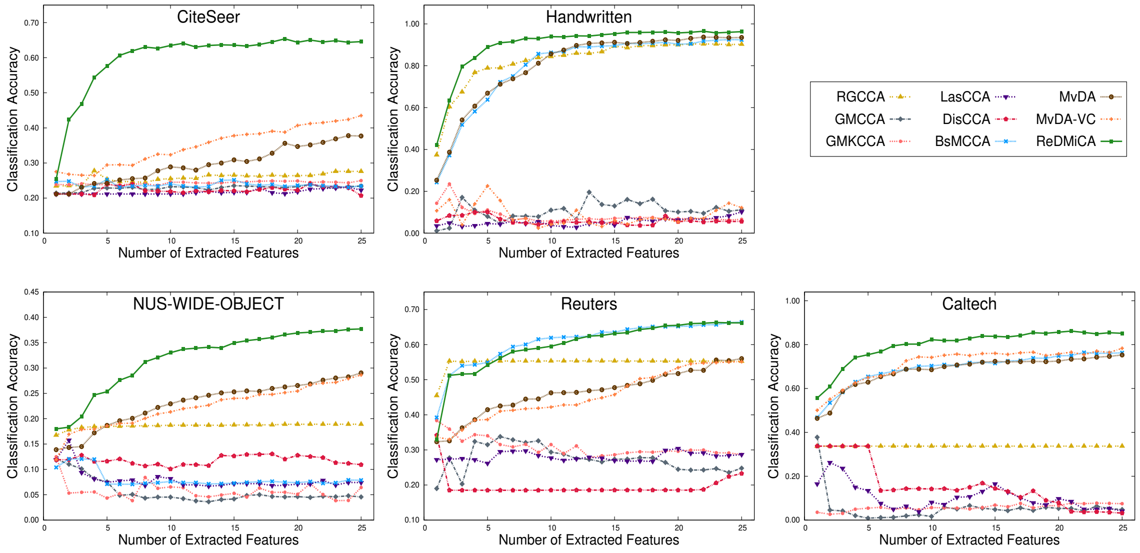


Figure 5.8: Variation of classification accuracy with respect to number of extracted features for different existing algorithms and proposed (ReDMiCA) algorithm on benchmark data sets for training-testing.

tidimensional data sets. On the other hand, the proposed ReDMiCA algorithm addresses this issue by using ridge regression optimization. Also, the significantly better performance of ReDMiCA is obtained due to the consideration of the supervised information of sample categories.

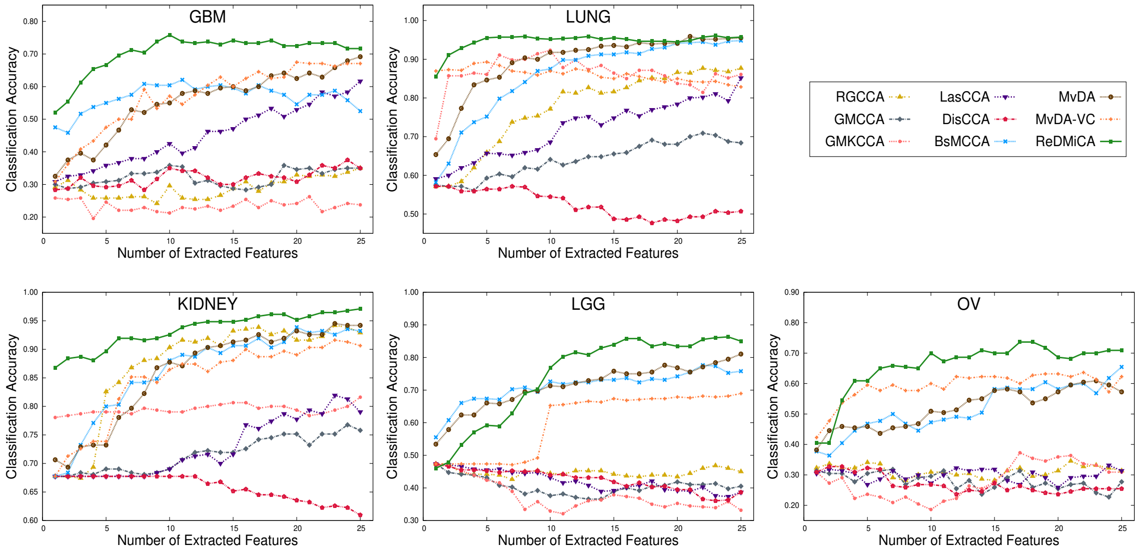


Figure 5.9: Variation of classification accuracy with respect to number of extracted features for different existing algorithms and proposed (ReDMiCA) algorithm on omics data sets for 10-fold CV.

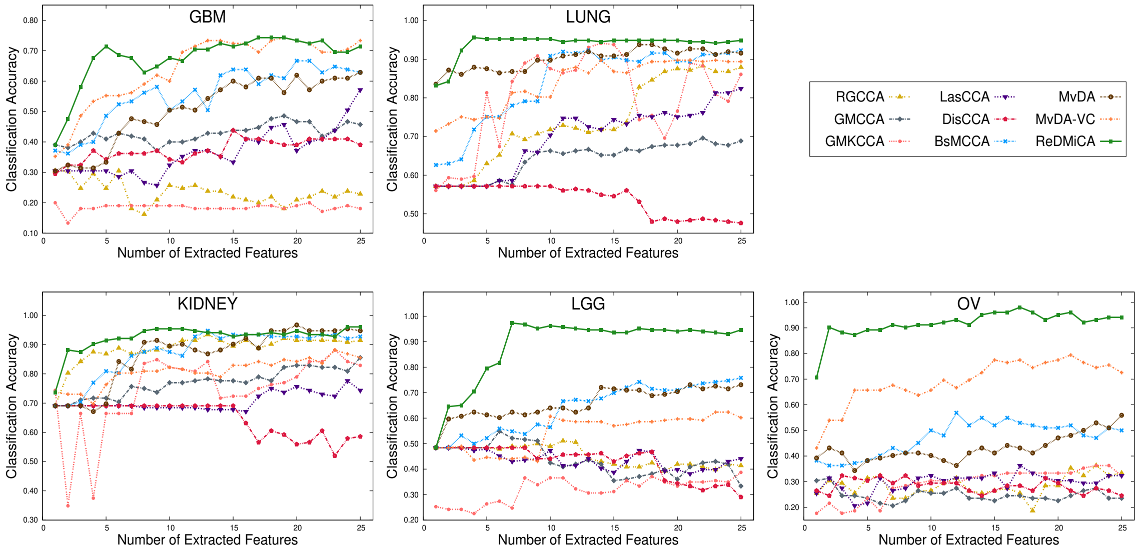
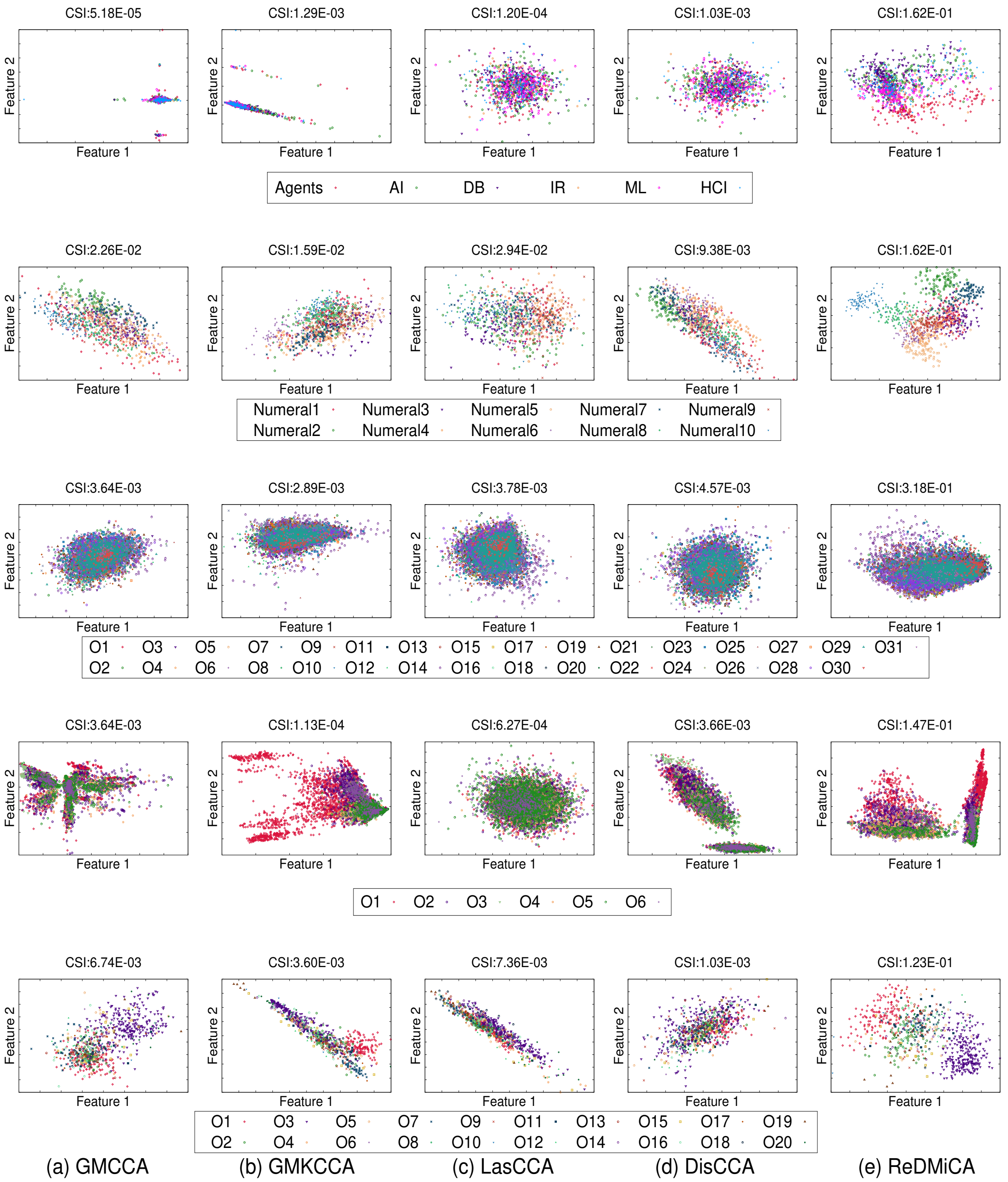


Figure 5.10: Variation of classification accuracy with respect to number of extracted features for different existing algorithms and proposed (ReDMiCA) algorithm on omics data sets for training-testing.

5.4.2 Comparative Performance Analysis

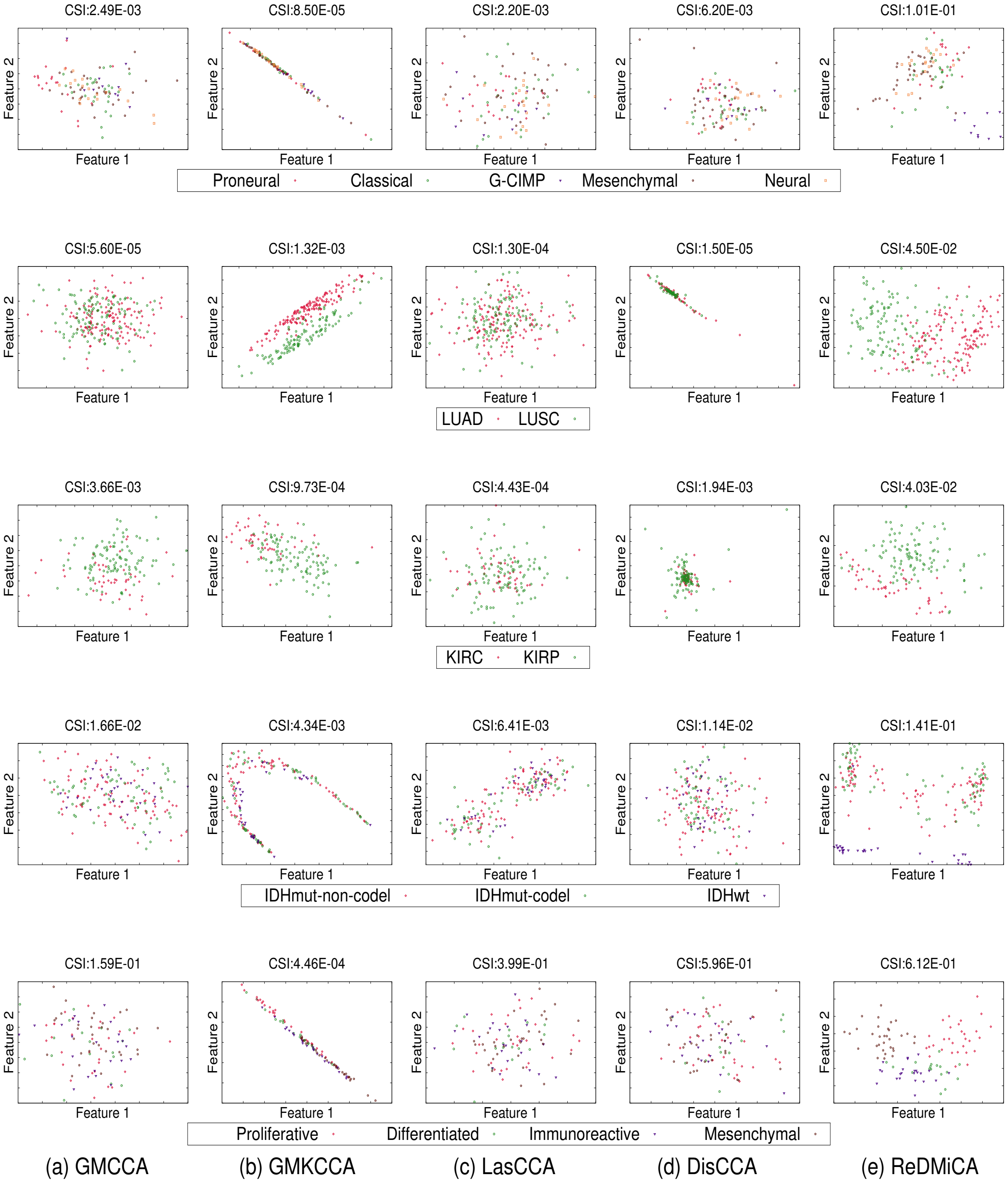
Finally, Figure 5.7, Figure 5.8, Figure 5.9, and Figure 5.10; Table 5.3, Table 5.4, Table 5.5, Table 5.6, Table 5.7, Table 5.8, Table 5.9, Table 5.11, and Table 5.10 analyze the performance of the proposed multimodal data integration algorithm with that of various state-



(a) GMCCA (b) GMKCCA (c) LasCCA (d) DisCCA (e) ReDMiCA

Figure 5.11: Scatter plots for state-of-the-art multimodal data integration algorithms (GMCCA, GMKCCA, LasCCA, and DisCCA) and proposed (ReDMiCA) algorithm on benchmark data sets, along with class separability index (top to bottom: CiteSeer, Handwritten, NW-OBJECT, Reuters, Caltech), each Oi denotes the i-th object class.

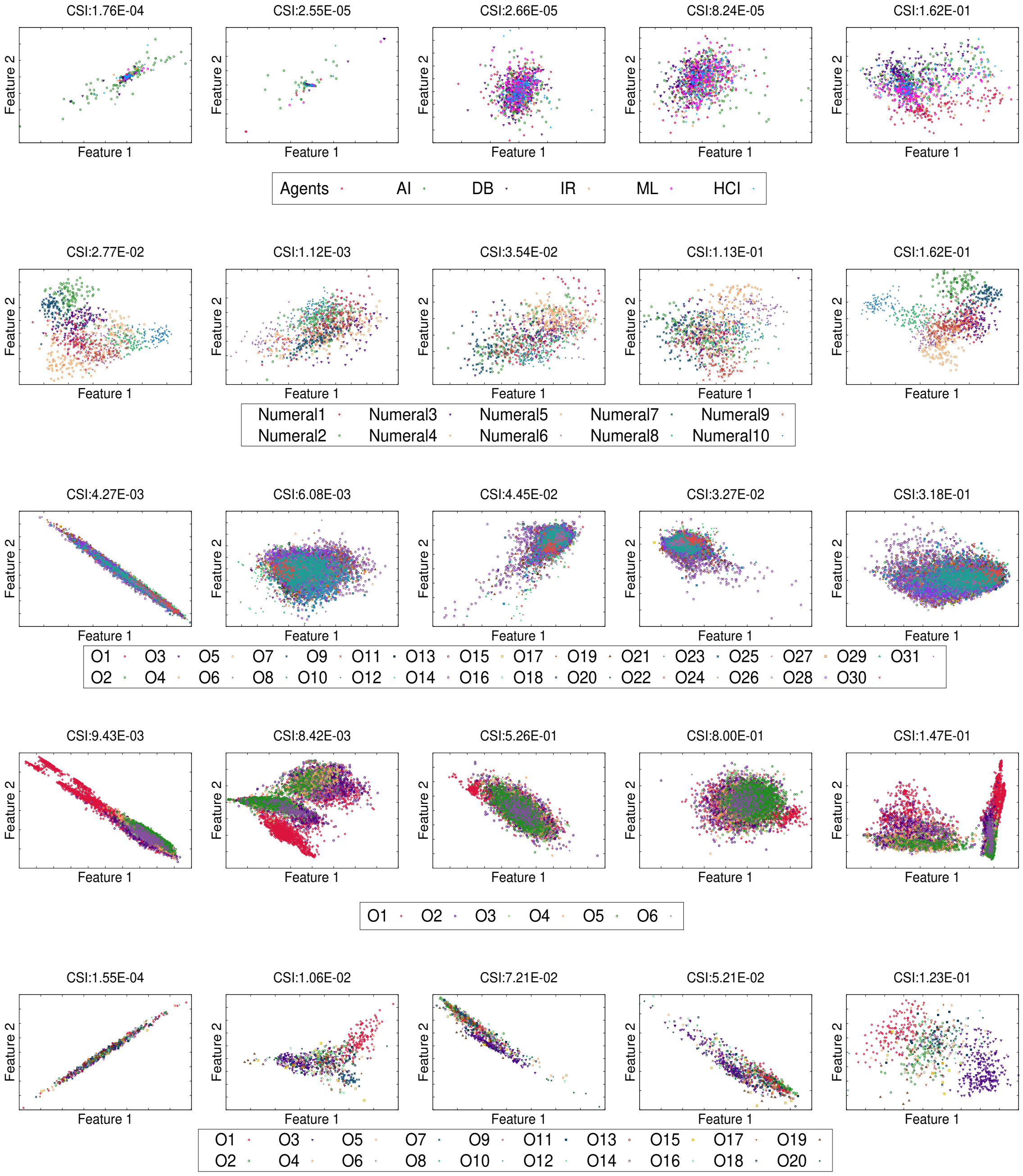
of-the-art MCCA based methods, namely, RGCCA [262], GMCCA [43], GMKCCA [43], large-scale generalized CCA (LasCCA) [84], distributed generalized CCA (DisCCA) [84], and block sparse MCCA (BsMCCA) [235]; two popular multidimensional data integration



(a) GMCCA (b) GMKCCA (c) LasCCA (d) DisCCA (e) ReDMiCA

Figure 5.12: Scatter plots for state-of-the-art multimodal data integration algorithms (GMCCA, GMKCCA, LasCCA, and DisCCA) and proposed (ReDMiCA) algorithm on omics data sets, along with class separability index (top to bottom: GBM, LUNG, KIDNEY, LGG, OV).

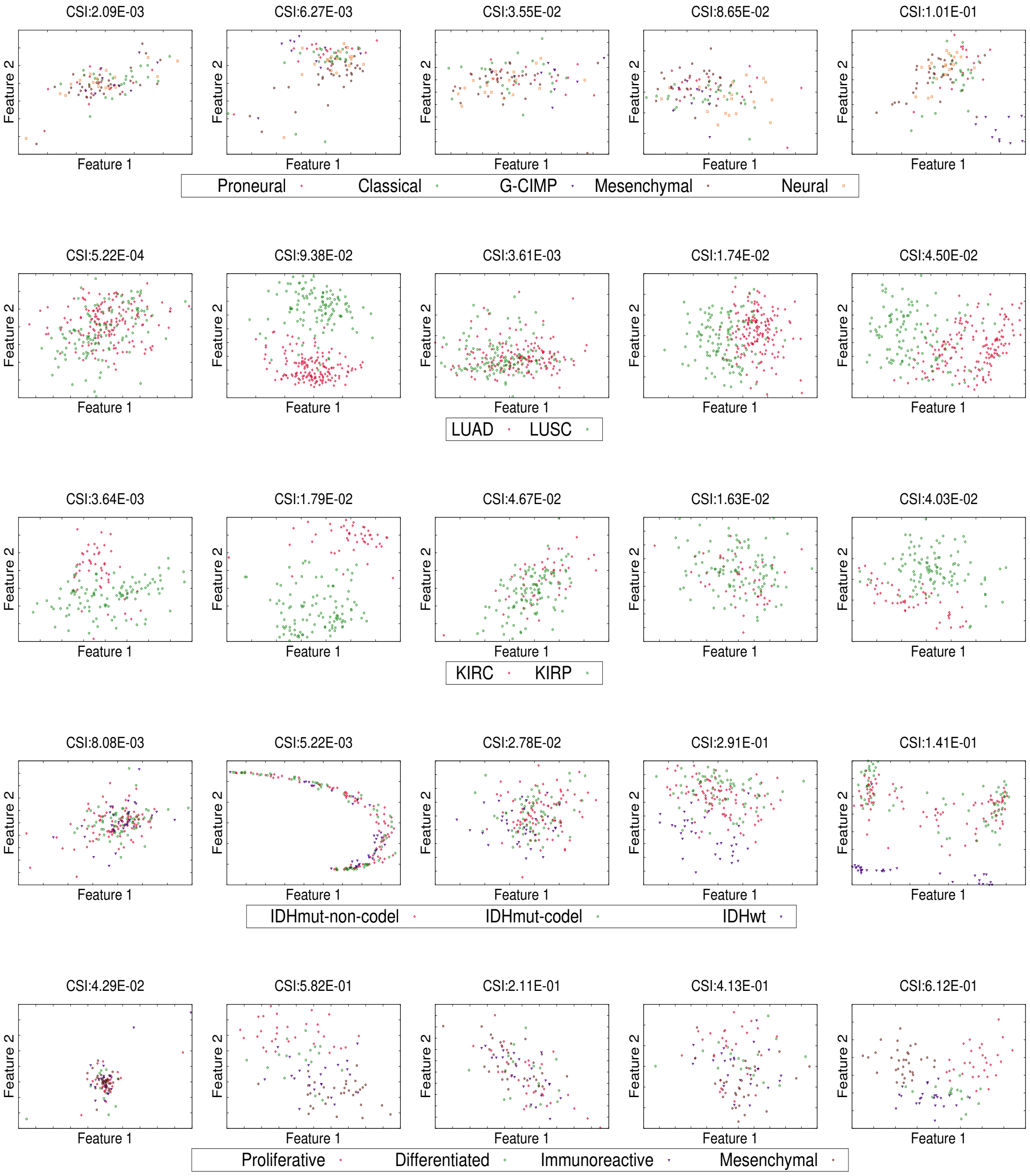
algorithms, namely, multi-view discriminant analysis (MvDA) [128] and multi-view discriminant analysis with view-consistency (MvDA-VC) [129]; and three deep learning-based algorithms, namely, deep multiset canonical correlation analysis (dMCCA) [244], task-optimal CCA (TOCCA) [55], and multimodal deep Boltzmann machines (MDBM) [247].



(a) RGCCA (b) BsMCCA (c) MvDA (d) MvDA-VC (e) ReDMiCA

Figure 5.13: Scatter plots for state-of-the-art multimodal data integration algorithms (RGCCA, BsMCCA, MvDA, and MvDA-VC) and proposed (ReDMiCA) algorithm on benchmark data sets, along with class separability index (top to bottom: CiteSeer, Handwritten, NW-OBJECT, Reuters, Caltech), each Oi denotes the i-th object class.

On the other hand, Figure 5.11, Figure 5.12, Figure 5.13, and Figure 5.14 show the scatter plots, along with the CSI using the first two extracted features of aforementioned algorithms on each data sets.



(a) RGCCA (b) BsMCCA (c) MvDA (d) MvDA-VC (e) ReDMiCA

Figure 5.14: Scatter plots for state-of-the-art multimodal data integration algorithms (RGCCA, BsMCCA, MvDA, and MvDA-VC) and proposed (ReDMiCA) algorithm on omics data sets, along with class separability index (top to bottom: GBM, LUNG, KIDNEY, LGG, OV).

5.4.2.1 MCCA Based Methods

Figure 5.7, Figure 5.8, Figure 5.9, and Figure 5.10; Table 5.3, Table 5.4, Table 5.5, Table 5.6, Table 5.7, and Table 5.8 demonstrate that the accuracy of the proposed ReDMiCA multimodal data integration algorithm is significantly higher as compared to that of existing MCCA based methods on both omics and benchmark data sets. All the results reported in Table 5.3, Table 5.4, and Table 5.5 confirm that the proposed algorithm attains the highest mean and median accuracy, irrespective of the data sets. Out of total 180 cases, the proposed algorithm attains significantly better (marked in bold) p-values than existing MCCA based methods in 177 cases, and better but not significant (marked in italics) p-values in 3 cases. From Figure 5.11, Figure 5.12 and the first two columns of Figure 5.13, Figure 5.14 it can be seen that the separation among various classes using the first two extracted features of the proposed algorithm is significantly better than that of the existing algorithms on omics and benchmark data sets.

Table 5.9: Classification Accuracy and Execution Time of Different Deep Learning Algorithms on Omics Data Sets

Data Sets	Different Algorithms	Accuracy (Train-Test)	Accuracy for 10-Fold CV Mean	Median	StdDev	Time (in sec.)
GBM	dMCCA	0.448	0.440	0.440	0.138	245.9
	TOCCA	0.381	0.406	0.367	0.116	235.6
	MDBM	0.581	0.435	0.419	0.134	13309.0
	ReDMiCA	**0.714**	**0.717**	**0.729**	0.047	2150.7
LUNG	dMCCA	0.593	0.607	0.589	0.051	11473.8
	TOCCA	0.571	0.571	0.571	0.000	940.2
	MDBM	0.879	0.613	0.429	0.238	9022.3
	ReDMiCA	**0.949**	**0.957**	**0.955**	0.032	6162.4
KIDNEY	dMCCA	0.862	0.690	0.677	0.041	44408.4
	TOCCA	0.684	0.713	0.677	0.070	585.5
	MDBM	0.691	0.710	0.677	0.102	13957.0
	ReDMiCA	**0.961**	**0.971**	**0.968**	0.028	6774.5
LGG	dMCCA	0.624	0.508	0.474	0.072	61032.4
	TOCCA	0.457	0.450	0.461	0.064	675.7
	MDBM	0.651	0.276	0.184	0.149	24204.6
	ReDMiCA	**0.946**	**0.850**	**0.842**	0.035	5958.2
OV	dMCCA	0.343	0.377	0.358	0.090	59090.0
	TOCCA	0.275	0.434	0.374	0.130	2213.1
	MDBM	0.373	0.445	0.418	0.127	31830.1
	ReDMiCA	**0.941**	**0.709**	**0.727**	0.075	8641.2

5.4.2.2 Others Multi-View Learning Algorithms

From Figure 5.7, Figure 5.8, Figure 5.9, and Figure 5.10, it is seen that the classification accuracy of the proposed algorithm is significantly higher, irrespective of the number of extracted features, as compared to that of both MvDA and MvDA-VC on five benchmark data, and GBM, LUNG, KIDNEY and OV data sets. In case of LGG data set, MvDA and MvDA-VC provide higher mean accuracy for 10-fold cross-validation than the proposed algorithm for lower number ($\leqslant$8) of extracted features. But, for higher number ($>$8) of features, the performance of the proposed algorithm improves drastically. As shown in Table 5.6, Table 5.7, and Table 5.8, out of the total 60 cases, the proposed algorithm attains significantly better (marked in bold) p-values than others two data integration methods in 34 cases, and better but not significant (marked in italics) p-values in 23 cases. The proposed algorithm is not significantly better than MvDA according to Friedman test on the GBM data set and MvDA-VC according to Friedman test on the Caltech and LUNG data sets. From the last three columns of Figure 5.13 and Figure 5.14, it is evident that different classes are remarkably separable using the first two extracted features of the proposed algorithm than these two existing multi-view learning algorithms on omics as well as benchmark data sets.

Table 5.10: Classification Accuracy and Execution Time of Different Deep Learning Algorithms on Benchmark Data Sets

Different Data Sets	dMCCA		TOCCA		MDBM		ReDMiCA	
	Accuracy	Time (sec)	Accuracy	Time (sec)	Accuracy	Time (sec)	Accuracy	Time (sec)
CiteSeer	0.212	21731.3	0.396	72415.7	0.178	54114.5	**0.646**	447.6
Handwritten	0.100	39127.7	**0.965**	1439.9	0.100	1812.7	0.963	1615.0
NW-OBJECT	0.178	250790.4	0.336	650417.0	0.261	1873753.2	**0.377**	1344.7
Reuters	0.517	109004.2	0.574	537830.2	0.468	1296693.8	**0.662**	11434.1
Caltech	0.337	1267362.4	0.802	58589.8	0.025	72458.5	**0.852**	8882.1

Table 5.11: Statistical Significance Analysis of Different Deep Learning Algorithms on Omics Data Sets

Data Sets	Different Algorithms	p-values for 10-Fold CV		
		Paired-t	Wilcoxon	Friedman
GBM	dMCCA	**3.50E-05**	**2.52E-03**	**1.57E-03**
	TOCCA	**4.40E-06**	**2.53E-03**	**1.57E-03**
	MDBM	**9.05E-05**	**2.53E-03**	**1.57E-03**
LUNG	dMCCA	**3.48E-10**	**2.46E-03**	**1.57E-03**
	TOCCA	**1.35E-11**	**2.42E-03**	**1.57E-03**
	MDBM	**5.15E-04**	**2.49E-03**	**1.57E-03**
KIDNEY	dMCCA	**1.97E-09**	**2.36E-03**	**1.57E-03**
	TOCCA	**2.46E-07**	**2.36E-03**	**1.57E-03**
	MDBM	**1.46E-05**	**3.19E-03**	**1.14E-02**
LGG	dMCCA	**1.94E-07**	**2.38E-03**	**1.57E-03**
	TOCCA	**1.10E-08**	**2.49E-03**	**1.57E-03**
	MDBM	**5.40E-07**	**2.45E-03**	**1.57E-03**
OV	dMCCA	**2.28E-06**	**2.53E-03**	**1.57E-03**
	TOCCA	**1.06E-04**	**3.46E-03**	**1.14E-02**
	MDBM	**1.20E-04**	**2.53E-03**	**1.57E-03**

5.4.2.3 Deep Learning-Based Methods

Finally, Table 5.9 and Table 5.10 compares the classification accuracy and execution time of the proposed algorithm with that of three deep learning-based methods on each data sets. The results presented in Table 5.9 and Table 5.10 demonstrate that the classification accuracy of the proposed algorithm is significantly higher as compared to that of various deep learning-based methods on all data sets, except Handwritten data, while its execution time is significantly lower. The TOCCA achieves 96.5% accuracy on Handwritten data set, whereas the proposed algorithm attains 96.3% accuracy. Although TOCCA performs well on benchmark data sets, it fails to achieve judicious results on omics data sets. The MDBM and dMCCA obtain 87.9% and 86.2% accuracy on LUNG and KIDNEY data sets, respectively, whereas both of them perform moderately on the GBM and LGG data sets. On the other hand, none of the deep learning-based methods performs better on the OV data set. Both MDBM and dMCCA provide poor performance on Handwritten, Caltech, and NW-OBJECT data sets due to the over training of these models. The results reported in Table 5.11 establish that the proposed algorithm attains significantly better p-values than the three deep learning-based methods, irrespective of the significance analysis and omics data sets used. All the results, reported here, establish the effectiveness of the proposed multiblock data integration algorithm over state-of-the-art approaches. The sequential extraction of relevant features from multiblock data enables the proposed algorithm to perform significantly better than existing methods.

5.5 Conclusion

A novel supervised sequential feature extraction algorithm has been proposed in this chapter. It integrates multimodal multidimensional data sets by solving the maximal correlation problem. A new block matrix representation has been introduced to determine the basis vectors of the MCCA. The proposed algorithm has addressed the 'large p and small n' issue of real-world multi-view data sets by using the ridge regression optimization technique, where regularization parameters have been varied within a certain range which helped to increase the search space of finding canonical variables. A theoretical analysis has been reported to manifest the connection between all canonical variables for each regularization parameter, which reduced the computational complexity as well as helped to generate correlated features sequentially. The proposed algorithm computes the canonical variable for a single modality having the lowest dimension with initial regularization parameter, and this canonical variable can be used to compute the canonical variables of all other modalities at different regularization parameter combinations.

The optimum values of regularization parameters have been estimated by computing the relevance and significance of the corresponding feature. To consider the supervised information, both relevance and significance measures have been computed based on the concept of the rough hypercuboid approach. The proposed algorithm can extract the desired number of relevant and significant features sequentially, without producing the complete set of possible features. The effectiveness of the proposed algorithm, along with a comparison with other algorithms, has been demonstrated on several omics and benchmark data sets.

One of the major problems in real-life multiblock dynamic data analysis is that all the modalities may not be available initially. The databases are generally updated incrementally. New modalities may be added to the existing modalities. So, it is necessary to develop a model that can generate the new feature from that of the existing modalities and the new modality without repeating the same procedure with the original data augmented by the new modality. In this regard, a new MCCA, termed as incremental MCCA (IMCCA) is proposed in the next chapter.

Chapter 6

Adaptive Generalized Multiset CCA for Incrementally Updated Multiblock Data

6.1 Introduction

A wide variety of applications from the brain-computer interface [161] to imaging genomics [155] have used multiset canonical correlation analysis (MCCA) for feature extraction. These applications involve either non-stationary or big data sets. The databases are generally updated continuously. They can be incremented in many ways. New instances may be added with the existing samples or new modalities may be considered for better analysis. For example, TCGA updates and releases the new data, both samples and modalities, twenty-two times in the last five years. Every day, a huge amount of data is being added to the existing databases. In such contexts, the algorithms for solving canonical correlation analysis (CCA) should be adaptive or incremental in nature. Incremental learning is a machine learning paradigm where the learning process takes place whenever new data is merged with or deleted from the existing data set and the solutions already obtained are only modified. In [310], an adaptive CCA based on matrix manifold has been presented, while a learning algorithm has been reported in [277] for adaptive CCA of several data sets. Both of them are applicable to the situation when new samples are being added with the existing samples and all the covariance matrices are required to update. However, they are not applicable when a new modality is available for the augmentation with the existing modalities. In [114], a one-pass learning approach has been introduced to address the problem of learning associated with incremental and decremental features. Recently, a safe classification approach has been proposed for exploiting augmented features or views [113]. However, none of these methods [113, 114] considers the covariance of individual modality as well as cross-covariance among different modalities. In [187], an incremental generalized CCA has been proposed for incremental updates of existing solutions based on new modalities, although it leads to approximate solutions. Moreover, all the adaptive CCA algorithms reported in [187, 277, 310] are unsupervised in nature.

In this regard, a new MCCA, termed as incremental MCCA (IMCCA), is introduced to integrate judiciously the information of sequentially arriving multidimensional variables. The proposed IMCCA incrementally updates the existing solutions, whenever a new modality is available for the analysis. The theoretical analysis presented in this chapter establishes that, unlike [187], the proposed model generates the exact solutions while updating canonical variables incrementally. The proposed IMCCA deals with the "large p-small n" problem of multidimensional data sets by using the ridge regression optimization technique. A theoretical analysis is presented, which helps to compute multiset canonical variables under ridge regression in an iterative way. Using the proposed IMCCA model, a new feature extraction algorithm, termed as SeFGeIM (Sequential Feature Generation using IMCCA), is introduced. The proposed SeFGeIM algorithm considers a new modality for the integration if it has relevant and significant information with respect to earlier modalities. It starts with the two most relevant modalities, and the remaining modalities are added sequentially according to their relevance. The optimum regularization parameters for the proposed algorithm are estimated based on the supervised information of sample categories. An analytical formulation enables the proposed algorithm generation of the required number of relevant and significant features from the multiblock dynamic data sets, without extracting all possible features. In fact, all the modalities may not be required to extract different features. The effectiveness of the proposed multiblock data integration approach, along with a comparative performance analysis with the state-of-the-art methods, is established on several real-life multiblock data. Some of the results of this chapter are reported in [184].

The rest of this chapter is organized as follows: Section 6.2 presents a new multiset CCA algorithm. In Section 6.3, a new feature extraction algorithm is presented for incrementally updated multiblock data. The effectiveness of the proposed multi-view data integration algorithm, along with a comparative performance analysis with state-of-the-art algorithms on different multi-view benchmark and omics data sets, is presented in Section 6.4. Concluding remarks are provided in Section 6.5.

6.2 Proposed Multiset CCA

This section presents a new multiset CCA, termed as incremental multiset CCA (IMCCA). It judiciously integrates the information of multidimensional multimodal data sets that are available sequentially one after another. When a new modality is available for the same set of samples, the proposed model generates a new set of features based on the new modality as well as the features extracted from the earlier modalities. It does not repeat the same procedure with the original data augmented by the new modality. Some important analytical formulations are reported next to explain the proposed multiset CCA model.

6.2.1 IMCCA: Incremental Multiset CCA

Let $\mathcal{X}_1 \in \Re^{m_1 \times n}$,$\mathcal{X}_2 \in \Re^{m_2 \times n}$,$\cdots$,$\mathcal{X}_\mathcal{M} \in \Re^{m_\mathcal{M} \times n}$ be $\mathcal{M}$ multidimensional data sets with $m_1, m_2, \cdots, m_\mathcal{M}$ variables, respectively, and n represents the number of common samples.

Let us consider that

$$\mathcal{A}_{\mathcal{M}} = \begin{bmatrix} \mathbf{0}^{[m_1]} & a_{12} & \cdots & a_{1\mathcal{M}} \\ a_{21} & \mathbf{0}^{[m_2]} & \cdots & a_{2\mathcal{M}} \\ \vdots & \vdots & \ddots & \vdots \\ a_{\mathcal{M}1} & a_{\mathcal{M}2} & \cdots & \mathbf{0}^{[m_{\mathcal{M}}]} \end{bmatrix};$$

where $\mathcal{A}_{\mathcal{M}} \in \Re^{\sum_{i=1}^{\mathcal{M}} m_i \times \sum_{i=1}^{\mathcal{M}} m_i}$ and $a_{ij} = \mathcal{C}_{ii}{}^{-1}\mathcal{C}_{ij}$; $\forall i \in \{1, 2, \cdots, \mathcal{M}\}$, $\forall j \in \{1, 2, \cdots, \mathcal{M}\}$, $i \neq j$ and $\mathbf{0}^{[k]}$ denotes a square null matrix with dimension k. The eigenvectors $\left[w_1^{\mathcal{M}}, w_2^{\mathcal{M}}, \cdots, w_{\mathcal{M}}^{\mathcal{M}}\right]^T$ of $\mathcal{A}_{\mathcal{M}}$ are the basis vectors of $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_{\mathcal{M}}$, that is,

$$\begin{bmatrix} \mathbf{0}^{[m_1]} & a_{12} & \cdots & a_{1\mathcal{M}} \\ a_{21} & \mathbf{0}^{[m_2]} & \cdots & a_{2\mathcal{M}} \\ \vdots & \vdots & \ddots & \vdots \\ a_{\mathcal{M}1} & a_{\mathcal{M}2} & \cdots & \mathbf{0}^{[m_{\mathcal{M}}]} \end{bmatrix} \begin{bmatrix} w_1^{\mathcal{M}} \\ w_2^{\mathcal{M}} \\ \vdots \\ w_{\mathcal{M}}^{\mathcal{M}} \end{bmatrix} = \rho^{\mathcal{M}} \begin{bmatrix} w_1^{\mathcal{M}} \\ w_2^{\mathcal{M}} \\ \vdots \\ w_{\mathcal{M}}^{\mathcal{M}} \end{bmatrix}. \tag{6.1}$$

Here, superscript $\mathcal{M}$ of $w_i^{\mathcal{M}}$ denotes that all $\mathcal{M}$ multidimensional variables are considered for the computation of the basis vector of i-th multidimensional variable. Now, a new multidimensional variable $\mathcal{X}_{(\mathcal{M}+1)} \in \Re^{m_{(\mathcal{M}+1)} \times n}$ is available. The basis vector $w_{(\mathcal{M}+1)}^{\mathcal{M}+1}$ corresponding to $\mathcal{X}_{(\mathcal{M}+1)}$ has to be computed as well as all the basis vectors $w_1^{\mathcal{M}}, w_2^{\mathcal{M}}, \cdots, w_{\mathcal{M}}^{\mathcal{M}}$ of existing $\mathcal{M}$ modalities have to be updated. Let us consider that the updated basis vectors are $w_1^{\mathcal{M}+1}, w_2^{\mathcal{M}+1}, \cdots, w_{\mathcal{M}}{}^{\mathcal{M}+1}$, which have to be modified by using the previous basis vectors $w_1^{\mathcal{M}}, w_2^{\mathcal{M}}, \cdots, w_{\mathcal{M}}^{\mathcal{M}}$ and the basis vector $w_{(\mathcal{M}+1)}^{\mathcal{M}+1}$ corresponding to $\mathcal{X}_{(\mathcal{M}+1)}$. As the number of multidimensional variables is $(\mathcal{M}+1)$, the matrix $\mathcal{A}_{\mathcal{M}}$ of (5.3.1) in Chapter 5 becomes $\mathcal{A}_{(\mathcal{M}+1)}$, where

$$\mathcal{A}_{(\mathcal{M}+1)} = \begin{bmatrix} \begin{bmatrix} \mathbf{0}^{[m_1]} & a_{12} & \cdots & a_{1\mathcal{M}} \\ a_{21} & \mathbf{0}^{[m_2]} & \cdots & a_{2\mathcal{M}} \\ \vdots & \vdots & \ddots & \vdots \\ a_{\mathcal{M}1} & a_{\mathcal{M}2} & \cdots & \mathbf{0}^{[m_{\mathcal{M}}]} \end{bmatrix} & \begin{bmatrix} a_{1(\mathcal{M}+1)} \\ a_{2(\mathcal{M}+1)} \\ \vdots \\ a_{\mathcal{M}(\mathcal{M}+1)} \end{bmatrix} \\ \begin{bmatrix} a_{(\mathcal{M}+1)1} & a_{(\mathcal{M}+1)2} & \cdots & a_{(\mathcal{M}+1)\mathcal{M}} \end{bmatrix} & \mathbf{0}^{[m_{(\mathcal{M}+1)}]} \end{bmatrix}$$

$$= \begin{bmatrix} \mathcal{A}_{\mathcal{M}} & \Theta_{\mathcal{M}} \\ \Phi_{\mathcal{M}} & \mathbf{0}^{[m_{(\mathcal{M}+1)}]} \end{bmatrix} = \begin{bmatrix} \mathcal{A}_{\mathcal{M}} + [\Theta_{\mathcal{M}}\Phi_{\mathcal{M}}]^{1/2} & \mathbf{0}^{[\sum_{i=1}^{\mathcal{M}} m_i, m_{(\mathcal{M}+1)}]} \\ \mathbf{0}^{[m_{(\mathcal{M}+1)}, \sum_{i=1}^{\mathcal{M}} m_i]} & [\Phi_{\mathcal{M}}\Theta_{\mathcal{M}}]^{1/2} \end{bmatrix}; \tag{6.2}$$

$$\text{where} \quad \Theta_{\mathcal{M}} = \begin{bmatrix} a_{1(\mathcal{M}+1)} \\ a_{2(\mathcal{M}+1)} \\ \cdots \\ a_{\mathcal{M}(\mathcal{M}+1)} \end{bmatrix};$$

$$\text{and} \qquad \Phi_{\mathcal{M}} = \begin{bmatrix} a_{(\mathcal{M}+1)1} & a_{(\mathcal{M}+1)2} & \cdots & a_{(\mathcal{M}+1)\mathcal{M}} \end{bmatrix};$$

$\mathbf{0}^{[k,\ell]}$ denotes a rectangular null matrix with dimension $k \times \ell$. According to (6.1), the eigenvectors of $\mathcal{A}_{(\mathcal{M}+1)}$ are the basis vectors $w_1^{\mathcal{M}+1}, w_2^{\mathcal{M}+1}, \cdots, w_{(\mathcal{M}+1)}^{\mathcal{M}+1}$ of all $(\mathcal{M}+1)$ multidimensional variables, that is,

$$\begin{bmatrix} \mathcal{A}_{\mathcal{M}} + [\Theta_{\mathcal{M}}\Phi_{\mathcal{M}}]^{1/2} & \mathbf{0}^{[\sum\limits_{i=1}^{\mathcal{M}} m_i, m_{(\mathcal{M}+1)}]} \\ \mathbf{0}^{[m_{(\mathcal{M}+1)}, \sum\limits_{i=1}^{\mathcal{M}} m_i]} & [\Phi_{\mathcal{M}}\Theta_{\mathcal{M}}]^{1/2} \end{bmatrix} \begin{bmatrix} \mathcal{W}_{\mathcal{M}}^{\mathcal{M}+1} \\ w_{(\mathcal{M}+1)}^{\mathcal{M}+1} \end{bmatrix} = \rho^{[\mathcal{M}+1]} \begin{bmatrix} \mathcal{W}_{\mathcal{M}}^{\mathcal{M}+1} \\ w_{(\mathcal{M}+1)}^{\mathcal{M}+1} \end{bmatrix} \tag{6.3}$$

$$\text{where} \quad \mathcal{W}_{\mathcal{M}}^{\mathcal{M}+1} = \begin{bmatrix} w_1^{\mathcal{M}+1} \\ w_2^{\mathcal{M}+1} \\ \cdots \\ w_{\mathcal{M}}^{\mathcal{M}+1} \end{bmatrix}. \tag{6.4}$$

From (6.3), it is evident that the eigenvectors of the matrices $\left[\mathcal{A}_{\mathcal{M}} + [\Theta_{\mathcal{M}}\Phi_{\mathcal{M}}]^{1/2}\right]$ and $[\Phi_{\mathcal{M}}\Theta_{\mathcal{M}}]^{1/2}$ are $\mathcal{W}_{\mathcal{M}}^{\mathcal{M}+1}$ and $w_{(\mathcal{M}+1)}^{\mathcal{M}+1}$, respectively, with corresponding eigenvalue $\rho^{\mathcal{M}+1}$, that is,

$$\left[\mathcal{A}_{\mathcal{M}} + [\Theta_{\mathcal{M}}\Phi_{\mathcal{M}}]^{1/2}\right] \mathcal{W}_{\mathcal{M}}^{\mathcal{M}+1} = \rho^{\mathcal{M}+1} \mathcal{W}_{\mathcal{M}}^{\mathcal{M}+1}; \tag{6.5}$$

$$\text{and} \qquad [\Phi_{\mathcal{M}}\Theta_{\mathcal{M}}]^{1/2} w_{(\mathcal{M}+1)}^{\mathcal{M}+1} = \rho^{\mathcal{M}+1} w_{(\mathcal{M}+1)}^{\mathcal{M}+1}$$

$$\Rightarrow [\Phi_{\mathcal{M}}\Theta_{\mathcal{M}}] w_{(\mathcal{M}+1)}^{\mathcal{M}+1} = \left[\rho^{\mathcal{M}+1}\right]^2 w_{(\mathcal{M}+1)}^{\mathcal{M}+1}. \tag{6.6}$$

From (6.2.1), it is seen that the basis vector $w_{(\mathcal{M}+1)}^{\mathcal{M}+1}$ of new multidimensional variable $\mathcal{X}_{(\mathcal{M}+1)}$ can be computed by calculating the eigenvector of the matrix $[\Phi_{\mathcal{M}}\Theta_{\mathcal{M}}]$. Also,

$$\Phi_{\mathcal{M}}\Theta_{\mathcal{M}} w_{(\mathcal{M}+1)}^{\mathcal{M}+1} = \left[\rho^{\mathcal{M}+1}\right]^2 w_{(\mathcal{M}+1)}^{\mathcal{M}+1}$$

$$\Rightarrow \Theta_{\mathcal{M}}\Phi_{\mathcal{M}}\Theta_{\mathcal{M}} w_{(\mathcal{M}+1)}^{\mathcal{M}+1} = \left[\rho^{\mathcal{M}+1}\right]^2 \Theta_{\mathcal{M}} w_{(\mathcal{M}+1)}^{\mathcal{M}+1}$$

$$\Rightarrow [\Theta_{\mathcal{M}}\Phi_{\mathcal{M}}]^{1/2} \Theta_{\mathcal{M}} w_{(\mathcal{M}+1)}^{\mathcal{M}+1} = \rho^{\mathcal{M}+1} \Theta_{\mathcal{M}} w_{(\mathcal{M}+1)}^{\mathcal{M}+1}$$

$$\Rightarrow [\Theta_{\mathcal{M}}\Phi_{\mathcal{M}}]^{1/2} = \rho^{\mathcal{M}+1} \Theta_{\mathcal{M}} w_{(\mathcal{M}+1)}^{\mathcal{M}+1} \left[w_{(\mathcal{M}+1)}^{\mathcal{M}+1}\right]^T \Theta_{\mathcal{M}}^{\dagger}; \tag{6.7}$$

where $A^{\dagger}$ denotes the pseudoinverse of A.

When the new multidimensional variable $\mathcal{X}_{(\mathcal{M}+1)}$ is added with the existing variables, the previous basis vectors $w_1^{\mathcal{M}}, w_2^{\mathcal{M}}, \cdots, w_{\mathcal{M}}^{\mathcal{M}}$ of preceding multidimensional variables $\mathcal{X}_1, \mathcal{X}_2, \cdots,$ $\mathcal{X}_{\mathcal{M}}$ have to be updated by using $w_{(\mathcal{M}+1)}^{\mathcal{M}+1}$ and $w_1^{\mathcal{M}}, w_2^{\mathcal{M}}, \cdots, w_{\mathcal{M}}^{\mathcal{M}}$. From (6.5), it can be observed that the updated basis vectors $w_1^{\mathcal{M}+1}, w_2^{\mathcal{M}+1}, \cdots, w_{\mathcal{M}}^{\mathcal{M}+1}$ of preceding multidimensional variables $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_{\mathcal{M}}$ are the eigenvectors of the matrix $\left[\mathcal{A}_{\mathcal{M}} + [\Theta_{\mathcal{M}}\Phi_{\mathcal{M}}]^{1/2}\right]$. From

(6.1), we get

$$\mathcal{A}_{\mathcal{M}} \mathcal{W}^{\mathcal{M}}_{\mathcal{M}} = \rho^{\mathcal{M}} \mathcal{W}^{\mathcal{M}}_{\mathcal{M}} \Rightarrow \mathcal{A}_{\mathcal{M}} = \rho^{\mathcal{M}} \mathcal{W}^{\mathcal{M}}_{\mathcal{M}} \left[\mathcal{W}^{\mathcal{M}}_{\mathcal{M}}\right]^T . \tag{6.8}$$

Adding (6.2.1) and (6.8), we get

$$\mathcal{A}_{\mathcal{M}} + \left[\Theta_{\mathcal{M}} \Phi_{\mathcal{M}}\right]^{1/2} = \rho^{\mathcal{M}} \mathcal{W}^{\mathcal{M}}_{\mathcal{M}} \left[\mathcal{W}^{\mathcal{M}}_{\mathcal{M}}\right]^T + \rho^{\mathcal{M}+1} \Theta_{\mathcal{M}} w^{\mathcal{M}+1}_{(\mathcal{M}+1)} \left[w^{\mathcal{M}+1}_{(\mathcal{M}+1)}\right]^T \Theta^{\dagger}_{\mathcal{M}}. \tag{6.9}$$

Hence, from (6.5) and (6.9), it is observed that the eigenvector of the matrix $\left[\rho^{\mathcal{M}} \mathcal{W}^{\mathcal{M}}_{\mathcal{M}} \left[\mathcal{W}^{\mathcal{M}}_{\mathcal{M}}\right]^T + \rho^{\mathcal{M}+1} \Theta_{\mathcal{M}} w^{\mathcal{M}+1}_{(\mathcal{M}+1)} \left[w^{\mathcal{M}+1}_{(\mathcal{M}+1)}\right]^T \Theta^{\dagger}_{\mathcal{M}}\right]$ is the updated basis vectors $w^{\mathcal{M}+1}_1$, $w^{\mathcal{M}+1}_2, \cdots, w^{\mathcal{M}+1}_{\mathcal{M}}$, which can be computed by using the previous basis vectors $w^{\mathcal{M}}_1$, $w^{\mathcal{M}}_2, \cdots$, $w^{\mathcal{M}}_{\mathcal{M}}$ of the preceding $\mathcal{M}$ multidimensional variables $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_{\mathcal{M}}$ and the basis vector $w^{\mathcal{M}+1}_{(\mathcal{M}+1)}$ of the newly added multidimensional variable $\mathcal{X}_{(\mathcal{M}+1)}$. Hence, when a new multidimensional variable is added, there is no need to compute the extracted feature set from the initial stage. The new features can be computed by using preceding features.

6.2.2 Validation of Proposed IMCCA

The proposed model is designed in such a way that, when a new data $\mathcal{X}_{(\mathcal{M}+1)}$ arrives, the algorithm will not repeat the same steps with the original data $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_{\mathcal{M}}$ augmented by the new data $\mathcal{X}_{(\mathcal{M}+1)}$. Rather, it starts with the basis vectors obtained using the previous set of data, and generates the new basis vectors. In fact, if the initial set of modalities and the new modality come together, then the proposed method generates the same set of basis vectors. To establish this characteristic, let us consider that all $(\mathcal{M}+1)$ multidimensional data arrive simultaneously. Hence, the eigenvectors of $\mathcal{A}_{(\mathcal{M}+1)}$ are the basis vectors of $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_{(\mathcal{M}+1)}$, that is,

$$\mathcal{A}_{(\mathcal{M}+1)} \boldsymbol{\mathcal{W}^{\mathcal{M}+1}_{(\mathcal{M}+1)}} = \begin{bmatrix} \mathbf{0}^{[m_1]} & a_{12} & \cdots & a_{1(\mathcal{M}+1)} \\ a_{21} & \mathbf{0}^{[m_2]} & \cdots & a_{2(\mathcal{M}+1)} \\ \vdots & \vdots & \ddots & \vdots \\ a_{(\mathcal{M}+1)1} & a_{(\mathcal{M}+1)2} & \cdots & \mathbf{0}^{[m_{(\mathcal{M}+1)}]} \end{bmatrix} \begin{bmatrix} w^{\mathcal{M}+1}_1 \\ w^{\mathcal{M}+1}_2 \\ \vdots \\ w^{\mathcal{M}+1}_{(\mathcal{M}+1)} \end{bmatrix};$$

where the matrix $\boldsymbol{\mathcal{W}^{\mathcal{M}+1}_{(\mathcal{M}+1)}}$ has $p = \min\{m_1, m_2, \cdots, m_{(\mathcal{M}+1)}\}$ columns, the i-th column being the i-th basis vector $\left[\mathcal{W}^{\mathcal{M}+1}_{(\mathcal{M}+1)}\right]_i$,$\forall i \in \{1, 2, \cdots, p\}$, of each multidimensional variable.

Next, let us assume that initially $\mathcal{M}$ number of multidimensional variables $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_{\mathcal{M}}$ are available, and their corresponding basis vectors have already been computed. Now, a new multidimensional variable $\mathcal{X}_{(\mathcal{M}+1)}$ is added with the existing $\mathcal{M}$ variables. According to (6.2.1), the eigenvector of $[\Phi_{\mathcal{M}} \Theta_{\mathcal{M}}]$ is the basis vector $w^{\mathcal{M}+1}_{(\mathcal{M}+1)}$ of $\mathcal{X}_{(\mathcal{M}+1)}$. On the other hand, it can be observed from (6.5) and (6.9) that the basis vectors of preceding $\mathcal{M}$ multidimensional variables can be updated by using previously obtained basis vectors of them and the basis vector of the newly added multidimensional variable. Let the updated basis vectors be $\mathcal{W}^{\mathcal{M}+1}_{\mathcal{M}}$. Let us also assume that $\widehat{\boldsymbol{\mathcal{W}}}\boldsymbol{^{\mathcal{M}+1}_{(\mathcal{M}+1)}} = \begin{bmatrix} \mathcal{W}^{\mathcal{M}+1}_{\mathcal{M}} \\ w^{\mathcal{M}+1}_{(\mathcal{M}+1)} \end{bmatrix}$, where the matrix

$\widehat{\boldsymbol{\mathcal{W}}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}}$ has also $p = \min\{m_1, m_2, \cdots, m_{(\mathcal{M}+1)}\}$ columns, the i-th column is the i-th basis vector $\left[\widehat{\mathcal{W}}^{\mathcal{M}+1}_{(\mathcal{M}+1)}\right]_i$ of each multidimensional variable.

To establish the characteristics of the proposed method, we need to show that $\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}}$ and $\widehat{\boldsymbol{\mathcal{W}}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}}$ are same. That means, if we project the $(\mathcal{M}+1)$ multidimensional variables in $\left[\mathcal{W}^{\mathcal{M}+1}_{(\mathcal{M}+1)}\right]_i$ and $\left[\widehat{\mathcal{W}}^{\mathcal{M}+1}_{(\mathcal{M}+1)}\right]_i$ directions, the correlation between the projected subspaces $\boldsymbol{\mathcal{U}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}} = \sum\limits_{j=1}^{\mathcal{M}+1} \left[\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{j}}\right]^T \mathcal{X}_j$ and $\widehat{\boldsymbol{\mathcal{U}}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}} = \sum\limits_{j=1}^{\mathcal{M}+1} \left[\widehat{\boldsymbol{\mathcal{W}}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{j}}\right]^T \mathcal{X}_j$ will be maximum or the angle between these two subspaces is minimum, that is,

$$\min_{\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}}, \widehat{\boldsymbol{\mathcal{W}}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}}} \|\boldsymbol{\mathcal{U}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}} - \widehat{\boldsymbol{\mathcal{U}}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}}\|_F^2; \tag{6.10}$$

subject to

$$\boldsymbol{\mathcal{U}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}} \left[\boldsymbol{\mathcal{U}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}}\right]^T = \widehat{\boldsymbol{\mathcal{U}}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}} \left[\widehat{\boldsymbol{\mathcal{U}}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}}\right]^T = I;$$

where I is the identity matrix with appropriate order. Now,

$$\|\boldsymbol{\mathcal{U}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}} - \widehat{\boldsymbol{\mathcal{U}}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}}\|_F^2 = 2I - 2\boldsymbol{\mathcal{U}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}} \left[\widehat{\boldsymbol{\mathcal{U}}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}}\right]^T. \tag{6.11}$$

Hence, using (6.11), the objective function of (6.10) can be reformulated as

$$\max_{\boldsymbol{\mathcal{W}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}}, \widehat{\boldsymbol{\mathcal{W}}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}}} \text{trace}\left(\boldsymbol{\mathcal{U}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}} \left[\widehat{\boldsymbol{\mathcal{U}}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}}\right]^T\right); \tag{6.12}$$

subject to

$$\boldsymbol{\mathcal{U}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}} \left[\boldsymbol{\mathcal{U}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}}\right]^T = \widehat{\boldsymbol{\mathcal{U}}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}} \left[\widehat{\boldsymbol{\mathcal{U}}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}}\right]^T = I.$$

Now,

$$\boldsymbol{\mathcal{U}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}} \left[\boldsymbol{\mathcal{U}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}}\right]^T = I \Rightarrow \boldsymbol{\mathcal{U}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}} = \left[\left[\boldsymbol{\mathcal{U}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}}\right]^T\right]^{-1}$$

$$\Rightarrow \boldsymbol{\mathcal{U}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}} \left[\widehat{\boldsymbol{\mathcal{U}}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}}\right]^T = \left[\left[\boldsymbol{\mathcal{U}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}}\right]^T\right]^{-1} \left[\widehat{\boldsymbol{\mathcal{U}}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}}\right]^T. \tag{6.13}$$

Hence, using (6.2.2), it is possible to prove that (6.12) will be maximum when

$$\left[\left[\boldsymbol{\mathcal{U}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}}\right]^T\right]^{-1} \left[\widehat{\boldsymbol{\mathcal{U}}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}}\right]^T = I \Rightarrow \boldsymbol{\mathcal{U}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}} = \widehat{\boldsymbol{\mathcal{U}}}^{\boldsymbol{\mathcal{M}+1}}_{\boldsymbol{(\mathcal{M}+1)}}.$$

So, the projected subspaces are same, that means, the proposed method extracts same set of basis vectors if all $(\mathcal{M}+1)$ multidimensional variables are available simultaneously.

6.2.3 IMCCA Under Ridge Regression Model

To compute the basis vectors, the inverse of the covariance matrix $\mathcal{C}_{ii}$ is needed; $\forall i \in \{1, 2, \cdots, (\mathcal{M}+1)\}$. If $n \ll m_i$, the covariance matrix $\mathcal{C}_{ii}$ becomes non-invertible, which leads to the invalid computation of MCCA [68]. Both shrinkage and regularization parameters are used to overcome this problem. The shrinkage parameter s_i is used to take care of the singularity problem of $\mathcal{C}_{ii}$ by reducing the off-diagonal elements, which can be computed as follows:

$$s_i = \frac{\sum\limits_{k \neq \ell} \hat{\mathcal{V}}([\mathcal{C}_{ii}]_{k\ell})}{\sum\limits_{k \neq \ell} [\mathcal{C}_{ii}^2]_{k\ell}} \tag{6.14}$$

where $\hat{\mathcal{V}}([\mathcal{C}_{ii}]_{k\ell})$ is the unbiased empirical variance of $[\mathcal{C}_{ii}]_{k\ell}$. Hence, to deal with the singularity issue, the covariance matrices can be redefined as follows:

$$[\tilde{\mathcal{C}}_{ii}]_{k\ell} = \begin{cases} (1 - s_i)[\mathcal{C}_{ii}]_{k\ell}; & \text{if } k \neq \ell \\ [\mathcal{C}_{ii}]_{k\ell}; & \text{otherwise.} \end{cases} \tag{6.15}$$

Moreover, a ridge regression optimization scheme is used to overcome the above problem, where a small positive quantity $\mathfrak{r}_i$, known as regularization parameter, is added to the diagonals of covariance matrix $\tilde{\mathcal{C}}_{ii}$. Let us assume that the ℓ-th dimension of the i-th multidimensional variable $\mathcal{X}_i[\ell]$ is contaminated with noise $\varepsilon_i[\ell]$, $\forall \ell \in \{1, 2, \cdots, m_i\}$, such that $\mathcal{E}[\varepsilon_i[\ell]] = 0$,$\mathcal{E}[\varepsilon_i[\ell]\varepsilon_i[k]^T] = 0$ for $\ell \neq k$, $\mathcal{E}[\varepsilon_i[\ell]\mathcal{X}_i[\ell]^T] = 0$ and $\mathcal{E}[\varepsilon_i[\ell]\varepsilon_i[\ell]^T] = \mathfrak{r}_i \geqslant 0$. Under these assumptions, the cross-covariance matrix of $\mathcal{X}_i$ and $\mathcal{X}_j$ is $\mathcal{C}_{ij}$ and the covariance matrix of $\mathcal{X}_i$ becomes $[\tilde{\mathcal{C}}_{ii} + \mathfrak{r}_i I]$. To estimate the basis vector w_i, the covariance matrix $\tilde{\mathcal{C}}_{ii}$ needs to be replaced by $[\tilde{\mathcal{C}}_{ii} + \mathfrak{r}_i I]$. This is similar to the ridge regression modification [278].

To estimate the optimal set of regularization parameters, a grid search optimization is performed, where each regularization parameter $\mathfrak{r}_i$ follows an arithmetic progression and is varied within a specified range. The optimal set of regularization parameters can be estimated in such a way that the correlation between multiset canonical variables is maximum. Let $\mathfrak{t}_i$ be the number of possible values of regularization parameter $\mathfrak{r}_i$, while d_i indicates the common difference for $\mathfrak{r}_i$. As $\mathfrak{r}_i$ is varied within a specified range $[\mathfrak{r}_{min}, \mathfrak{r}_{max}]$, the inverse of covariance matrix of each multidimensional variable has to be computed $\mathfrak{t}_i$ times. As the diagonal elements of $\tilde{\mathcal{C}}_{ii}$ are only changed by adding $\mathfrak{r}_i$, the eigenvalues of $[\tilde{\mathcal{C}}_{ii} + (\mathfrak{r}_i + k_i d_i)I]$, $\forall k_i \in \{0, 1, \cdots, (\mathfrak{t}_i - 1)\}$, are changed, but the corresponding eigenvectors remain same [183]. Also, there exists a relation between the eigenvalues of $[\tilde{\mathcal{C}}_{ii} + \mathfrak{r}_i I]$ and that of $[\tilde{\mathcal{C}}_{ii} + (\mathfrak{r}_i + k_i d_i)I]$, which is given by,

$$\Delta_{i k_i} = \Delta_i + k_i d_i I; \tag{6.16}$$

where $\Delta_{i k_i}$ is the diagonal matrix, whose diagonal elements are the eigenvalues of $[\tilde{\mathcal{C}}_{ii} + (\mathfrak{r}_i + k_i d_i)I]$, $\Delta_i = \Delta_{i0}$. Let the corresponding eigenvectors of the matrix $[\tilde{\mathcal{C}}_{ii} + (\mathfrak{r}_i + k_i d_i)I]$ be the columns of Ω_i. Based on spectral decomposition, the covariance matrix $[\tilde{\mathcal{C}}_{ii} + (\mathfrak{r}_i + k_i d_i)I]$

can be expressed as follows [269]:

$$[\tilde{\mathcal{C}}_{ii} + (\mathfrak{r}_i + \mathscr{k}_i \mathscr{d}_i) I] = \Omega_i \Delta_{i\mathscr{k}_i} \Omega_i^T = \Omega_i [\Delta_i + \mathscr{k}_i \mathscr{d}_i I] \Omega_i^T = \sum_{\ell=1}^{m_i} (\delta_{i\ell} + \mathscr{k}_i \mathscr{d}_i) \omega_{i\ell} \omega_{i\ell}^T; \tag{6.17}$$

and the inverse covariance matrix $[\tilde{\mathcal{C}}_{ii} + (\mathfrak{r}_i + \mathscr{k}_i \mathscr{d}_i) I]^{-1}$ can be computed as follows:

$$[\tilde{\mathcal{C}}_{ii} + (\mathfrak{r}_i + \mathscr{k}_i \mathscr{d}_i) I]^{-1} = \sum_{\ell=1}^{m_i} \frac{1}{(\delta_{i\ell} + \mathscr{k}_i \mathscr{d}_i)} \omega_{i\ell} \omega_{i\ell}^T; \tag{6.18}$$

where the ℓ-th element $\delta_{i\ell}$ of diagonal matrix Δ_i denotes the ℓ-th eigenvalue of the matrix $[\tilde{\mathcal{C}}_{ii} + \mathfrak{r}_i I]$. The ℓ-th column of the matrix Ω_i represents the orthogonalized eigenvector $\omega_{i\ell}$ corresponding to the eigenvalue $\delta_{i\ell}$, $\forall \ell \in \{1, 2, \cdots, m_i\}$. From (6.18), it can be observed that there is no need to compute the eigenvalue for every $\mathfrak{r}_i$ of each $\mathcal{X}_i$. It is sufficient to calculate the eigenvalues $\delta_{i\ell}$ and eigenvectors $\omega_{i\ell}$ of covariance matrix corresponding to the initial value of $\mathfrak{r}_i$.

Moreover, the ℓ-th element of each diagonal matrix $[\Delta_i + \mathscr{k}_i \mathscr{d}_i I]$ is in arithmetic progression, as the regularization parameters follow an arithmetic progression. Hence, the ℓ-th element of each diagonal matrix $[\Delta_i + \mathscr{k}_i \mathscr{d}_i I]^{-1}$ follows harmonic progression, that is, the ℓ-th element of all diagonal matrices $[\Delta_i + \mathscr{k}_i \mathscr{d}_i I]^{-1}$ be $\frac{1}{\delta_{i\ell}}, \frac{1}{\delta_{i\ell} + \mathscr{d}_i}, \frac{1}{\delta_{i\ell} + 2\mathscr{d}_i}, \cdots, \frac{1}{\delta_{i\ell} + (\mathfrak{t}_i - 1)\mathscr{d}_i}$. Now,

$$\frac{1}{\delta_{i\ell} + \mathscr{k}_i \mathscr{d}_i} = \frac{1}{\delta_{i\ell}} - \sum_{j=1}^{\mathscr{k}_i} \frac{\mathscr{d}_i}{(\delta_{i\ell} + (j-1)\mathscr{d}_i)(\delta_{i\ell} + j\mathscr{d}_i)}. \tag{6.19}$$

So, using (5.3.2) of Chapter 5, the inverse covariance matrix of (6.18) can be expressed as

$$[\tilde{\mathcal{C}}_{ii} + (\mathfrak{r}_i + \mathscr{k}_i \mathscr{d}_i) I]^{-1} = [\tilde{\mathcal{C}}_{ii} + (\mathfrak{r}_i + (\mathscr{k}_i - 1)\mathscr{d}_i) I]^{-1} - \Upsilon_{i\mathscr{k}_i}$$

$$= [\tilde{\mathcal{C}}_{ii} + \mathfrak{r}_i I]^{-1} - \sum_{j=1}^{\mathscr{k}_i} \Upsilon_{ij} = \Omega_i \Delta_i^{-1} \Omega_i^T - \sum_{j=1}^{\mathscr{k}_i} \Upsilon_{ij} \tag{6.20}$$

$$\text{where} \quad \Upsilon_{i\mathscr{k}_i} = \Omega_i \left[\sum_{\ell=1}^{m_i} \widehat{\mathcal{C}_\ell} \mathcal{V}_{i\mathscr{k}_i} \widehat{\mathcal{D}_\ell} \right] \Omega_i^T;$$

and $\mathcal{V}_{i\mathscr{k}_i} \in \Re^{m_i}$ be a row vector, where

$$\mathcal{V}_{i\mathscr{k}_i} = \begin{bmatrix} \frac{\mathscr{d}_i}{(\delta_{i1} + (\mathscr{k}_i - 1)\mathscr{d}_i)(\delta_{i1} + \mathscr{k}_i \mathscr{d}_i)} \\ \frac{\mathscr{d}_i}{(\delta_{i2} + (\mathscr{k}_i - 1)\mathscr{d}_i)(\delta_{i2} + \mathscr{k}_i \mathscr{d}_i)} \\ \vdots \\ \frac{\mathscr{d}_i}{(\delta_{im_i} + (\mathscr{k}_i - 1)\mathscr{d}_i)(\delta_{im_i} + \mathscr{k}_i \mathscr{d}_i)} \end{bmatrix}^T;$$

$\forall \mathscr{k}_i \in \{1, 2, \cdots, (\mathfrak{t}_i - 1)\}$; $\widehat{\mathcal{C}_\ell}$ and $\widehat{\mathcal{D}_\ell}$ be a column vector and a square matrix of dimension

m_i, respectively, where

$$\widehat{\mathcal{C}}_\ell[j] = \begin{cases} 1 & \text{if} \quad \ell = j, \\ 0 & \text{otherwise;} \end{cases}$$

$$\text{and} \quad \widehat{\mathcal{D}}_\ell[j,t] = \begin{cases} 1 & \text{if} \quad \ell = j = t, \\ 0 & \text{otherwise;} \end{cases}$$

$\forall \ell, j, t \in \{1, 2, \cdots, m_i\}$. From (6.2.3), it is observed that the covariance matrix of each $\mathcal{X}_i$, corresponding to every $\mathfrak{r}_i$, can be computed from the covariance matrix corresponding to initial value of $\mathfrak{r}_i$. Hence, the matrix $[\Phi_{\mathcal{M}}\Theta_{\mathcal{M}}]$ of (6.2.1), corresponding to r-th regularization parameter of $(\mathcal{M}+1)$-th modality, can be expressed as

$$[\Phi_{\mathcal{M}}\Theta_{\mathcal{M}}]_r = \sum_{i=1}^{\mathcal{M}} \left[\tilde{\mathcal{C}}_{(\mathcal{M}+1)(\mathcal{M}+1)} + (\mathfrak{r}_{(\mathcal{M}+1)} + k_{(\mathcal{M}+1)} d_{(\mathcal{M}+1)}) I\right]^{-1}$$

$$\mathcal{C}_{(\mathcal{M}+1)i} \left[\tilde{\mathcal{C}}_{ii} + (\mathfrak{r}_i + k_i d_i) I\right]^{-1} \mathcal{C}_{i(\mathcal{M}+1)}$$

$$= \sum_{i=1}^{\mathcal{M}} \left(\left[\tilde{\mathcal{C}}_{(\mathcal{M}+1)(\mathcal{M}+1)} + (\mathfrak{r}_{(\mathcal{M}+1)} + (k_{(\mathcal{M}+1)} - 1) d_{(\mathcal{M}+1)}) I\right]^{-1} - \Upsilon_{(\mathcal{M}+1)_{k_{(\mathcal{M}+1)}}}\right) \mathcal{C}_{(\mathcal{M}+1)i}$$

$$\left(\left[\tilde{\mathcal{C}}_{ii} + (\mathfrak{r}_i + (k_i - 1) d_i) I\right]^{-1} - \Upsilon_{i_{k_i}}\right) \mathcal{C}_{i(\mathcal{M}+1)}$$

$$= [\Phi_{\mathcal{M}}\Theta_{\mathcal{M}}]_{(r-1)} - \tilde{\mathcal{G}}_{(\mathcal{M}+1)_r} - \hat{\mathcal{G}}_{(\mathcal{M}+1)_r} + \bar{\mathcal{G}}_{(\mathcal{M}+1)_r}$$

$$= [\Phi_{\mathcal{M}}\Theta_{\mathcal{M}}]_1 + \sum_{j=1}^{r} \left(\bar{\mathcal{G}}_{(\mathcal{M}+1)_j} - \tilde{\mathcal{G}}_{(\mathcal{M}+1)_j} - \hat{\mathcal{G}}_{(\mathcal{M}+1)_j}\right) \tag{6.21}$$

$$\text{where} \quad \bar{\mathcal{G}}_{(\mathcal{M}+1)_r} = \Upsilon_{(\mathcal{M}+1)_{k_{(\mathcal{M}+1)}}} \sum_{i=1}^{\mathcal{M}} \mathcal{C}_{(\mathcal{M}+1)i} \Upsilon_{i_{k_i}} \mathcal{C}_{i(\mathcal{M}+1)}; \tag{6.22}$$

$$\tilde{\mathcal{G}}_{(\mathcal{M}+1)_r} = \Upsilon_{(\mathcal{M}+1)_{k_{(\mathcal{M}+1)}}} \sum_{i=1}^{\mathcal{M}} \mathcal{C}_{(\mathcal{M}+1)i} \Omega_i [\Delta_i + (k_i - 1) d_i I]^{-1} \Omega_i^T \mathcal{C}_{i(\mathcal{M}+1)}; \tag{6.23}$$

$$\text{and} \quad \hat{\mathcal{G}}_{(\mathcal{M}+1)_r} = \Omega_{(\mathcal{M}+1)} [\Delta_i + (k_{(\mathcal{M}+1)} - 1) d_{(\mathcal{M}+1)} I]^{-1} \Omega_{(\mathcal{M}+1)}^T \sum_{i=1}^{\mathcal{M}} \mathcal{C}_{(\mathcal{M}+1)i} \Upsilon_{i_{k_i}} \mathcal{C}_{i(\mathcal{M}+1)}; \tag{6.24}$$

$\forall k_i \in \{1, 2, \cdots, (\mathfrak{t}_i - 1)\}$ and $\forall r \in \{1, 2, \cdots, \mathfrak{t}_{(\mathcal{M}+1)}\}$ represents the number of all possible values of regularization parameter of $(\mathcal{M}+1)$-th multidimensional variable.

From (6.2.1) and (6.2.3), we get the following relation:

$$[\Phi_{\mathcal{M}}\Theta_{\mathcal{M}}]_r = \left[\rho_1^{\mathcal{M}+1}(t)\right]^2 \left[w^{\mathcal{M}+1}_{(\mathcal{M}+1)_1}(t)\right] \left[w^{\mathcal{M}+1}_{(\mathcal{M}+1)_1}(t)\right]^T + \sum_{j=1}^{r} \left(\bar{\mathcal{G}}_{(\mathcal{M}+1)_j} - \tilde{\mathcal{G}}_{(\mathcal{M}+1)_j} - \hat{\mathcal{G}}_{(\mathcal{M}+1)_j}\right). \tag{6.25}$$

From (6.25), it is clear that the eigenvalues and eigenvectors of $[\tilde{\mathcal{C}}_{(\mathcal{M}+1)(\mathcal{M}+1)} + \mathfrak{r}_{(\mathcal{M}+1)}I]$ are enough to compute $[\Phi_{\mathcal{M}}\Theta_{\mathcal{M}}]_r$. There is no need to compute the eigenvalues and eigenvectors of $[\tilde{\mathcal{C}}_{(\mathcal{M}+1)(\mathcal{M}+1)} + (\mathfrak{r}_{(\mathcal{M}+1)} + k_{(\mathcal{M}+1)}d_{(\mathcal{M}+1)})I]$ corresponding to each value of $\mathfrak{r}_{(\mathcal{M}+1)}$ for computing $[\Phi_{\mathcal{M}}\Theta_{\mathcal{M}}]_r$. Hence, when a new multidimensional variable is added, the eigenvectors of $[\tilde{\mathcal{C}}_{(\mathcal{M}+1)(\mathcal{M}+1)} + \mathfrak{r}_{(\mathcal{M}+1)}I]$ have to be computed only to get the basis vector corresponding to any regularization parameter. Moreover, it is necessary to update the previously obtained basis vectors of preceding multidimensional variables. The relation between basis vectors, corresponding to r-th and initial regularization parameters, of $(\mathcal{M}+1)$-th multidimensional variable is shown in (6.25). So, at first, the basis vector of $(\mathcal{M}+1)$-th multidimensional variable corresponding to initial regularization parameter has to be computed. Then, the basis vectors of preceding multidimensional variables are updated according to (6.9).

6.3 Proposed Feature Extraction Algorithm

This section presents some analytical formulations required for the sequential generation of canonical variables when a new modality arrives. Moreover, in real-life data analysis, all the available modalities may not be relevant, some of them may provide even noisy and inconsistent information. Hence, the proposed method is designed in such a way that it considers multidimensional variables incrementally if they have relevant and significant information with respect to previously considered modalities.

6.3.1 Sequential Generation of Canonical Variables

When a new multidimensional variable is added, it is possible to extract all p number of features, where $p = \min\{m_1, m_2, \cdots, m_{(\mathcal{M}+1)}\}$, by computing the eigenvalue-eigenvector pairs of $[\Phi_{\mathcal{M}}\Theta_{\mathcal{M}}]_r$ using the Jacobi method, with a computational complexity $\mathcal{O}(p^3)$. However, in real-life high dimensional multimodal data analysis, the value of p is large, while a small fraction $\mathcal{D} << p$ is enough to deal with a certain problem. In multimodal data analysis, the goal is thus to extract a reduced set of most relevant features. This is an important problem in machine learning and termed as feature selection. So, in place of generating all p eigenvalue-eigenvector pairs using the Jacobi method, if each eigenvalue-eigenvector pair corresponding to matrix $[\Phi_{\mathcal{M}}\Theta_{\mathcal{M}}]_r$ is extracted sequentially, the quality of each generated feature can be evaluated, eventually, $\mathcal{D}$ features can be generated.

In the proposed algorithm, the Power method is used to compute each eigenvalue-eigenvector pair of $[\Phi_{\mathcal{M}}\Theta_{\mathcal{M}}]_r$ sequentially. The first eigenvalue-eigenvector pair is enough to compute any t-th eigenvalue-eigenvector pair as described below. The analytical formulation reported next establishes the relation between t-th and $(t+1)$-th eigenvalue-eigenvector pairs, which helps to generate correlated features sequentially. Now, the t-th eigenvalue of the matrix $[\Phi_{\mathcal{M}}\Theta_{\mathcal{M}}]_r$ is $\left[\rho_r^{\mathcal{M}+1}(t)\right]^2$ and the corresponding eigenvector is $w^{\mathcal{M}+1}_{(\mathcal{M}+1)_r}(t)$. So, using the Deflation method, we get

$$\left[[\Phi_{\mathcal{M}}\Theta_{\mathcal{M}}]_r - \left[\rho_r^{\mathcal{M}+1}(t)\right]^2 \left[w_{(\mathcal{M}+1)_r}^{\mathcal{M}+1}(t)\right]\left[w_{(\mathcal{M}+1)_r}^{\mathcal{M}+1}(t)\right]^T\right]\left[w_{(\mathcal{M}+1)_r}^{\mathcal{M}+1}(t+1)\right]$$

$$= \left[\rho_r^{\mathcal{M}+1}(t+1)\right]^2 \left[w_{(\mathcal{M}+1)_r}^{\mathcal{M}+1}(t+1)\right]. \tag{6.26}$$

Hence, from (6.3.1), it is proved that the $(t+1)$-th eigenvalue-eigenvector pair $\left\{\left[\rho_r^{\mathcal{M}+1}(t+1)\right]^2, \left[w_{(\mathcal{M}+1)_r}^{\mathcal{M}+1}(t+1)\right]\right\}$ of the matrix $[\Phi_{\mathcal{M}}\Theta_{\mathcal{M}}]_r$ is same as the first eigenvalue-eigenvector pair of the matrix $\left[[\Phi_{\mathcal{M}}\Theta_{\mathcal{M}}]_r - \left[\rho_r^{\mathcal{M}+1}(t)\right]^2 \left[w_{(\mathcal{M}+1)_r}^{\mathcal{M}+1}(t)\right]\left[w_{(\mathcal{M}+1)_r}^{\mathcal{M}+1}(t)\right]^T\right]$. For calculating $(t+1)$-th eigenvalue-eigenvector pair, the matrix $[\Phi_{\mathcal{M}}\Theta_{\mathcal{M}}]_r$ can be calculated as follows:

$$[\Phi_{\mathcal{M}}\Theta_{\mathcal{M}}]_r(t+1) = [\Phi_{\mathcal{M}}\Theta_{\mathcal{M}}]_r - \sum_{\ell=1}^{t}\left[\rho_r^{\mathcal{M}+1}(\ell)\right]^2 \left[w_{(\mathcal{M}+1)_r}^{\mathcal{M}+1}(\ell)\right]\left[w_{(\mathcal{M}+1)_r}^{\mathcal{M}+1}(\ell)\right]^T$$

$$= [\Phi_{\mathcal{M}}\Theta_{\mathcal{M}}]_1 + \sum_{j=1}^{r}\left(\bar{\mathcal{G}}_{(\mathcal{M}+1)_j} - \tilde{\mathcal{G}}_{(\mathcal{M}+1)_j} - \hat{\mathcal{G}}_{(\mathcal{M}+1)_j}\right) - \sum_{\ell=1}^{t}\left[\rho_r^{\mathcal{M}+1}(\ell)\right]^2 \left[w_{(\mathcal{M}+1)_r}^{\mathcal{M}+1}(\ell)\right]\left[w_{(\mathcal{M}+1)_r}^{\mathcal{M}+1}(\ell)\right]^T \tag{6.27}$$

where $[\Phi_{\mathcal{M}}\Theta_{\mathcal{M}}]_r = [\Phi_{\mathcal{M}}\Theta_{\mathcal{M}}]_r(1), \forall r \in \{2, 3, \cdots, \mathfrak{t}_{(\mathcal{M}+1)}\}, \forall t \in \{1, 2, \cdots, (\mathcal{D}-1)\}$ and $\mathcal{D} \leqslant p$. From (6.3.1), it is clear that, when a new multidimensional variable $\mathcal{X}_{(\mathcal{M}+1)}$ is added, the first basis vector corresponding to the initial regularization parameter has to be computed and using that basis vector it is possible to compute any t-th basis vector corresponding to any regularization parameter of $\mathcal{X}_{(\mathcal{M}+1)}, \forall t \in \{2, 3 \cdots, \mathcal{D}\}$. The previous basis vectors of preceding multidimensional variables $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_{\mathcal{M}}$ have to be updated for each regularization parameter of $\mathcal{X}_{(\mathcal{M}+1)}$. The update of basis vectors is also done sequentially. Hence, from (6.9), it can be established that the $(t+1)$-th updated basis vectors of preceding multidimensional variables will be the $(t+1)$-th eigenvector of the matrix $\left[\rho^{\mathcal{M}} \mathcal{W}_{\mathcal{M}}^{\mathcal{M}} \mathcal{W}_{\mathcal{M}}^{\mathcal{M}^T} + \rho^{\mathcal{M}+1}\Theta_{\mathcal{M}} w_{(\mathcal{M}+1)}^{\mathcal{M}+1}\left[w_{(\mathcal{M}+1)}^{\mathcal{M}+1}\right]^T \Theta_{\mathcal{M}}^{\dagger}\right]_r$. Hence,

$$\left[\rho^{\mathcal{M}} \mathcal{W}_{\mathcal{M}}^{\mathcal{M}} \mathcal{W}_{\mathcal{M}}^{\mathcal{M}^T} + \rho^{\mathcal{M}+1}\Theta_{\mathcal{M}} w_{(\mathcal{M}+1)}^{\mathcal{M}+1}\left[w_{(\mathcal{M}+1)}^{\mathcal{M}+1}\right]^T \Theta_{\mathcal{M}}^{\dagger}\right]_r(t+1)$$

$$= \left[\rho^{\mathcal{M}} \mathcal{W}_{\mathcal{M}}^{\mathcal{M}} \mathcal{W}_{\mathcal{M}}^{\mathcal{M}^T} + \rho^{\mathcal{M}+1}\Theta_{\mathcal{M}} w_{(\mathcal{M}+1)}^{\mathcal{M}+1}\left[w_{(\mathcal{M}+1)}^{\mathcal{M}+1}\right]^T \Theta_{\mathcal{M}}^{\dagger}\right]_r$$

$$- \sum_{\ell=1}^{t}\left[\rho_r^{\mathcal{M}+1}(\ell)\right]^2 \left[\mathcal{W}_{\mathcal{M}}^{\mathcal{M}+1}{}_r(\ell)\right]\left[\mathcal{W}_{\mathcal{M}}^{\mathcal{M}+1}{}_r(\ell)\right]^T. \tag{6.28}$$

From (6.3.1), it is seen that the update of the t-th basis vectors of preceding multidimensional variables can be done by using all previously updated basis vectors of preceding multidimensional variables. Hence, the basis vectors of preceding multidimensional variables can be updated sequentially.

Let $\{\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_{(\mathcal{M}+1)}\}$ be the ordered list of $(\mathcal{M}+1)$ multidimensional data sets. The proposed IMCCA based feature generation algorithm starts with first two multidimensional variables $\{\mathcal{X}_1, \mathcal{X}_2\}$ and computes the largest eigenvalue $\left[\rho_r^2(t)\right]^2$ and eigenvector $w^2_{2\,r}(t)$ of the matrix $[\Phi_1\Theta_1]_r(t)$, where $w^2_{2\,r}(t)$ is the t-th basis vector of $\mathcal{X}_2$; $\forall r \in \{1, 2, \cdots, \prod_{\ell=1}^{2} \mathrm{t}_\ell\}$. On the other hand, the t-th basis vector $w^2_{1\,r}(t)$ of $\mathcal{X}_1$ can be computed as

$$w^2_{1\,r}(t) = \left(\Omega_1\Delta_1^{-1}\Omega_1^T - \sum_{s=1}^{k_1} \Upsilon_{1_s}\right) \mathcal{C}_{12} w^2_{2\,r}(t). \tag{6.29}$$

The algorithm to compute the basis vector $w^{j+1}_{(j+1)}$ of newly added multidimensional variable $\mathcal{X}_{(j+1)}$ and to update of the basis vectors of preceding all j ($\forall j \in \{2, 3, \cdots, \mathcal{M}\}$) multidimensional variables, is presented in Algorithm 6.1.

Algorithm 6.1 Algorithm for dynamic multiblock data

Input: Cross-covariance matrix $\mathcal{C}_{i(j+1)}$ of $\mathcal{X}_i$ and $\mathcal{X}_{j+1}$, eigenvalues $\Delta_{(j+1)}$ of covariance matrix $\tilde{\mathcal{C}}_{(j+1)(j+1)}$, along with corresponding eigenvectors $\Omega_{(j+1)}$, and the optimal solution $\{\rho^j, \mathcal{W}^j_j\}$ of preceding j multidimensional variables; $\forall i \in \{1, 2, \cdots, j\}$.

Output: The t-th basis vectors $w_i^{j+1}{}_r(t)$ of all $(j+1)$-th multidimensional variables corresponding to r-th regularization parameter, $\forall i \in \{1, 2, \cdots, (j+1)\}$.

```
for each r-th regularization parameters, where $\forall r \in \{1, 2, \cdots, \mathrm{t}_{(j+1)}\}$ for each t-th extracted feature do
    (i) Calculate $[\Phi_j\Theta_j]_r(t)$ using (6.2.3) if $t = 1$, otherwise using (6.3.1).
    (ii) Calculate largest eigenvalue $\left[\rho_r^{j+1}(t)\right]^2$ and eigenvector $w^{j+1}_{(j+1)_r}(t)$ of the matrix $[\Phi_j\Theta_j]_r(t)$, where $w^{j+1}_{(j+1)_r}(t)$ is the t-th basis vector of $(j+1)$-th multidimensional variable.
    (iii) Update the t-th basis vector $w_\ell^{j+1}{}_r(t)$ of the preceding multidimensional variables using (6.3.1), where $\forall \ell \in \{1, 2, \cdots, j\}$.
end for
```

6.3.2 Selection of Multidimensional Variables

The proposed feature extraction algorithm, termed as SeFGeIM, is designed in such a way that it considers the relevance of multidimensional variables while adding them sequentially. The largest eigenvalue of the covariance matrix of a modality represents its relevance value. It is the variance in the direction of the largest spread of the data. So, if the data is projected towards the direction of the largest spread or first principal axis, then the eigenvalue of the covariance matrix represents the variance of the data in that direction. As both eigenvalues and eigenvectors of each covariance matrix have to be computed to calculate the inverse of the covariance matrix for each multidimensional variable, the multidimensional variables

Algorithm 6.2 Proposed Algorithm: SeFGeIM

```
Input: $(\mathcal{M}+1)$ multidimensional variables $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_{(\mathcal{M}+1)}$.
Output: A set $\mathbb{S}^{\mathcal{M}+1}$ of $\mathcal{D}$ selected features.
Calculate the cross-covariance matrix $\mathcal{C}_{ij}$ of $\mathcal{X}_i$ and $\mathcal{X}_j$, $\forall i, j \in \{1, 2, \cdots, (\mathcal{M}+1)\}$ and $i < j$.
Calculate the covariance matrix $\tilde{\mathcal{C}}_{ii}$ of $\mathcal{X}_i$,$\forall i \in \{1, 2, \cdots, (\mathcal{M}+1)\}$.
Calculate eigenvalues $\delta_{i\ell}$ of $\tilde{\mathcal{C}}_{ii}$, along with corresponding eigenvectors $\omega_{i\ell}$,$\forall i \in \{1, 2, \cdots, (\mathcal{M}+1)\}$, $\forall \ell \in \{1, 2, \cdots, m_i\}$.
Construct the diagonal matrix $\Delta_i$, whose diagonal elements are $\delta_{i\ell}$, and the square matrix $\Omega_i$, whose each column is $\omega_{i\ell}$.
Order the multidimensional variables according their largest eigenvalues of the covariance matrices. Let $\{\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_{(\mathcal{M}+1)}\}$ be the order list.
for each $i = 2, \cdots, (\mathcal{M}+1)$ do
    (I) Initialize $\mathbb{S}^i = \varnothing$ and $t = 1$.
    (II) for each $t \leqslant \mathcal{D}$ do
        (i) Initialize $\mathbb{C}^i = \varnothing$.
        (ii) Compute the $t$-th basis vector $w^i_{k_r}(t)$ using (6.29) if $i = 2$; otherwise, using Algorithm 6.1; $\forall j \in \{1, 2, \cdots, (i-1)\}$ and $\forall k \in \{1, 2, \cdots, i\}$.
        (iii) for each $r$-th combinations of regularization parameters, where $\forall r \in \{1, 2, \cdots, \prod_{\ell=1}^{i} \mathfrak{t}_\ell\}$ if $i = 2$; otherwise, for all $r$-th regularization parameters, where $\forall r \in \{1, 2, \cdots, \mathfrak{t}_i\}$ do
            (a) Calculate the $t$-th canonical variable $\mathcal{U}_{j_r}(t)$; $\forall j \in \{1, 2, \cdots, i\}$ using (6.30).
            (b) Extract the $t$-th feature $\mathcal{F}^i_r(t)$ using (6.31).
            (c) Calculate the relevance $\gamma_{\mathcal{F}^i_r(t)}(\mathbb{D})$ of the feature $\mathcal{F}^i_r(t)$.
            (d) If $t > 1$, calculate the significance $\sigma_{\{\mathcal{F}^i_r(t), \mathcal{F}^i_\ell\}}(\mathbb{D}, \mathcal{F}^i_r(t))$ of the extracted feature $\mathcal{F}^i_r(t)$.
            (e) Add $\mathcal{F}^i_r(t)$ to $\mathbb{C}^i$ if its significance is non-zero with respect to all of the selected features of $\mathbb{S}^i$. In effect, $\mathbb{C}^i = \mathbb{C}^i \bigcup \mathcal{F}^i_r(t)$.
        (iv) end for
        (v) If $\mathbb{C}^i \neq \varnothing$, select a feature as $t$-th feature $\mathcal{F}^i_r(t)$ from all the features of $\mathbb{C}^i$, which maximizes the condition (6.32) when $t = 1$, otherwise (6.33). As a result of that, $\mathbb{S}^i = \mathbb{S}^i \bigcup \mathcal{F}^i_r(t)$.
        (vi) For $i > 2$, if the value of objective function ((6.32) for $t = 1$ and (6.33) otherwise) of the $t$-th feature of $\mathbb{S}^{i-1}$ is greater than that of $\mathbb{S}^i$, then $\mathbb{S}^i = \mathbb{S}^i \bigcup \mathcal{F}^{i-1}_r(t)$.
        (vii) Set $t = t + 1$.
    (III) end for
end for
Stop.
```

are arranged, in descending order, according to their largest eigenvalues of the covariance matrices.

Let $\{\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_{(\mathcal{M}+1)}\}$ be the ordered list of $(\mathcal{M}+1)$ multidimensional data sets, where each $\mathcal{X}_i \in \Re^{m_i \times n}$; m_i and n represent the number of features and samples, respectively, $\forall i \in \{1, 2, \cdots, (\mathcal{M}+1)\}$. Let us assume that each attribute is centered to have zero mean across the samples, and each regularization parameter $\mathfrak{r}_i$ is bounded by $[\mathfrak{r}_{min}, \mathfrak{r}_{max}]$. Let $\mathfrak{t}_i$ denote the number of all possible values of $\mathfrak{r}_i$ within that range. Let the t-th canonical variable

$$\mathcal{U}_{j_r}(t) = \left[w^i_{j_r}(t)\right]^T \mathcal{X}_j; \tag{6.30}$$

$\forall j \in \{1, 2, \cdots, i\}$, where $w^i_{j_r}(t)$ denotes the t-th basis vector of the j-th multidimensional variable and

$$\mathcal{F}^j_r(t) = \sum_{\ell=1}^{j} \mathcal{U}_{\ell_r}(t) \tag{6.31}$$

be the t-th extracted feature with r-th combination of regularization parameters of $\{\mathfrak{r}_{\mathcal{k}}\}$, $\forall \mathcal{k} \in \{1, 2, \cdots, j\}$, where all j multidimensional variables are considered, $\forall j \in \{2, 3, \cdots, i\}$. The relevance of the feature $\mathcal{F}^j_r(t)$ with respect to the sample categories $\mathbb{D}$ is denoted by $\gamma_{\mathcal{F}^j_r(t)}(\mathbb{D})$. Let $\sigma_{\{\mathcal{F}^j_r(t), \mathcal{F}^j_\ell\}}(\mathbb{D}, \mathcal{F}^j_r(t))$ denote the significance of the feature $\mathcal{F}^j_r(t)$ with respect to already-selected feature $\mathcal{F}^j_\ell \in \mathbb{S}^j$,$\mathbb{S}^j$ being the set of $\mathcal{D}$ selected features where all j multidimensional variables are considered and initially $\mathbb{S}^j \leftarrow \emptyset$. The optimal regularization parameters for each extracted feature can be selected by using the relevance and significance of that feature. Let us assume that the set $\mathbb{C}^j$ contains all the t-th extracted features which are computed by using all r-th combinations of regularization parameters, $\forall t \in \{1, 2, \cdots, \mathcal{D}\}$ where all j multidimensional variables are considered. For $t = 1$, the most relevant feature is selected from the set $\mathbb{C}^j$ and is included to $\mathbb{S}^j$, that is,

$$\mathcal{F}^j(t) = \underset{\mathcal{F}^j_r(t) \in \mathbb{C}^j}{\arg\max} \left\{\gamma_{\mathcal{F}^j_r(t)}(\mathbb{D})\right\}; \tag{6.32}$$

while for $t > 1$, the feature which has maximum relevance among the features of $\mathbb{C}^j$ and significance with respect to the features of $\mathbb{S}^j$ is selected as follows:

$$\mathcal{F}^j(t) = \underset{\mathcal{F}^j_r(t) \in \mathbb{C}^j}{\arg\max} \left\{ \gamma_{\mathcal{F}^j_r(t)}(\mathbb{D}) + \frac{1}{t-1} \sum_{\mathcal{F}^j_\ell \in \mathbb{S}^j} \sigma_{\{\mathcal{F}^j_r(t), \mathcal{F}^j_\ell\}}(\mathbb{D}, \mathcal{F}^j_r(t)) \right\}. \tag{6.33}$$

The proposed algorithm starts with first two multidimensional variables $\mathcal{X}_1$ and $\mathcal{X}_2$ from the ordered list, and produces a feature set $\mathbb{S}^2$. Then, other modalities are considered sequentially one after another. If the t-th feature of $\mathbb{S}^j$ has higher relevance and significance value than that of $\mathbb{S}^{j+1}$, $\forall j \in \{2, 3, \cdots, \mathcal{M}\}$, then the t-th feature of $\mathbb{S}^j$ is considered instead of $\mathbb{S}^{j+1}$. So, if a multidimensional variable $\mathcal{X}_{(j+1)}$ is relevant in extracting the t-th feature, then only it is considered, otherwise, the multidimensional variables $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_j$ are integrated to extract the t-th feature. So, each feature is extracted by integrating different number of multidimensional variables. The problem of generating a set of most significant and relevant feature set $\mathbb{S}^{\mathcal{M}+1}$ from the selected multiblock data sets is addressed by Algorithm 6.2. In the current research work, both significance and relevance of an extracted feature are computed by using the concept of rough hypercuboid approach [172].

6.3.3 Complexity Analysis

Let $\{\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_{(\mathcal{M}+1)}\}$ be the $(\mathcal{M}+1)$ multidimensional data sets, with c classes and n samples, where each $\mathcal{X}_i \in \Re^{m_i \times n}$ and m_i represents the number of features in $\mathcal{X}_i$. Let us assume that the regularization parameter $\mathfrak{r}_i$ has $\mathfrak{t}_i$ possible values. Let $\{\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_{(\mathcal{M}+1)}\}$ is the order list, which is rearranged according their largest eigenvalues of covariance matrices. Let $q = \max\{m_1, m_2, \cdots, m_{(\mathcal{M}+1)}\}$, $p = \min\{m_1, m_2, \cdots, m_{(\mathcal{M}+1)}\}$, and y denotes the second largest dimension among $\{m_1, m_2, \cdots, m_{(\mathcal{M}+1)}\}$, where the number of extracted features $\mathcal{D} << p$. Let $\tau = \mathfrak{t}_1\mathfrak{t}_2 + \sum\limits_{\ell=3}^{(\mathcal{M}+1)} \mathfrak{t}_\ell$.

All the cross-covariance matrices $\{\mathcal{C}_{ij}\}$ can be computed with a complexity $\mathcal{O}(\sum\limits_{i<j} m_i m_j n) \approx \mathcal{O}(qyn)$; whereas the total time complexity to compute all the covariance matrices $\{\mathcal{C}_{ii}\}$ is $\mathcal{O}(\sum\limits_{i} m_i^2 n) \approx \mathcal{O}(q^2 n)$, $\forall i, j \in \{1, 2, \cdots, (\mathcal{M}+1)\}$. All the eigenvalues $\delta_{i\ell}$, along with corresponding eigenvectors $\omega_{i\ell}$, are computed with computational complexity $\mathcal{O}(\sum\limits_{i} m_i^3) \approx \mathcal{O}(q^3)$; $\forall \ell \in \{1, 2, \cdots, m_i\}$, in step 3. On the other hand, step 4 and step 5 have constant time complexity of $\mathcal{O}(1)$. Thus, the total computational complexity of these five steps is $\mathcal{O}(qyn + q^2 n + q^3) \approx \mathcal{O}(q^3)$ as $n << q$.

In step 6, there is a loop which is executed $\mathcal{M}$ times. The first step of this loop has constant time complexity, which is given as $\mathcal{O}(1)$ and the next step has another loop, which is executed $\mathcal{D}$ times. Again the first step of this loop has constant time complexity, which is given by $\mathcal{O}(1)$. The complexity to compute $[\Phi_{i-1}\Theta_{i-1}]_r(t)$ is $\mathcal{O}(q^3 + \tau q^3) \approx \mathcal{O}(\tau q^3)$. The eigenvector of the matrix $[\Phi_{i-1}\Theta_{i-1}]_r(t)$ can be calculated with computational complexity $\mathcal{O}(m_i^2)$. If $i = 2$, the t-th basis vector $w^i_{j_r}(t)$ of j-th multidimensional variable can be computed with time complexity $\mathcal{O}(m_j^2)$, otherwise updation of the t-th basis vector $w^i_{j_r}(t)$ of the preceding multidimensional variables can be done with complexity $\mathcal{O}((\sum\limits_{k}^{j} m_k)^3)$, where $\forall j \in \{1, 2, \cdots, (i-1)\}$. The next step has another loop, which is executed $\prod\limits_{\ell=1}^{i} \mathfrak{t}_\ell$ times if $i = 2$; otherwise, $\mathfrak{t}_i$ times, $\forall i \in \{2, 3, \cdots, (\mathcal{M}+1)\}$. In step 6(II)(iii)(a), the t-th canonical variable $\mathcal{U}_{j_r}(t)$ can be computed with time complexity $\mathcal{O}(\sum\limits_{i} m_i n) \approx \mathcal{O}(qn)$. Hence, a feature $\mathcal{F}^i_r(t)$ can be extracted with computational complexity $\mathcal{O}(n)$. The step to compute both significance and relevance of a feature has same time complexity, which is given by $\mathcal{O}(cn)$. Step 6(II)(iii)(e) has constant time complexity of $\mathcal{O}(1)$. The selection of a feature from $\mathfrak{t}_1\mathfrak{t}_2$ candidate features, for $i = 2$, otherwise $(\tau - \mathfrak{t}_1\mathfrak{t}_2)$ candidate features, by maximizing both relevance and significance, has complexity $\mathcal{O}(\tau)$. The last step of the loop has constant time complexity of $\mathcal{O}(1)$. So, the total complexity to execute the loop $\mathcal{D}$ times is $\mathcal{O}(\mathcal{D}(\tau q^3 + m_i^2 + m_j^2 + (\sum\limits_{k}^{j} m_k)^3 + qn + n + cn + \tau)) \approx \mathcal{O}(\mathcal{D}\tau q^3)$. Hence, the proposed multi-block data integration algorithm has computational complexity of $\mathcal{O}(q^3 + \mathcal{D}\mathcal{T}q^3) \approx \mathcal{O}(\mathcal{D}\mathcal{T}q^3)$.

6.4 Performance Analysis

The performance of the proposed IMCCA based sequential feature generation algorithm, termed as SeFGeIM, is extensively studied and compared with that of several state-of-the-art multi-view data integration algorithms. To evaluate the performance of different algorithms, support vector machine with linear kernels is used. Five benchmark data sets, namely, CiteSeer, Handwritten, NUS-WIDE-OBJECT (NW-OBJECT), Reuters, and Caltech; and five cancer data sets, namely, glioblastoma multiforme (GBM), lung (LUNG), kidney (KIDNEY), lower grade glioma (LGG) and ovarian serous cystadenocarcinoma (OV), are considered in the current research work. All ten data sets are summarized in Table 5.1 and Table 5.2 of Chapter 5 and briefly described in Appendix A. Each of the regularization parameters of the proposed SeFGeIM algorithm is bounded in between 0.0 and 1.0 and varied with a difference of 0.1. The proposed algorithm is implemented in C language and run in Ubuntu 14.04 LTS having machine configuration Intel(R) Core(TM) i7-4790 CPU @ 3.60GHz×8 and 32 GB RAM. The source code of the proposed algorithm is available at `https://www.isical.ac.in/~bibl/results/sefgeim/sefgeim.html`.

Both 10-fold cross-validation (CV) and training-testing are performed to assess the performance of different algorithms. To analyze the statistical significance of the derived results in 10-fold CV, paired-t test (one-tailed), Wilcoxon signed rank test (one-tailed) and Friedman test (one-tailed), with 95% confidence level, are used to compute the p-values. For training-testing, the randomly selected 50% samples from each class are used for training and the rest are used for testing purpose for each of the data sets. For each data set, 25 top-ranked correlated features are selected for the analysis.

6.4.1 Gain in Execution Time

When a new modality is available for the analysis, the existing MCCA based method generates the canonical variables for all the modalities considering the original data augmented by the new modality. One of the important features of the proposed SeFGeIM algorithm is that it is based on the IMCCA model. The proposed IMCCA can generate the new canonical variables as well as modify the existing variables from the knowledge of canonical variables of the earlier modalities and the new modality only, without repeating the same procedure like existing MCCA. In effect, a significant gain in execution time can be

Table 6.1: Gain in Execution Time

Approach/Data	CiteSeer	Handwritten	NW-OBJECT	Reuters	Caltech
MCCA	6932.4	114.4	90.8	13922.7	19819.1
IMCCA	19.5	48.5	69.5	596.0	243.8
Approach/Data	GBM	LUNG	KIDNEY	LGG	OV
MCCA	2562.5	37931.0	5940.9	12578.5	997.2
IMCCA	139.7	251.4	219.4	287.6	284.1

achieved by the IMCCA when the number of features in a modality is high. Table 6.1 compares the performance of MCCA and IMCCA with respect to execution time (in second) on ten data sets. From the results, it can be seen that the IMCCA needs significantly lesser execution time than the MCCA, particularly for the data sets having modalities with large number of features.

6.4.2 Performance With Multiple Modalities

The superiority of the IMCCA over MCCA is its ability to handle dynamic data. In order to establish the relationship between the performance of the IMCCA based proposed SeFGeIM algorithm and the number of data modalities, extensive experimentation is carried out on five benchmark and five omics data sets. Figure 6.1 presents the variation of classification accuracy obtained by the proposed SeFGeIM algorithm with respect to the number of modalities or views, both for benchmark and omics data sets. The results are reported for 10-fold CV.

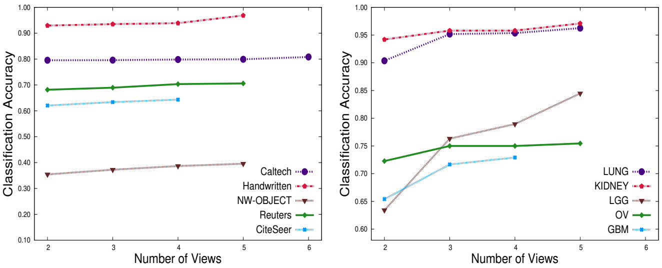


Figure 6.1: Variation of classification accuracy with respect to number of modalities / views (left: benchmark data; right: omics data).

6.4.3 Comparative Performance Analysis

Finally, Figure 6.2, Figure 6.3, Figure 6.4, Figure 6.5, Figure 6.8, Figure 6.9, Figure 6.10, and Figure 6.11 along with Table 6.2, Table 6.3, Table 6.4, Table 6.5, and Table 6.6 analyze the performance of the proposed multimodal data integration algorithm, termed as SeFGeIM, with that of (i) different criteria of the MCCA, namely, SUMCOR, MAXVAR, generalized variance (GENVAR), minimum variance (MINVAR), and sum of squared correlations (SSQCOR) [135]; (ii) various state-of-the-art MCCA based methods, namely, RGCCA [262], GMCCA [43], GMKCCA [43], large-scale generalized CCA (LasCCA) [84], distributed generalized CCA (DisCCA) [84], block sparse MCCA (BsMCCA) [235], and ReDMiCA [185] presented in Chapter 5; (iii) two popular multidimensional data integration algorithms, namely, multi-view discriminant analysis (MvDA) [128] and MvDA with view-consistency (MvDA-VC) [129]; (iv) three multi-view incremental algorithms, namely, live generalized canonical correlation analysis (LiveGCANO) [187], one-pass learning with incremental and decremental features (OPID) [114], and safe classification with augmented features (SAC) [113]; and (v) three deep learning-based algorithms, namely, deep MCCA (dMCCA) [244], deep multi-view learning via task-optimal CCA (TOCCA) [55], and multimodal deep Boltzmann machines (MDBM) [247].

Table 6.2: Classification Accuracy and Execution Time of Different Incremental Learning Algorithms on Benchmark and Omics Data Sets

Different	Data	Accuracy	Accuracy for 10-Fold CV			Time	Data	Accuracy	Accuracy for 10-Fold CV			Time
Algorithms	Sets	(Train-Test)	Mean	Median	StdDev	(in sec.)	Sets	(Train-Test)	Mean	Median	StdDev	(in sec.)
LiveGCANO	CiteSeer	0.213	0.177	0.368	0.000	68.7	GBM	0.533	0.567	0.542	0.095	101.1
OPID		0.462	0.401	0.393	0.034	15.3		0.648	0.725	**0.708**	0.045	52.0
SAC		0.450	0.385	0.377	0.045	58.6		0.638	0.667	0.646	0.092	507.1
SeFGeIM		**0.652**	**0.643**	**0.646**	0.027	93.5		**0.724**	**0.729**	0.671	0.166	639.7
LiveGCANO	Handwritten	0.063	0.114	0.113	0.042	10.0	LUNG	0.469	0.521	0.521	0.064	132.4
OPID		0.937	0.951	0.958	0.022	7.8		0.949	0.943	0.946	0.039	21.7
SAC		0.941	0.948	0.948	0.016	15.1		**0.963**	0.950	0.946	0.036	47.0
SeFGeIM		**0.966**	**0.969**	**0.970**	0.009	848.5		0.941	**0.963**	**0.955**	0.024	915.1
LiveGCANO	NW-OBJECT	0.076	0.108	0.112	0.022	482.5	KIDNEY	0.461	0.565	0.587	0.135	102.8
OPID		0.265	0.265	0.263	0.004	58.1		0.947	0.965	**0.968**	0.024	12.1
SAC		0.257	0.264	0.265	0.007	10030.1		**0.974**	0.948	0.935	0.027	43.3
SeFGeIM		**0.393**	**0.396**	**0.399**	0.008	669.5		0.967	**0.971**	0.955	0.018	911.9
LiveGCANO	Reuters	0.123	0.085	0.064	0.066	6363.7	LGG	0.269	0.387	0.374	0.065	104.5
OPID		0.562	0.591	0.589	0.012	6206.9		0.769	**0.847**	0.829	0.055	13.0
SAC		0.579	0.584	0.580	0.016	4347.3		0.909	0.842	**0.868**	0.054	43.3
SeFGeIM		**0.669**	**0.706**	**0.703**	0.014	5960.2		**0.946**	0.845	0.766	0.052	876.2
LiveGCANO	Caltech	0.016	0.026	0.026	0.007	317.6	OV	0.216	0.268	0.255	0.073	94.6
OPID		0.767	0.780	0.780	0.018	30.0		0.451	0.586	0.591	0.131	13.2
SAC		0.776	0.784	0.789	0.017	260.3		0.627	0.636	0.614	0.113	42.0
SeFGeIM		**0.891**	**0.808**	**0.809**	0.009	1943.8		**0.951**	**0.755**	**0.645**	0.096	984.1

Table 6.3: Statistical Significance Analysis of Different Algorithms on CiteSeer, Handwritten, GBM, and LUNG Data Sets

Different Algorithms		Data Sets	*p*-values for 10-Fold CV			Data Sets	*p*-values for 10-Fold CV		
			Paired-*t*	Wilcoxon	Friedman		Paired-*t*	Wilcoxon	Friedman
MCCA	SUMCOR	CiteSeer	**3.59E-05**	**2.53E-03**	**1.57E-03**	GBM	**6.62E-06**	**2.52E-03**	**1.57E-03**
	GENVAR		**5.58E-10**	**2.53E-03**	**1.57E-03**		**9.84E-06**	**2.46E-03**	**1.57E-03**
	MAXVAR		**2.60E-05**	**2.53E-03**	**1.57E-03**		*5.67E-02*	*5.46E-02*	**1.96E-02**
	MINVAR		**5.80E-06**	**2.53E-03**	**1.57E-03**		**2.87E-02**	**3.30E-02**	*9.56E-02*
	SSQCOR		**1.93E-06**	**2.53E-03**	**1.57E-03**		**2.95E-03**	**1.77E-02**	**1.14E-02**
RGCCA			**3.99E-05**	**3.82E-03**	**2.70E-03**		**1.05E-04**	**3.79E-03**	**2.70E-03**
GMCCA			**4.40E-09**	**2.52E-03**	**1.57E-03**		**1.60E-04**	**3.42E-03**	**1.14E-02**
GMKCCA			**9.00E-11**	**2.47E-03**	**1.57E-03**		**1.75E-05**	**3.37E-03**	**1.14E-02**
LasCCA			**5.60E-10**	**2.52E-03**	**1.57E-03**		**1.73E-02**	**3.27E-02**	**1.96E-02**
DisCCA			**1.95E-11**	**2.53E-03**	**1.57E-03**		**6.73E-06**	**2.47E-03**	**1.57E-03**
BsMCCA			**1.51E-10**	**2.53E-03**	**1.57E-03**		**3.72E-02**	**2.06E-02**	*5.78E-02*
ReDMiCA			*1.29E-01*	*1.98E-01*	*1.57E-01*		*4.04E-01*	*1.85E-01*	*3.17E-01*
MvDA			**6.91E-09**	**2.53E-03**	**1.57E-03**		*2.59E-01*	*1.29E-01*	*9.56E-02*
MvDA-VC			**2.30E-07**	**2.53E-03**	**1.57E-03**		*1.94E-01*	*6.85E-02*	*9.56E-02*
LiveGCANO			**7.49E-09**	**2.50E-03**	**1.57E-03**		**1.07E-02**	**3.70E-02**	**1.14E-02**
OPID			**1.62E-09**	**2.53E-03**	**1.57E-03**		*4.68E-01*	*2.76E-01*	*3.17E-01*
SAC			**3.10E-10**	**2.53E-03**	**1.57E-03**		*1.47E-01*	*1.71E-01*	*9.56E-02*
MCCA	SUMCOR	Handwritten	**9.73E-09**	**2.52E-03**	**1.57E-03**	LUNG	**3.81E-12**	**2.46E-03**	**1.57E-03**
	GENVAR		**1.93E-06**	**2.39E-03**	**1.57E-03**		**9.88E-07**	**2.52E-03**	**1.57E-03**
	MAXVAR		**3.52E-11**	**2.53E-03**	**1.57E-03**		**8.01E-06**	**2.53E-03**	**1.57E-03**
	MINVAR		**1.49E-13**	**2.53E-03**	**1.57E-03**		**5.68E-06**	**2.50E-03**	**1.57E-03**
	SSQCOR		**8.50E-13**	**2.52E-03**	**1.57E-03**		**2.72E-06**	**2.53E-03**	**1.57E-03**
RGCCA			**3.82E-14**	**2.52E-03**	**1.57E-03**		**3.55E-05**	**2.50E-03**	**1.57E-03**
GMCCA			**3.82E-14**	**2.52E-03**	**1.57E-03**		**4.34E-07**	**2.52E-03**	**1.57E-03**
GMKCCA			**1.96E-15**	**2.49E-03**	**1.57E-03**		**2.54E-03**	**3.84E-03**	**2.70E-03**
LasCCA			**6.04E-13**	**2.52E-03**	**1.57E-03**		**2.10E-04**	**2.47E-03**	**1.57E-03**
DisCCA			**2.20E-16**	**2.38E-03**	**1.57E-03**		**3.59E-08**	**2.53E-03**	**1.57E-03**
BsMCCA			**1.44E-12**	**2.53E-03**	**1.57E-03**		**1.19E-02**	**8.98E-03**	**8.15E-03**
ReDMiCA			5.00E-01	5.34E-01	7.06E-01		*2.17E-01*	*2.40E-01*	6.55E-01
MvDA			**1.20E-03**	**3.71E-03**	**2.70E-03**		*6.85E-02*	**3.80E-02**	*2.06E-01*
MvDA-VC			**5.99E-03**	**8.66E-03**	**1.96E-02**		*1.11E-01*	*6.03E-02*	*2.57E-01*
LiveGCANO			**2.01E-13**	**2.46E-03**	**1.57E-03**		**3.35E-09**	**2.49E-03**	**1.57E-03**
OPID			**1.03E-02**	**1.38E-02**	**1.14E-02**		**3.75E-02**	**3.91E-02**	*1.80E-01*
SAC			**8.47E-04**	**3.98E-03**	**1.14E-02**		*5.54E-02*	**4.24E-02**	*1.57E-01*

The classification accuracy on test samples of each data set in case of training-testing, as well as the mean, median, and standard deviation of accuracy for 10-fold cross-validation are reported in Table 6.2, for both benchmark and omics data sets. To analyze the performance of the proposed algorithm statistically, the *p*-values computed using three statistical tests are also reported in these tables. While both Figure 6.2, Figure 6.4, Figure 6.8, and Figure 6.10 are with respect to 10-fold CV, the corresponding results for training-testing are reported in Figure 6.3, Figure 6.5, Figure 6.9, and Figure 6.11. On the other hand, Figure 6.6 and Figure 6.7 show the scatter plots using the first two extracted features, along with the class separability index, of the proposed algorithm on five benchmark and five omics data sets, while the corresponding plots of aforementioned algorithms are reported in Figure 5.5, Figure 5.6, Figure 5.11, and Figure 5.12, of Chapter 5.

Table 6.4: Statistical Significance Analysis of Different Algorithms on NW-OBJECT, Reuters, KIDNEY, and LGG Data Sets

Different Algorithms		Data Sets	p-values for 10-Fold CV			Data Sets	p-values for 10-Fold CV		
			Paired-t	Wilcoxon	Friedman		Paired-t	Wilcoxon	Friedman
MCCA	SUMCOR	NW-OBJECT	**1.55E-11**	**2.53E-03**	**1.57E-03**	KIDNEY	**5.60E-09**	**2.36E-03**	**1.57E-03**
	GENVAR		**2.49E-16**	**2.53E-03**	**1.57E-03**		**9.13E-07**	**2.46E-03**	**1.57E-03**
	MAXVAR		**7.83E-14**	**2.53E-03**	**1.57E-03**		**5.12E-06**	**2.38E-03**	**1.57E-03**
	MINVAR		**2.20E-16**	**2.52E-03**	**1.57E-03**		**1.70E-06**	**2.50E-03**	**1.57E-03**
	SSQCOR		**5.92E-14**	**2.53E-03**	**1.57E-03**		**1.29E-03**	**3.98E-03**	**1.14E-02**
RGCCA			**2.20E-16**	**2.53E-03**	**1.57E-03**		**1.38E-02**	**1.78E-02**	*9.56E-02*
GMCCA			**1.15E-15**	**2.52E-03**	**1.57E-03**		**1.64E-06**	**2.50E-03**	**1.57E-03**
GMKCCA			**1.39E-12**	**2.53E-03**	**1.57E-03**		**7.76E-05**	**2.49E-03**	**1.57E-03**
LasCCA			**2.20E-16**	**2.53E-03**	**1.57E-03**		**3.21E-04**	**2.52E-03**	**1.57E-03**
DisCCA			**1.78E-13**	**2.53E-03**	**1.57E-03**		**2.94E-07**	**2.38E-03**	**1.57E-03**
BsMCCA			**4.16E-15**	**2.53E-03**	**1.57E-03**		**5.27E-04**	**3.30E-03**	**2.70E-03**
ReDMiCA			**6.79E-06**	**2.53E-03**	**1.57E-03**		5.00E-01	5.00E-01	1.00E+00
MvDA			**1.84E-12**	**2.53E-03**	**1.57E-03**		**1.01E-04**	**2.88E-03**	**2.70E-03**
MvDA-VC			**7.49E-11**	**2.53E-03**	**1.57E-03**		**1.93E-03**	**6.94E-03**	**8.15E-03**
LiveGCANO			**2.05E-11**	**2.53E-03**	**1.57E-03**		**3.53E-06**	**2.50E-03**	**1.57E-03**
OPID			**7.23E-13**	**2.52E-03**	**1.57E-03**		*1.72E-01*	*1.59E-01*	*3.17E-01*
SAC			**2.90E-11**	**2.53E-03**	**1.57E-03**		**1.24E-02**	**1.74E-02**	**3.39E-02**
MCCA	SUMCOR	Reuters	**1.28E-07**	**2.53E-03**	**1.57E-03**	LGG	**7.20E-08**	**2.50E-03**	**1.57E-03**
	GENVAR		**1.25E-13**	**2.53E-03**	**1.57E-03**		**1.09E-07**	**2.50E-03**	**1.57E-03**
	MAXVAR		**5.22E-13**	**2.53E-03**	**1.57E-03**		**1.35E-08**	**2.52E-03**	**1.57E-03**
	MINVAR		**3.25E-13**	**2.53E-03**	**1.57E-03**		**2.60E-08**	**2.50E-03**	**1.57E-03**
	SSQCOR		**3.62E-15**	**2.53E-03**	**1.57E-03**		**2.19E-07**	**2.50E-03**	**1.57E-03**
RGCCA			**4.23E-12**	**2.53E-03**	**1.57E-03**		**5.17E-06**	**2.53E-03**	**1.57E-03**
GMCCA			**1.11E-13**	**2.53E-03**	**1.57E-03**		**4.22E-07**	**2.50E-03**	**1.57E-03**
GMKCCA			**4.95E-14**	**2.52E-03**	**1.57E-03**		**1.49E-09**	**2.50E-03**	**1.57E-03**
LasCCA			**3.05E-11**	**2.53E-03**	**1.57E-03**		**6.70E-09**	**2.45E-03**	**1.57E-03**
DisCCA			**2.49E-08**	**2.53E-03**	**1.57E-03**		**1.29E-08**	**2.52E-03**	**1.57E-03**
BsMCCA			**1.03E-11**	**2.53E-03**	**1.57E-03**		**4.22E-05**	**2.50E-03**	**1.57E-03**
ReDMiCA			**6.83E-03**	**1.09E-02**	**1.14E-02**		6.47E-01	*4.58E-01*	1.00E+00
MvDA			**9.56E-10**	**2.53E-03**	**1.57E-03**		**3.99E-04**	**3.76E-03**	**2.70E-03**
MvDA-VC			**6.29E-10**	**2.53E-03**	**1.57E-03**		*1.62E-01*	*1.53E-01*	5.27E-01
LiveGCANO			**8.27E-11**	**2.53E-03**	**1.57E-03**		**2.82E-10**	**2.50E-03**	**1.57E-03**
OPID			**5.24E-10**	**2.53E-03**	**1.57E-03**		5.46E-01	5.56E-01	1.00E+00
SAC			**3.70E-09**	**2.53E-03**	**1.57E-03**		*4.30E-01*	*4.52E-01*	7.39E-01

6.4.3.1 Various Criteria of MCCA

Figure 6.2, Figure 6.3, Figure 6.4, and Figure 6.5, show the variation of mean classification accuracy with respect to the number of extracted features for various criteria of the MCCA and the proposed algorithm, on both benchmark and omics data sets. All these results convey that the performance of the proposed feature extraction algorithm is significantly higher as compared to that of the various criteria of the MCCA, irrespective of the generated features and data sets used. The results reported in Table 5.3, Table 5.4, and Table 5.5 of Chapter 5 demonstrate that, among the five criteria of the MCCA, the SUMCOR attains the highest accuracy of 0.581 on CiteSeer, 0.870 on Handwritten, 0.303 on NW-OBJECT, and 0.575 on Reuters, while MAXVAR provides the highest classification accuracy of 0.733

Table 6.5: Statistical Significance Analysis of Different Algorithms on Caltech and OV Data Sets

Different Algorithms		Data Sets	p-values for 10-Fold CV			Data Sets	p-values for 10-Fold CV		
			Paired-t	Wilcoxon	Friedman		Paired-t	Wilcoxon	Friedman
MCCA	SUMCOR	Caltech	**2.58E-07**	**2.53E-03**	**1.57E-03**	OV	**2.21E-07**	**2.52E-03**	**1.57E-03**
	GENVAR		**1.62E-08**	**2.53E-03**	**1.57E-03**		**9.17E-08**	**2.50E-03**	**1.57E-03**
	MAXVAR		**2.06E-07**	**2.52E-03**	**1.57E-03**		**5.16E-04**	**3.71E-03**	**2.70E-03**
	MINVAR		**7.87E-09**	**2.52E-03**	**1.57E-03**		**2.13E-03**	**5.86E-03**	**4.68E-03**
	SSQCOR		**1.19E-07**	**2.52E-03**	**1.57E-03**		**4.05E-03**	**8.58E-03**	**3.39E-02**
RGCCA			**2.20E-16**	**2.49E-03**	**1.57E-03**		**1.63E-07**	**2.53E-03**	**1.57E-03**
GMCCA			**2.20E-16**	**2.52E-03**	**1.57E-03**		**1.93E-08**	**2.52E-03**	**1.57E-03**
GMKCCA			**2.20E-16**	**2.47E-03**	**1.57E-03**		**1.41E-05**	**2.50E-03**	**1.57E-03**
LasCCA			**1.94E-15**	**2.38E-03**	**1.57E-03**		**1.42E-06**	**2.50E-03**	**1.57E-03**
DisCCA			**7.62E-11**	**2.53E-03**	**1.57E-03**		**5.44E-08**	**2.39E-03**	**1.57E-03**
BsMCCA			**7.87E-09**	**2.52E-03**	**1.57E-03**		*5.69E-02*	*6.95E-02*	*5.78E-02*
ReDMiCA			*1.45E-01*	*1.93E-01*	*2.06E-01*		*1.69E-01*	*1.66E-01*	*2.06E-01*
MvDA			**2.79E-03**	**5.36E-03**	**1.14E-02**		**1.89E-02**	**1.88E-02**	*9.56E-02*
MvDA-VC			**3.69E-03**	**3.79E-03**	**2.70E-03**		**2.79E-03**	**8.30E-03**	*5.78E-02*
LiveGCANO			**2.20E-16**	**2.52E-03**	**1.57E-03**		**1.12E-06**	**2.50E-03**	**1.57E-03**
OPID			**1.15E-04**	**2.53E-03**	**1.57E-03**		**1.88E-03**	**5.76E-03**	**4.68E-03**
SAC			**1.05E-03**	**2.52E-03**	**1.57E-03**		**1.73E-02**	**2.32E-02**	*5.78E-02*

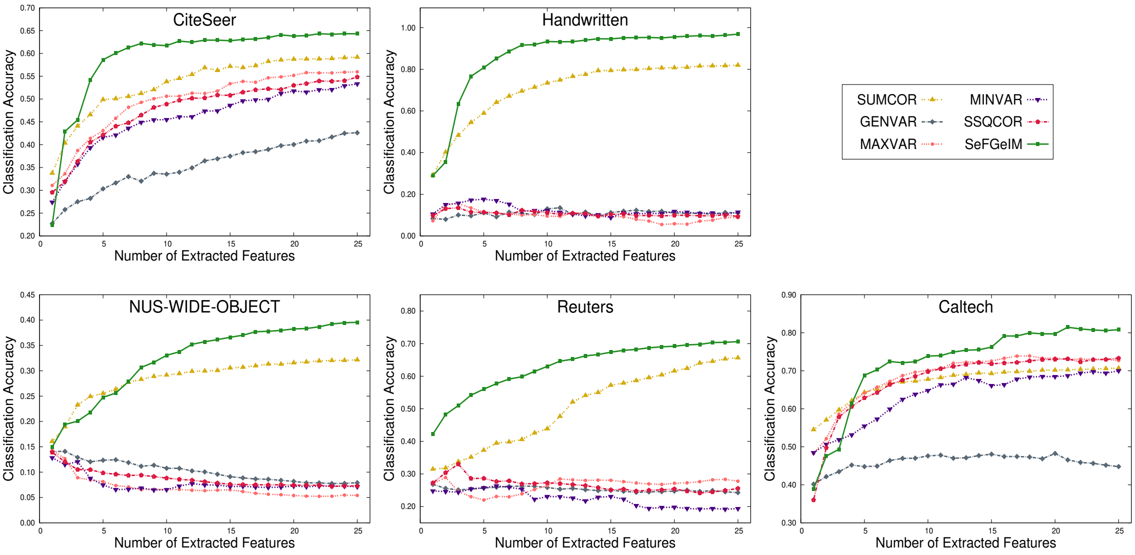


Figure 6.2: Variation of classification accuracy with respect to number of extracted features for different criteria of the MCCA and proposed (SeFGeIM) algorithm on benchmark data sets for 10-fold CV.

on Caltech data set. However, the proposed algorithm achieves the highest classification accuracy on all the benchmark data sets reported in Table 6.2. Similarly, for omics data sets, the proposed SeFGeIM algorithm attains higher classification accuracy than the five existing criteria of the MCCA. In case of 10-fold CV, the proposed algorithm attains significantly better p-values (marked in bold) than different criteria of the MCCA in 147 cases, out of total 150 cases, considering 95% confidence level. On the other hand, the

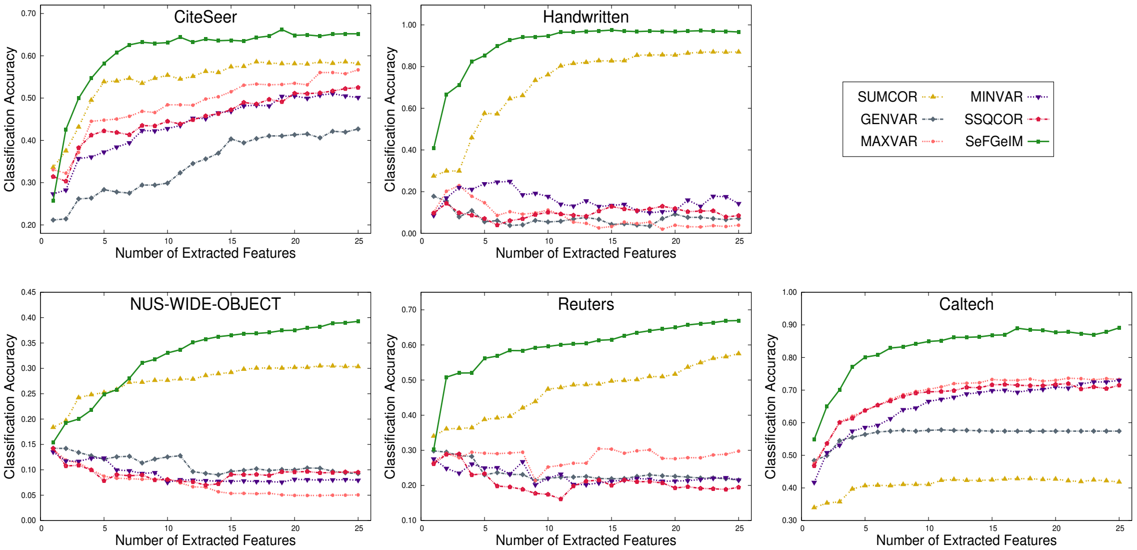


Figure 6.3: Variation of classification accuracy with respect to number of extracted features for different criteria of the MCCA and proposed (SeFGeIM) algorithm on benchmark data sets for training-testing.

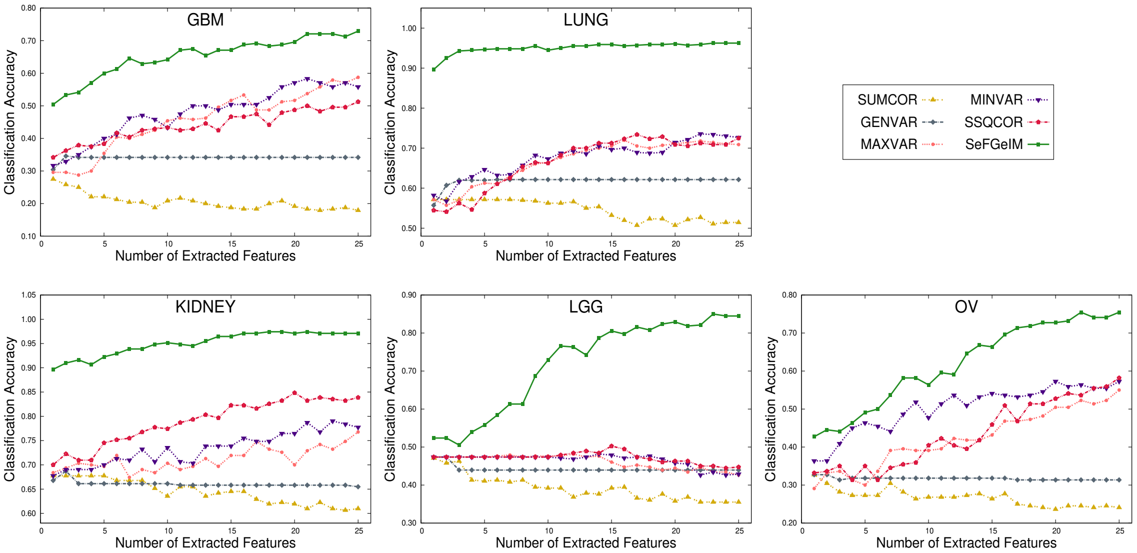


Figure 6.4: Variation of classification accuracy with respect to number of extracted features for different criteria of the MCCA and proposed (SeFGeIM) algorithm on omics data sets for 10-fold CV.

proposed algorithm provides better but not significant (marked in italics) p-values in only 3 cases, for MAXVAR using paired-t test and Wilcoxon signed rank test, and MINVAR using Friedman test on GBM data set. As mentioned in Section 5.4.1 of Chapter 5, all five criteria of MCCA cannot handle the 'large p and small n' issue of multidimensional data sets. On

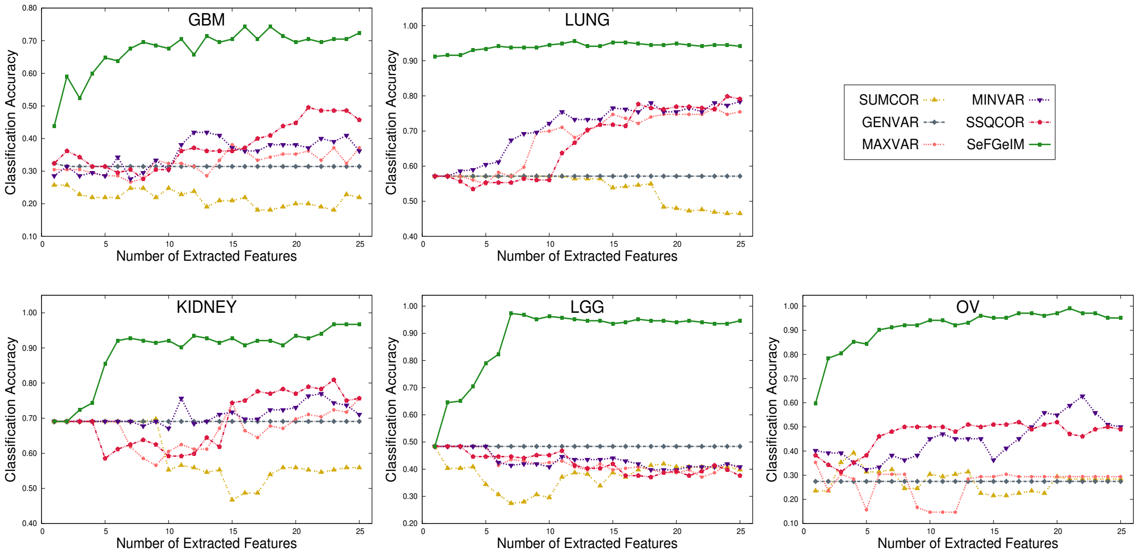
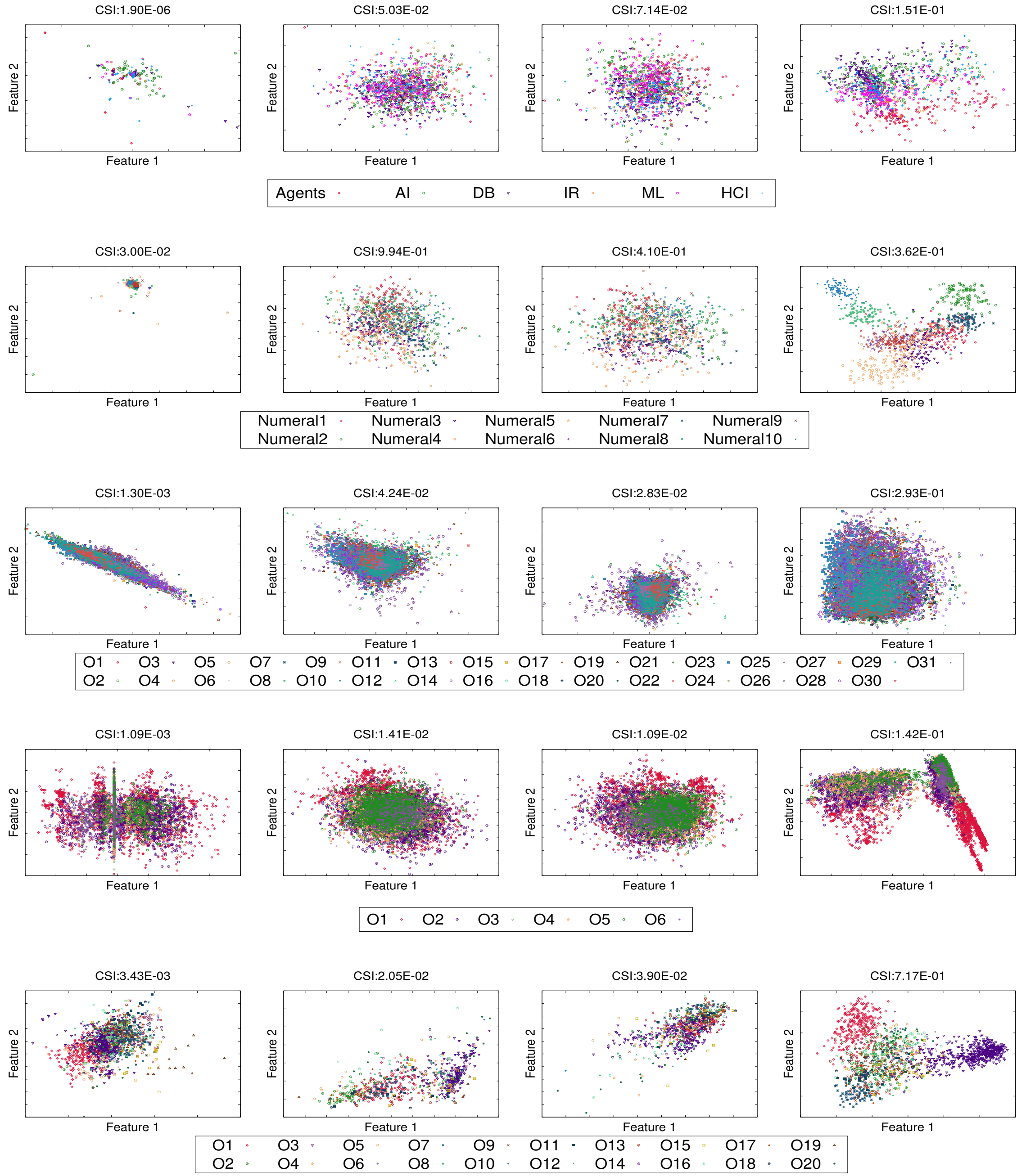


Figure 6.5: Variation of classification accuracy with respect to number of extracted features for different criteria of the MCCA and proposed (SeFGeIM) algorithm on omics data sets for training-testing.

the other hand, the SeFGeIM addresses this issue by using ridge regression optimization. As omics data sets suffer from high-dimension low-sample size problem, different criteria of MCCA do not perform well enough, while SeFGeIM obtains higher classification accuracy. Also, the significantly better performance of SeFGeIM is obtained due to the consideration of supervised information of sample categories. Moreover, SeFGeIM considers only the relevant views for the analysis. Thus, it obtains significantly higher classification accuracy than the various criteria of MCCA.

6.4.3.2 MCCA Based Methods

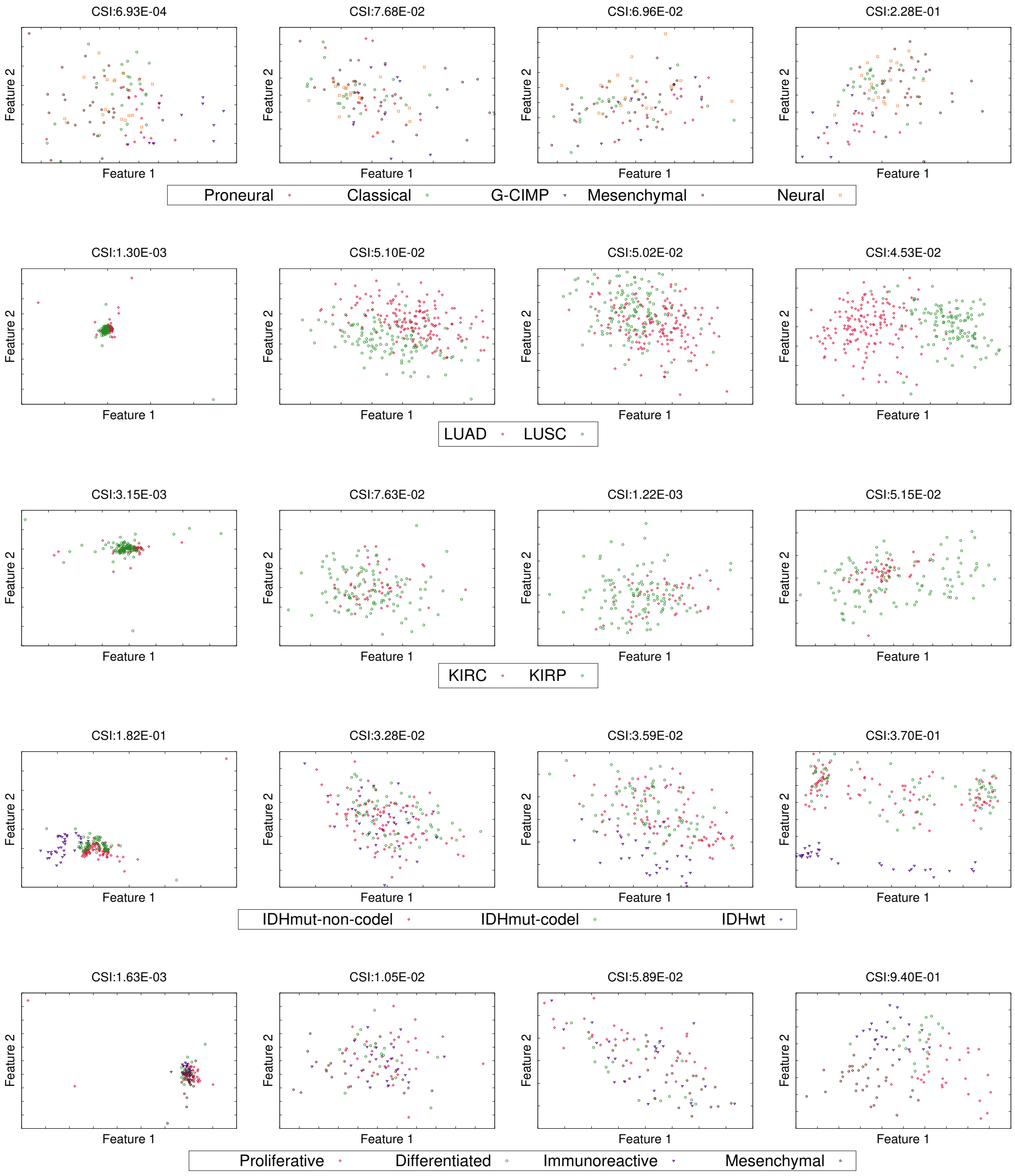
Figure 6.8 and Figure 6.10 present the variation of mean classification accuracy with respect to the number of extracted features for state-of-the-art methods as well as the proposed algorithm, in case of 10-fold CV on both benchmark and omics data sets, while the corresponding results for training-testing are reported in Figure 6.9 and Figure 6.11. The results reported here demonstrate that the mean accuracy of the proposed sequential feature generation algorithm is significantly higher as compared to that of the existing MCCA based methods, namely, RGCCA, GMCCA, GMKCCA, LasCCA, DisCCA, and BsMCCA, on both benchmark and omics data sets, irrespective of the number of generated features. All the results reported in Table 5.3, Table 5.4, and Table 5.5 of Chapter 5 and Table 6.2 confirm that the classification accuracy obtained by the proposed algorithm is higher than that of the six MCCA based algorithms, in case of training-testing, for all benchmark and omics data sets. In case of 10-fold CV, the proposed algorithm also attains the highest mean and median accuracy with respect to six MCCA based algorithms, irrespective of the data sets used. To analyze the performance of the proposed algorithm statistically, the p-values computed using three statistical tests are also reported in Table 6.3, Table 6.4,



(a) LiveGCANO (b) OPID (c) SAC (d) SeFGeIM

Figure 6.6: Scatter plots for different incremental learning algorithms (LiveGCANO, OPID, and SAC) and proposed (SeFGeIM) algorithm on benchmark data sets, along with class separability index (top to bottom: CiteSeer, Handwritten, NW-OBJECT, Reuters, Caltech), each Oi denotes the i-th object class.

and Table 6.5. Out of the total of 180 cases, the SeFGeIM algorithm attains significantly better p-values (marked in bold) than existing MCCA based methods in 175 cases. On the



(a) LiveGCANO (b) OPID (c) SAC (d) SeFGeIM

Figure 6.7: Scatter plots for different incremental learning algorithms (LiveGCANO, OPID, and SAC) and proposed (SeFGeIM) algorithm on omics data sets, along with class separability index (top to bottom: GBM, LUNG, KIDNEY, LGG, OV).

other hand, the proposed algorithm provides better but not significant (marked in italics) p-value in only 5 cases, with respect to RGCCA using Friedman test on the KIDNEY data set and BsMCCA using all three significance tests on OV data set and Friedman test on

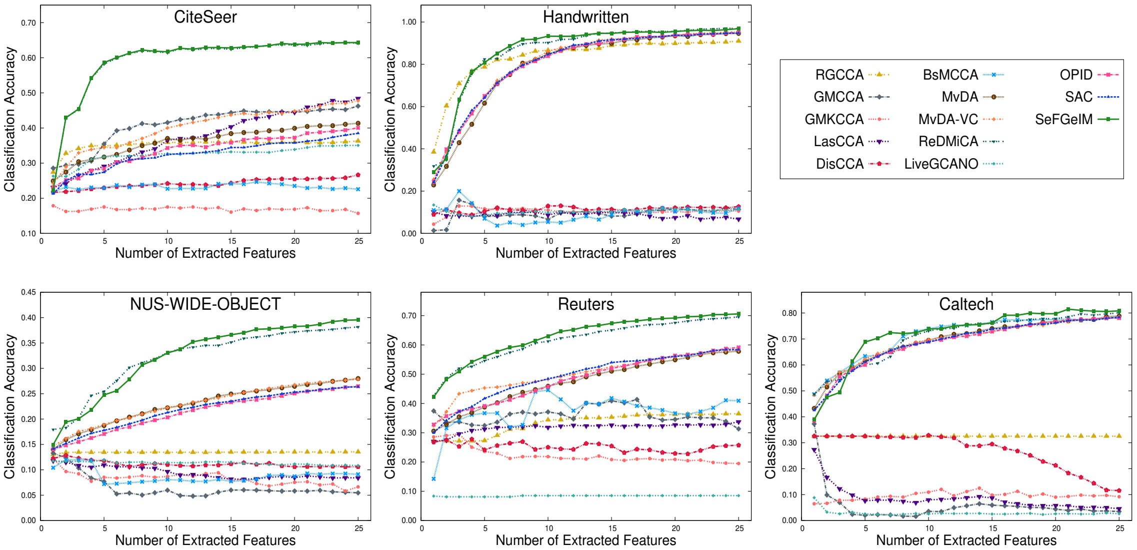


Figure 6.8: Variation of classification accuracy with respect to number of extracted features for different existing algorithms and proposed (SeFGeIM) algorithm on benchmark data sets for 10-fold CV.

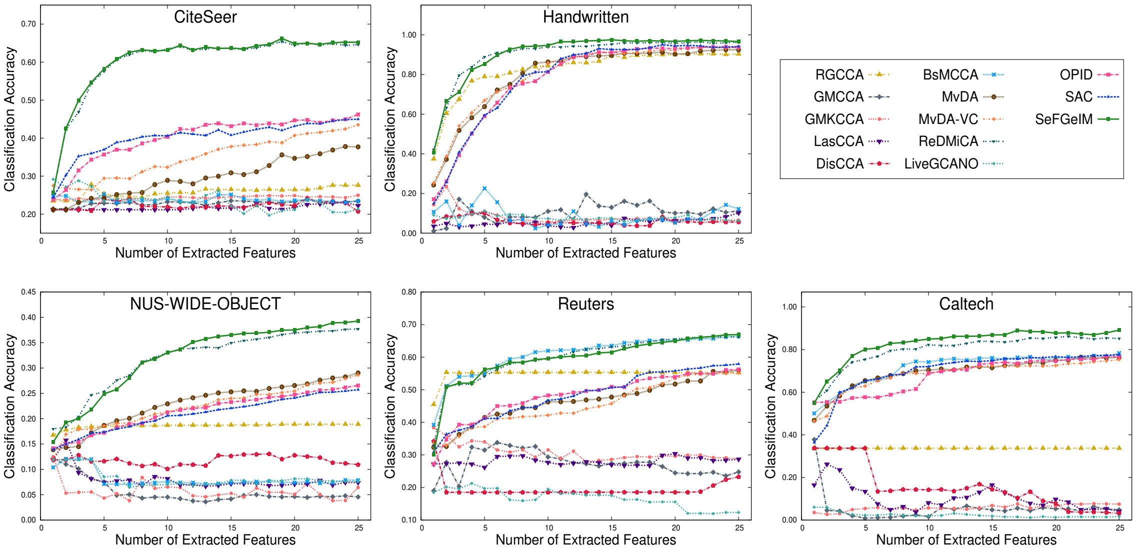


Figure 6.9: Variation of classification accuracy with respect to number of extracted features for different existing algorithms and proposed (SeFGeIM) algorithm on benchmark data sets for training-testing.

the GBM data set.

The last column of Figure 6.6 and Figure 6.7 depict the scatter plots with respect to the first two extracted features, for the proposed algorithm, while the corresponding results for different criteria of the MCCA are reported in the first five columns of Figure

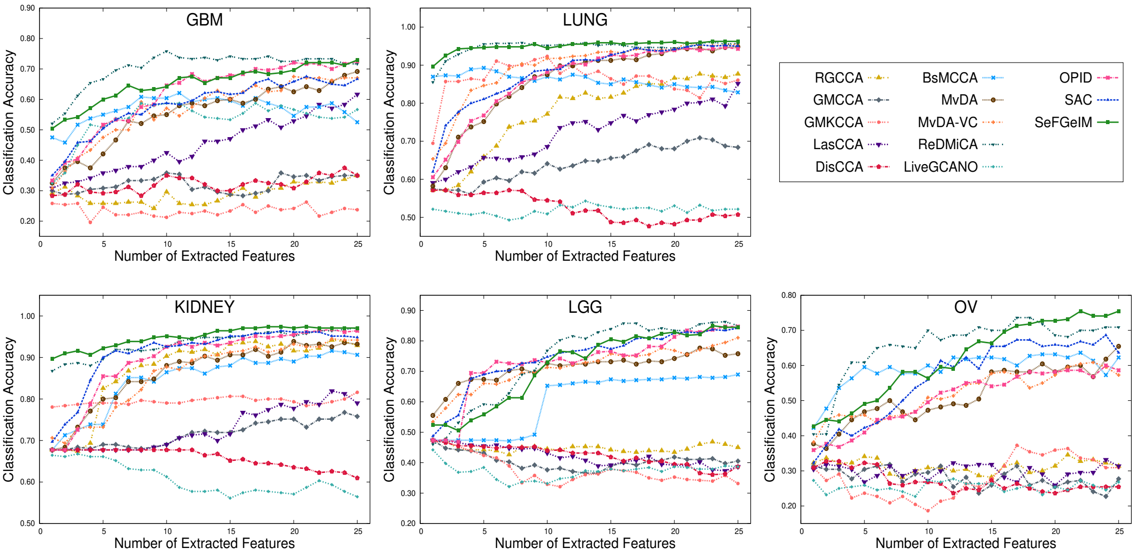


Figure 6.10: Variation of classification accuracy with respect to number of extracted features for different existing algorithms and proposed (SeFGeIM) algorithm on omics data sets for 10-fold CV.

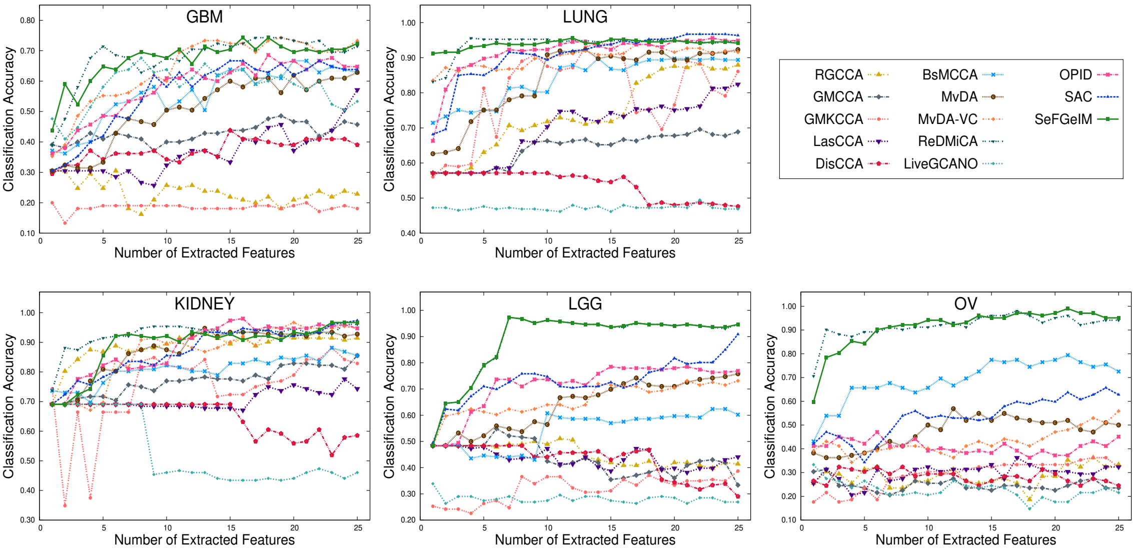


Figure 6.11: Variation of classification accuracy with respect to number of extracted features for different existing algorithms and proposed (SeFGeIM) algorithm on omics data sets for training-testing.

5.5 and Figure 5.6 of Chapter 5. All these results shows that the proposed algorithm is able to separate different classes of LUNG and OV data sets using the first two extracted features only, which is also evident from the corresponding class separability index values. On the other hand, the corresponding results of different criteria indicate that the existing

MCCA criteria are not able to separate different classes, irrespective of the data sets used. Moreover, the execution time of the proposed algorithm is significantly lower than that of most of the criteria of the MCCA, except GENVAR criterion.

Although the execution time of the proposed feature generation algorithm is higher than that of all the MCCA based methods on the data sets with a smaller number of samples, the proposed SeFGeIM algorithm needs significantly lower execution time than RGCCA, GMCCA and GMKCCA, for both NW-OBJECT and Reuters data sets where the number of samples is huge. Comparing the results of the last column of Figure 6.6 and Figure 6.7 with the first four columns of Figure 5.11 of Chapter 5 and the first two columns of Figure 5.12 of Chapter 5, it can be seen that the separation among various classes using the first two extracted features of the proposed algorithm is significantly better than that of the existing MCCA algorithms on both omics and benchmark data sets. Also, the existing algorithms are not incremental in nature.

The algorithm proposed in Chapter 5, termed as ReDMiCA (Regularized Discriminant Multi-View CCA) [185], is also an MCCA based multi-view learning algorithm. The comparative performance analysis of ReDMiCA and SeFGeIM are presented in Figure 6.8, Figure 6.9, Figure 6.10, and Figure 6.11. As the proposed SeFGeIM algorithm considers only the relevant views for the analysis, it obtains higher classification accuracy than other existing multi-view learning algorithms. From the results reported in Table 5.3, Table 5.4, and Table 5.5 of Chapter 5 and Table 6.2, it can be seen that the proposed SeFGeIM algorithm attains the higher classification accuracy of training-testing in 8 cases, while the ReDMiCA algorithm achieves it only for the LUNG data set and for LGG data set, both ReDMiCA and SeFGeIM achieve same classification accuracy. The results corresponding to 10-fold CV indicate that the proposed algorithm attains the higher mean accuracy in 7 cases out of total 10 cases each. On the other hand, for Handwritten and KIDNEY data sets the mean accuracy is same. From the results reported in Table 6.3, Table 6.4, and Table 6.5, it is evident that the proposed SeFGeIM algorithm attains significantly better p-values (marked in bold) than ReDMiCA in 6 cases, out of the total 30 cases, and better but not significant p-values (marked in italics) in 15 cases, considering 95% confidence level.

6.4.3.3 Multi-View Learning Algorithms

From Figure 6.8, Figure 6.9, Figure 6.10, and Figure 6.11 it is also seen that the mean classification accuracy of the proposed SeFGeIM algorithm is significantly higher, irrespective of the number of extracted features, as compared to that of two existing multimodal data integration methods, namely, MvDA and MvDA-VC on five benchmark data, and GBM, LUNG, KIDNEY, and OV data sets. In case of LGG data set, both MvDA and MvDA-VC provide higher mean accuracy for 10-fold cross-validation than the proposed algorithm for lower number ($\leqslant$9) of extracted features. The presence of highly redundant or similar features at the initial stages of the SeFGeIM algorithm leads to the lower classification accuracy in some of the folds of the LGG data. However, the proposed algorithm is able to alleviate this problem by generating relevant and significant features at the latter stages. In effect, the performance of the SeFGeIM improves drastically for higher number ($>$9) of features.

As shown in Table 5.3, Table 5.4, and Table 5.5 of Chapter 5 and Table 6.2, the proposed algorithm achieves higher accuracy, in case of training-testing, than both MvDA and

MvDA-VC for all the cases. In case of 10-fold CV, out of the total 60 cases, the SeFGeIM attains significantly better p-values (marked in bold) than two existing multimodal data integration methods in 44 cases, better but not significant p-values (marked in italics) in 15 cases, considering 95% confidence level. The proposed algorithm is not significantly better than MvDA-VC according to Friedman test on the LGG data set. Also, both the methods need lesser execution time than the proposed algorithm in most of the cases. Comparing the results of the last column of Figure 6.6 and Figure 6.7 with the 3rd and 4th columns of Figure 5.12 of Chapter 5, it is evident that different classes are more separable using the first two extracted features of the proposed algorithm than that of these two existing multi-view learning algorithms on omics as well as benchmark data sets. Moreover, unlike the proposed algorithm, both the existing algorithms are not incremental in nature.

6.4.3.4 Multi-View Incremental Learning Algorithms

The results reported in Figure 6.8, Figure 6.9, Figure 6.10, and Figure 6.11 clearly establish the fact that the proposed SeFGeIM algorithm provides better performance than three incremental multi-view data integration methods, namely, LiveGCANO, OPID, and SAC, in most of the cases. However, for the LGG data set, both OPID and SAC attain the higher mean accuracy than the proposed algorithm for lower number ($\leqslant$12) of features, while all of them provide similar performance for higher number ($>$12) of features. From the results presented in Table 6.2, it can be seen that the proposed algorithm attains the higher classification accuracy of training-testing in 8 cases, while the SAC algorithm achieves it only for the LUNG and KIDNEY data sets. The results corresponding to 10-fold CV indicate that the proposed algorithm attains the higher mean accuracy in 9 cases and higher median accuracy in 5 cases, out of total 10 cases each. From the results reported in Table 6.3, Table 6.4, and Table 6.5, it is evident that the proposed SeFGeIM algorithm attains significantly better p-values (marked in bold) than the three incremental methods in 71 cases, out of the total 90 cases, and better but not significant p-values (marked in italics) in 15 cases, considering 95% confidence level. The proposed algorithm is not significantly better than the SAC according to Friedman test on the LGG data set. Also, the OPID obtains the higher classification accuracy than the proposed algorithm for this case. Hence, the p-values with respect to the OPID, computed according to all three significance tests, are not better on the LGG data set. Although the execution time for the existing approaches is significantly lower than that of the proposed algorithm on the data sets having lower number of samples, it is comparable or even significantly higher than that of the SeFGeIM algorithm when the data sets like Reuters and NW-OBJECT have a large number of samples. Finally, the comparative analysis of the scatter plots presented in Figure 6.6 and Figure 6.7 confirm that the proposed algorithm can separate different classes better than the existing approaches.

6.4.3.5 Deep Learning Based Methods

Finally, the performance of the proposed SeFGeIM algorithm is compared with that of three deep learning-based methods, namely, dMCCA [244], TOCCA [55], and MDBM [247]. Table 6.6 presents the statistical significance analysis on five omics data sets. The results reported in Table 6.6 establish that the proposed algorithm attains significantly better p-

Table 6.6: Statistical Significance Analysis of Different Deep Learning Algorithms on Omics Data Sets

Data Sets	Different Algorithms	p-values for 10-Fold CV		
		Paired-t	Wilcoxon	Friedman
GBM	dMCCA	**2.26E-03**	**8.30E-03**	**1.14E-02**
	TOCCA	**3.27E-04**	**3.46E-03**	**1.14E-02**
	MDBM	**1.77E-04**	**2.53E-03**	**1.57E-03**
LUNG	dMCCA	**2.17E-10**	**2.46E-03**	**1.57E-03**
	TOCCA	**1.17E-12**	**2.20E-03**	**1.57E-03**
	MDBM	**5.27E-04**	**2.40E-03**	**1.57E-03**
KIDNEY	dMCCA	**4.49E-10**	**1.80E-03**	**1.57E-03**
	TOCCA	**2.46E-07**	**2.24E-03**	**1.57E-03**
	MDBM	**1.24E-05**	**2.83E-03**	**1.14E-02**
LGG	dMCCA	**2.49E-06**	**2.49E-03**	**1.57E-03**
	TOCCA	**3.38E-08**	**2.46E-03**	**1.57E-03**
	MDBM	**1.98E-06**	**2.46E-03**	**1.57E-03**
OV	dMCCA	**7.00E-07**	**2.53E-03**	**1.57E-03**
	TOCCA	**1.73E-05**	**2.53E-03**	**1.57E-03**
	MDBM	**3.80E-06**	**2.53E-03**	**1.57E-03**

values than the three deep learning-based methods, irrespective of the significance analysis and omics data sets used. Also, the results presented in Table 5.9 and Table 5.10 of Chapter 5 and Table 6.2 demonstrate that the classification accuracy of the proposed algorithm is significantly higher as compared to that of various deep learning-based methods in most of the cases. The TOCCA algorithm performs well on benchmark data sets, but it fails to achieve judicious results on omics data sets. The MDBM and dMCCA obtain 87.9% and 86.2% accuracy on LUNG and KIDNEY data sets, respectively, whereas both of them perform moderately on the LGG data set. On the other hand, none of the deep learning-based methods performs well on the OV data set. Both MDBM and dMCCA provide poor performance on Handwritten, Caltech, and NW-OBJECT data sets due to the over training of these models. The execution time required for the deep learning-based algorithms is also significantly higher as compared to that of the proposed algorithm.

All the results, reported here, establish the effectiveness of the proposed incremental multiblock data integration algorithm over the state-of-the-art data integration approaches. The better performance of the proposed SeFGeIM algorithm is achieved due to the following facts.

1. Instead of considering all the given modalities, the proposed algorithm considers only relevant ones to extract significant and relevant features.

2. The features are extracted sequentially based on the supervised information of sample categories.

3. The theory of rough hypercuboid approach is used to evaluate the quality of an extracted feature, in terms of its relevance and significance. It helps to address the uncertainty associated with real-life multimodal data.

In effect, a desired set of significant and relevant features is being generated sequentially, by incrementally incorporating relevant modalities, using the proposed IMCCA based feature extraction algorithm.

6.5 Conclusion

The main contribution of this chapter is three-fold, namely,

1. introduction of a novel MCCA, termed as incremental MCCA (IMCCA), which can update its solutions adaptively wherever a new modality is available for the analysis;
2. development of a new feature selection algorithm, based on IMCCA, judiciously integrating the information of multiblock data sets; and
3. demonstrating its success in benchmark as well as multi-omics data analysis.

The proposed IMCCA model deals with the "curse of dimensionality" problem due to "large p-small n" characteristics of real-life multimodal data sets, by using ridge regression optimization technique with shrinkage estimation. The proposed feature extraction algorithm, based on the IMCCA model, considers a new modality for the analysis if it has relevant and significant information with respect to existing modalities. The quality of the extracted features depends on the supervised information of sample categories. Analytical formulation facilitates the generation of relevant and significant features from multiblock dynamic data sets with significantly lower computational costs. The effectiveness of the proposed IMCCA based algorithm, along with a comparison with other algorithms, has been demonstrated on several real-life multiblock data sets.

Both ReDMiCA algorithm presented in Chapter 5 and SeFGeIM algorithm presented in Chapter 6 are based on SUMCOR criterion. Moreover, both the algorithms do not exploit the geometry of the data set. On the other hand, the MAXVAR criterion has lower computational overhead as compared to the SUMCOR criterion. In this regard, a new supervised feature extraction algorithm, termed as GraDiM, is presented in next chapter, which integrates dynamic multi-view data sets by using MAXVAR criterion and the knowledge of the graph.

Chapter 7

Graph Discriminant Multiset CCA for Adaptive Multi-View Learning

7.1 Introduction

Over the last few years, a keen interest in using complementary data associated with a specific problem has been developed. Different data sources are likely to contain distinct and thus partly independent information. The integration of orthogonal attributes from a broad range of views is supposed to provide better predictions than any single view [128, 247]. In this context, there has been a growing interest to integrate multi-view data. As mentioned in Chapter 5 and Chapter 6, multiset canonical correlation analysis (MCCA) [110] is an effective method to study the inter-dependence among multiple views. The goal of MCCA can be achieved by optimizing several criteria. In [61, 100, 135, 262], several criteria have been studied to extend canonical correlation analysis (CCA) for three or more sets of views. The commonly used criterion, which is the natural extension of CCA, is the sum of correlations (SUMCOR), which finds a common structure in multiple views via imposing pairwise similarities between the canonical variables. Although SUMCOR is NP-hard [229], it has got maximum attention in recent years [131]. The algorithms developed in both Chapter 5 and Chapter 6 use the SUMCOR criterion. On the other hand, the maximum variance (MAXVAR) criterion provides a conceptually simple solution among different formulations of MCCA [135]. Thus, it has successfully received recognition in the last few years [43, 115]. Instead of considering the pairwise similarity between the canonical variables, one can seek a common latent representation, which has a minimum dissimilarity of all views. The MAXVAR reduces the number of constraints that are associated with SUMCOR to a single constraint. Thus, it provides a conceptually simple algebraic solution, which reduces the computational cost.

The geometry of the multi-view data can provide the structural information of the data set. This structural information or prior knowledge of the samples can be encoded by a graph. The incorporation of the geometrical information along with the categorical knowledge may escalate the performance of the algorithm. Recently, graph-aware regularizers have manifested propitious achievement across a span of machine learning applications, such as dimensionality reduction, data reconstruction, clustering, and classification [44, 125, 237]. However, the SUMCOR framework of MCCA does not utilize the geometry of the multi-view data which may be available a priori or can be established using a certain domain of knowledge. This prior knowledge can be encoded by a graph, and be invoked as a regularizer to enrich the MAXVAR framework of MCCA. In this context, CCA with structural information induced by a graph has been reported in [44], but it is limited to analyzing two views only. In [43], a novel graph-regularized MCCA (GMCCA) algorithm has been proposed, to minimize the dissimilarity among the views based on the MAXVAR criterion.

As mentioned in Chapter 6, the brain-computer interface [161] to imaging genomics [155] involve either non-stationary or big data sets; either new instances may be added to the existing samples or new views may be considered for better analysis. Thus, the algorithms which integrate multi-view data should be adaptive or incremental in nature. The algorithms presented in [277, 310, 319] are applicable to the situation when new instances are being added with the existing samples and all the covariance matrices are required to update. However, they are not applicable when a new view is available for the augmentation with existing views. Recently, incremental generalized CCA has been proposed in [187] for incremental updates of existing solutions based on new modalities, although it leads to approximate solutions. Moreover, all the adaptive CCA algorithms reported in [187, 277, 310, 319] are unsupervised in nature. Both [319] and [310] integrate two views. On the other hand, the CCA generalization considered in [277] is equivalent to the MAXVAR criterion proposed in [135].

In this regard, the chapter introduces a new feature extraction algorithm, termed as GraDiM (Graph Discriminant Multi-View CCA), for multi-view data analysis. It incorporates the geometrical knowledge along with the categorical information of the data set using the MAXVAR criterion. The proposed algorithm is dynamic in nature, that is, it incrementally updates the existing solutions, whenever a new view is available for the analysis. On the other hand, the algorithm is designed in such a way that if all the views are present at the beginning of the data analysis, the algorithm starts with the two most relevant modalities, and the remaining modalities are added sequentially according to their relevance. The proposed algorithm addresses the singularity issue of the covariance matrices by using the ridge regression optimization technique. The optimum regularization parameters for the proposed algorithm are estimated based on the supervised information of sample categories. An analytical formulation demonstrates that the proposed algorithm can generate the required number of relevant and significant features from multi-view dynamic data sets, without extracting all possible features. In fact, all the views may not be required to extract different features. If the new view has relevant and significant information with respect to earlier views, then only the new view is incorporated in the integration process. The effectiveness of the proposed multi-view data integration algorithm, along with a comparative performance analysis with state-of-the-art algorithms, is established on several real-life multi-block data sets.

The rest of the chapter is organized as follows: Section 7.2 outlines the basic principle of GMCCA. Section 7.3 presents the proposed multi-view adaptive algorithm based on MAXVAR criterion. The effectiveness of the proposed multi-view data integration algorithm, along with a comparative performance analysis with state-of-the-art algorithms on different multi-view benchmark and omics data sets, is presented in Section 7.4. Concluding remarks is provided in Section 7.5.

7.2 Basics of MAXVAR Criterion and Graph-Regularized MCCA

This section presents the fundamental concepts in the theories of MCCA using the MAXVAR criterion and GMCCA. The objective of the MCCA is to extract the most correlated latent features from $\mathcal{M} \geqslant 3$ views of n samples, $\{\mathcal{X}_i \in \Re^{m_i \times n}\}_{i=1}^{\mathcal{M}}$, where m_i is the dimension of the i-th view. Without loss of generality, it is assumed that each multidimensional variable $\mathcal{X}_i$ is centered to have zero mean across the samples, that is, $\mathcal{E}[\mathcal{X}_i] = 0, \forall i \in \{1, 2, \cdots, \mathcal{M}\}$. The main objective of the MCCA is to find optimal basis vectors $\{\mathcal{W}_i \in \Re^{m_i \times p}\}_{i=1}^{\mathcal{M}}$ that maximize some merit functions under certain constraints, where $p = \min\{m_i, n\}$. Instead of considering the pairwise similarity between the canonical variables as considered in the SUMCOR criterion, one can seek a common latent representation, which has a minimum dissimilarity with all the multidimensional variables:

$$\min_{\{\mathcal{W}_i\}_{i=1}^{\mathcal{M}}, \mathcal{S}} \sum_{i=1}^{\mathcal{M}} \|\mathcal{W}_i{}^T \mathcal{X}_i - \mathcal{S}\|_F^2 = \max_{\{\mathcal{W}_i\}_{i=1}^{\mathcal{M}}, \mathcal{S}} \sum_{i=1}^{\mathcal{M}} \operatorname{Tr}\left(\mathcal{W}_i{}^T \mathcal{X}_i \mathcal{S}^T\right);$$

$$\text{subject to} \quad \mathcal{S}\mathcal{S}^T = \mathcal{I}; \tag{7.1}$$

where $\mathcal{S} \in \Re^{p \times n}$ is a common latent representation of the multidimensional variables; A^T and $\operatorname{Tr}(A)$ denote the transpose and trace of a matrix A, respectively, and $\mathcal{I}$ denotes the identity matrix with an appropriate order. The above approach is known as the MAXVAR criterion. From (7.2), it can be seen that the MAXVAR also finds highly correlated reduced-dimensional views as the SUMCOR does. Moreover, it reduces the number of constraints which are associated with the SUMCOR, $\mathcal{W}_i{}^T \mathcal{C}_{ii} \mathcal{W}_i = \mathcal{I}; \forall i \in \{1, 2, \cdots, \mathcal{M}\}$ to a single constraint $\mathcal{S}\mathcal{S}^T = \mathcal{I}$, where $\mathcal{C}_{ii} \in \Re^{m_i \times m_i}$ denotes the covariance matrix of $\mathcal{X}_i$. Thus, it provides a conceptually simple algebraic solution, which reduces the computational cost. Let us assume that $\mathcal{X}_i$ has full row rank and the solution of (7.2) with respect to $\mathcal{W}_i$ is

$$\mathcal{W}_i = \left[\mathcal{X}_i{}^\dagger\right]^T \mathcal{S}^T; \qquad \text{where} \quad \mathcal{X}_i{}^\dagger = \mathcal{X}_i{}^T \left(\mathcal{X}_i \mathcal{X}_i{}^T\right)^{-1}; \tag{7.2}$$

where $A^\dagger$ denotes the pseudoinverse of A. By substituting (7.2) from (7.2), an optimal solution $\mathcal{S}_{\text{opt}}$ can be obtained by solving the following problem:

$$\mathcal{S}_{\text{opt}} = \arg\max_{\mathcal{S}} \sum_{i=1}^{\mathcal{M}} \operatorname{Tr}\left(\mathcal{S}\mathcal{X}_i{}^\dagger \mathcal{X}_i \mathcal{S}^T\right)$$

$$= \arg\max_{\mathcal{S}} \operatorname{Tr}\left[\mathcal{S}\left(\sum_{i=1}^{\mathcal{M}} \mathcal{X}_i{}^\dagger \mathcal{X}_i\right)\mathcal{S}^T\right]; \text{ subject to } \mathcal{S}\mathcal{S}^T = \mathcal{I}. \tag{7.3}$$

The optimal solution $\mathcal{S}_{\text{opt}}$ is the first p principal eigenvectors of $\widehat{\mathcal{Y}} = \sum\limits_{i=1}^{\mathcal{M}} \mathcal{X}_i{}^\dagger \mathcal{X}_i \in \Re^{n \times n}$ [90].

In [43], the geometrical knowledge has been consolidated with the MAXVAR criterion to extract features from the multi-view data sets. Let the Laplacian matrix $\mathcal{L}_{\mathcal{G}_i}$ of a graph

$\mathcal{G}$ of the i-th view be defined as $\mathcal{L}_{\mathcal{G}_i} = \mathcal{D}_i - \mathcal{W}_i$, where $\mathcal{D}_i \in \Re^{n\times n}$ and $\mathcal{W}_i \in \Re^{n\times n}$ denote the degree matrix and the adjacency matrix, respectively. Now, the Laplacian matrix $\mathcal{L}_{\mathcal{G}}$, where all $\mathcal{M}$ views are considered, is determined as $\mathcal{L}_{\mathcal{G}} = \sum\limits_{i=1}^{\mathcal{M}} \mathcal{L}_{\mathcal{G}_i}$. Such additional geometrical information of the samples can be invoked as a regularizer in the MAXVAR criterion of the MCCA as follows:

$$\min_{\{\mathcal{W}_i\}_{i=1}^{\mathcal{M}},\mathcal{S}} \sum_{i=1}^{\mathcal{M}} \|\mathcal{W}_i^T \mathcal{X}_i - \mathcal{S}\|_F^2 + \gamma \mathrm{Tr}(\mathcal{S}\mathcal{L}_{\mathcal{G}}\mathcal{S}^T);$$

$$\text{subject to} \quad \mathcal{S}\mathcal{S}^T = \mathcal{I}; \tag{7.4}$$

where the coefficient $\gamma \geqslant 0$ trades off minimizing the distance between the canonical variables and smoothness of the samples over the graph $\mathcal{G}$. Specifically, when $\gamma = 0$, (7.2) reduces to the classical MCCA using the MAXVAR criterion in (7.2) and as γ increases, (7.2) relies more heavily on the extra graph knowledge to find the canonical variables. Let us consider each covariance matrix $\mathcal{C}_{ii}$ has full rank. Now, the partial derivative of (7.2) with respect to each $\mathcal{W}_i$, and setting the vectors of derivative to zero, we obtain the following equation:

$$\mathcal{S}_{\text{opt}} = \arg\max_{\mathcal{S}} \ \mathrm{Tr}\left[\mathcal{S}\left(\sum_{i=1}^{\mathcal{M}} \left(\mathcal{X}_i^{\dagger}\mathcal{X}_i - \gamma\mathcal{L}_{\mathcal{G}_i}\right)\right)\mathcal{S}^T\right];$$

$$\text{subject to} \quad \mathcal{S}\mathcal{S}^T = \mathcal{I}. \tag{7.5}$$

Similar to the standard MCCA using the MAXVAR criterion, the optimal solution $\mathcal{S}_{\text{opt}}$ be obtained by the p leading eigenvectors of the matrix $\mathcal{Y} = \sum\limits_{i=1}^{\mathcal{M}} \left(\mathcal{X}_i^{\dagger}\mathcal{X}_i - \gamma\mathcal{L}_{\mathcal{G}_i}\right)$.

In real-world applications, $n \ll m_i$, thus the computation of eigenvalue decomposition (EVD) of $\widehat{\mathcal{Y}}$ or $\mathcal{Y}$ is easier than the computation associated with the SUMCOR criterion. Besides, even if the multidimensional variable $\mathcal{X}_i$ is sparse, computing $\left(\mathcal{X}_i\mathcal{X}_i^T\right)^{-1}$ will create a large dense matrix and control the sparsity problem associated with real-life high dimensional multimodal data sets. But, $n \ll m_i$ makes the covariance matrix $\mathcal{C}_{ii} = \mathcal{X}_i\mathcal{X}_i^T$ non-invertible, which leads to the invalid computation of MCCA.

7.3 GraDiM: Proposed Multiset CCA

This section presents a new sequential feature extraction algorithm, which integrates the information of multi-view data sets that are available sequentially one after another. When a new view is available for the same set of samples, the proposed algorithm generates a new set of features based on the new view as well as the features extracted from the earlier views. It does not repeat the same procedure with the original data augmented by the new view. Moreover, the proposed algorithm encodes the geometrical information of the samples by a graph, and invokes this knowledge as a regularizer to boost the maximum variance MCCA framework. The prominence of the proposed algorithm can be established by some important analytical formulations, which are explained next.

7.3.1 Multiset Ridge Regression Model

Let $\{\mathcal{X}_i \in \Re^{m_i \times n}\}_{i=1}^{\mathcal{M}}$ be $\mathcal{M}$ multi-view data sets with m_i variables and n represents the number of samples. From (7.2), it is seen that the inverse of the covariance matrix $\mathcal{C}_{ii} \left(= \mathcal{X}_i\mathcal{X}_i^T\right)$ is needed to compute the basis vector $\mathcal{W}_i$; $\forall i \in \{1, 2, \cdots, \mathcal{M}\}$. If $n \ll m_i$, the covariance matrix $\mathcal{C}_{ii}$ becomes non-invertible, which leads to the invalid computation of MCCA [68]. To overcome this problem, a ridge regression optimization scheme is used by adding a small positive quantity $\mathfrak{r}_i$, known as regularization parameter, to the diagonals of the covariance matrix $\mathcal{C}_{ii}$. Let us assume that the ℓ-th dimension of the i-th multidimensional variable $\mathcal{X}_i[\ell]$ is contaminated with noise $\varepsilon_i[\ell]$, $\forall \ell \in \{1, 2, \cdots, m_i\}$ and $\forall i \in \{1, 2, \cdots, \mathcal{M}\}$, such that $\mathcal{E}\left[\varepsilon_i[\ell]\right] = 0$, $\mathcal{E}\left[\varepsilon_i[\ell]\varepsilon_i[\mathscr{k}]^T\right] = 0$ for $\ell \neq \mathscr{k}$, $\mathcal{E}\left[\varepsilon_i[\ell]\mathcal{X}_i[\ell]^T\right] = 0$ and $\mathcal{E}[\varepsilon_i[\ell]\varepsilon_i[\ell]^T] = \mathfrak{r}_i \geqslant 0$. Under these assumptions, the covariance matrix of $\mathcal{X}_i$ becomes $[\mathcal{C}_{ii} + \mathfrak{r}_i I]$. This modification is similar to ridge regression optimization [278]. The optimal set of regularization parameters can be estimated in such a way that the summation of the dissimilarity between each multidimensional variable with optimal subspace $\mathcal{S}_{\text{opt}}$ is minimum. To estimate the optimal set of regularization parameters, a grid search optimization is performed, where each regularization parameter $\mathfrak{r}_i$ follows an arithmetic progression and is varied within a specified range $[\mathfrak{r}_{min}, \mathfrak{r}_{max}]$. Let $\mathscr{d}_i$ be the common difference for regularization parameter $\mathfrak{r}_i$, while the parameter $\mathfrak{t}_i$ indicates the number of possible values of $\mathfrak{r}_i$. Hence, to compute the matrix $\mathcal{X}_i^{\dagger}$ of (7.2), the inverse of the corresponding covariance matrix has to be computed $\mathfrak{t}_i$ times, where

$$\mathcal{X}_{i\mathscr{k}_i}^{\dagger} = \mathcal{X}_i^T \left[\mathcal{C}_{ii} + (\mathfrak{r}_i + \mathscr{k}_i \mathscr{d}_i)I\right]^{-1} \tag{7.6}$$

$\forall \mathscr{k}_i \in \{0, 1, \cdots, (\mathfrak{t}_i - 1)\}$. According to [174], as the diagonal elements of $\mathcal{C}_{ii}$ are only changed by adding $\mathfrak{r}_i$, the eigenvalues of $[\mathcal{C}_{ii} + (\mathfrak{r}_i + \mathscr{k}_i\mathscr{d}_i)I]$ are changed, but the corresponding eigenvectors remain same. Also, there exists a relation between the eigenvalues of $[\mathcal{C}_{ii} + \mathfrak{r}_i I]$ and that of $[\mathcal{C}_{ii} + (\mathfrak{r}_i + \mathscr{k}_i\mathscr{d}_i)I]$, which is given by

$$\Delta_{i\mathscr{k}_i} = \Delta_i + \mathscr{k}_i\mathscr{d}_i I; \tag{7.7}$$

where $\Delta_{i\mathscr{k}_i}$ denotes a diagonal matrix, whose diagonal elements are the eigenvalues of $[\mathcal{C}_{ii} + (\mathfrak{r}_i + \mathscr{k}_i\mathscr{d}_i)I]$, $\Delta_i = \Delta_{i0}$ and the corresponding eigenvectors of $[\mathcal{C}_{ii} + (\mathfrak{r}_i + \mathscr{k}_i\mathscr{d}_i)I]$ be the columns of Ω_i. Based on the spectral decomposition, the covariance matrix $[\mathcal{C}_{ii} + (\mathfrak{r}_i + \mathscr{k}_i\mathscr{d}_i)I]$ and its inverse can be expressed as follows [269]:

$$[\mathcal{C}_{ii} + (\mathfrak{r}_i + \mathscr{k}_i\mathscr{d}_i)I] = \Omega_i\Delta_{i\mathscr{k}_i}\Omega_i^T = \Omega_i[\Delta_i + \mathscr{k}_i\mathscr{d}_i I]\Omega_i^T; \tag{7.8}$$

$$\text{and} \quad [\mathcal{C}_{ii} + (\mathfrak{r}_i + \mathscr{k}_i\mathscr{d}_i)I]^{-1} = \Omega_i[\Delta_i + \mathscr{k}_i\mathscr{d}_i I]^{-1}\Omega_i^T. \tag{7.9}$$

Now, the inverse of the diagonal matrix $[\Delta_i + \mathscr{k}_i\mathscr{d}_i I]$ can be computed as

$$[\Delta_i + \mathscr{k}_i\mathscr{d}_i I]^{-1} = \Delta_i^{-1} - \Delta_i^{-1}\mathscr{k}_i\mathscr{d}_i I\Delta_i^{-1}\left(I + \mathscr{k}_i\mathscr{d}_i I\Delta_i^{-1}\right)^{-1}$$

$$= \Delta_i^{-1} - \Delta_i^{-1} \mathcal{k}_i \mathcal{d}_i \Delta_i^{-1} \left(I + \mathcal{k}_i \mathcal{d}_i \Delta_i^{-1}\right)^{-1} = \Delta_i^{-1} - \Delta_i^{-1} \nabla_{i\mathcal{k}_i}; \tag{7.10}$$

$$\text{where} \quad \nabla_{i\mathcal{k}_i} = \mathcal{k}_i \mathcal{d}_i \Delta_i^{-1} \left(I + \mathcal{k}_i \mathcal{d}_i \Delta_i^{-1}\right)^{-1}. \tag{7.11}$$

Hence, using (7.9) and (7.3.1), the matrix ${\mathcal{X}_{i\mathcal{k}_i}}^{\dagger}$ of (7.6) becomes

$${\mathcal{X}_{i\mathcal{k}_i}}^{\dagger} = {\mathcal{X}_i}^T \Omega_i \left[\Delta_i^{-1} - \Delta_i^{-1} \nabla_{i\mathcal{k}_i}\right] \Omega_i^T = {\mathcal{X}_i}^T \Omega_i \Delta_i^{-1} \Omega_i^T - {\mathcal{X}_i}^T \Omega_i \Delta_i^{-1} \nabla_{i\mathcal{k}_i} \Omega_i^T = {\mathcal{X}_i}^{\dagger} - \mathcal{B}_{i\mathcal{k}_i}; \tag{7.12}$$

$$\text{where} \quad \mathcal{B}_{i\mathcal{k}_i} = {\mathcal{X}_i}^T \Omega_i \Delta_i^{-1} \nabla_{i\mathcal{k}_i} \Omega_i^T. \tag{7.13}$$

From (7.12), it is clear that ${\mathcal{X}_{i\mathcal{k}_i}}^{\dagger}$ is dependent on ${\mathcal{X}_i}^{\dagger}$. If the eigenvalues and the corresponding eigenvectors of $\mathcal{C}_{ii}$ are determined to compute ${\mathcal{X}_i}^{\dagger}$, there is no need to calculate the eigenvalues and the corresponding eigenvectors of the covariance matrix associated with other regularization parameters for each multidimensional variable. As the matrix ${\mathcal{X}_{i\mathcal{k}_i}}^{\dagger}$ has to be computed $\mathfrak{t}_i$ times for each multidimensional variables, the total number of all possible combination of regularization parameters is $\mathcal{T} = \prod\limits_{\ell=1}^{\mathcal{M}} \mathfrak{t}_\ell$. Hence, the matrix $\mathcal{Y}$ has to be determined $\mathcal{T}$ times, which is given by

$$\mathcal{Y}_r = \sum_{i=1}^{\mathcal{M}} \left(\mathcal{X}_{ir}^{\dagger} \mathcal{X}_i - \gamma \mathcal{L}_{\mathcal{G}_i}\right) = \sum_{i=1}^{\mathcal{M}} \left(\left({\mathcal{X}_i}^{\dagger} - \mathcal{B}_{ir}\right) \mathcal{X}_i - \gamma \mathcal{L}_{\mathcal{G}_i}\right)$$

$$= \sum_{i=1}^{\mathcal{M}} {\mathcal{X}_i}^{\dagger} \mathcal{X}_i - \sum_{i=1}^{\mathcal{M}} \mathcal{B}_{ir} \mathcal{X}_i - \sum_{i=1}^{\mathcal{M}} \gamma \mathcal{L}_{\mathcal{G}_i} = \mathcal{Y} - \sum_{i=1}^{\mathcal{M}} \mathcal{B}_{ir} \mathcal{X}_i; \tag{7.14}$$

$\forall r \in \{1, 2, \cdots, \mathcal{T}\}$. Hence, from (7.3.1), it is evident that each $\mathcal{Y}_r$ matrix can be computed with the help of the matrix $\mathcal{Y}$. The p principal eigenvectors of $\mathcal{Y}_r$ is the optimal solution $\mathcal{S}_{\text{opt}_r}$ corresponding to the r-th combination of regularization parameters.

7.3.2 Sequential Generation of Canonical Variables

In real-world high dimensional multi-view data analysis, the value of p is large. So, the computation of all p principal eigenvectors of $\mathcal{Y}_r$ is computationally expensive. In real-life applications, a small number of extracted features is typically effective to perform a certain task. So, instead of computing all possible eigenvectors of $\mathcal{Y}_r$, if eigenvectors are evaluated sequentially, then only the required number of features can be extracted. In this regard, the Power method [90] is used to approximate the dominant eigenvalue and corresponding eigenvector of the matrix $\mathcal{Y}_r$, while other eigenvalues associated with the corresponding eigenvectors can be approximated based on the Deflation method [293]. Let the t-th eigenvalue and corresponding eigenvector of $\mathcal{Y}_r$ be $\rho(t)$ and $\mathcal{S}_{\text{opt}_r}(t)$, respectively, associated with the r-th combination of regularization parameters, which is the dominant eigenvalue-eigenvector pair of the matrix $\mathcal{Y}_r(t)$. Then

$$\mathcal{Y}_r(t) \mathcal{S}_{\text{opt}_r}(t) = \rho(t) \mathcal{S}_{\text{opt}_r}(t); \tag{7.15}$$

where $t < p$. Now, using Deflation method, the $(t+1)$-th eigenvalue-eigenvector pair of the matrix $\mathcal{Y}_r$ will be the dominant eigenvalue and corresponding eigenvector of the matrix $\mathcal{Y}_r(t+1)$, where

$$\mathcal{Y}_r(t+1) = \mathcal{Y}_r(t) - \rho(t)\mathcal{S}_{\text{opt}_r}(t)\left[\mathcal{S}_{\text{opt}_r}(t)\right]^T = \mathcal{Y}_r(1) - \sum_{\ell=1}^{t}\rho(\ell)\mathcal{S}_{\text{opt}_r}(\ell)\left[\mathcal{S}_{\text{opt}_r}(\ell)\right]^T. \quad (7.16)$$

As $\mathcal{Y}_r(1)$ is nothing but $\mathcal{Y}_r$, (7.16) can be redefined by using (7.3.1), which is expressed as follows:

$$\mathcal{Y}_r(t+1) = \mathcal{Y} - \sum_{i=1}^{\mathcal{M}}\mathcal{B}_{ir}\mathcal{X}_i - \sum_{\ell=1}^{t}\rho(\ell)\mathcal{S}_{\text{opt}_r}(\ell)\left[\mathcal{S}_{\text{opt}_r}(\ell)\right]^T. \quad (7.17)$$

From (7.17), it is clear that every $(t+1)$-th eigenvalue-eigenvector pair of the matrix $\mathcal{Y}_r$, corresponding to each r-th regularization parameter, can be determined from all previously computed eigenvalue-eigenvector pairs of the matrix $\mathcal{Y}_r$. Thus, the required number of features is extracted sequentially using (7.17).

7.3.3 Multi-View Incremental Data Analysis

In real-world applications, a huge amount of data is being added to the existing databases continuously. Hence, the algorithm has to be adaptive or incremental in nature to solve certain problems associated with such a dynamic database. Moreover, all the available views of real-life data sets may not be relevant. Some of them may provide noisy or even inconsistent information with respect to other views. So, it is necessary to evaluate the quality of a new view before considering it for feature extraction. In this regard, the proposed algorithm not only integrates dynamic data sets, but also evaluates the relevance of each multidimensional variable before the integration. The proposed algorithm considers a new multidimensional variable if it has relevant and significant information with respect to earlier views. When a new view is available for the same set of samples, the proposed algorithm generates a new set of features based on the new view as well as the features extracted from the earlier views. It does not repeat the same procedure with the original data augmented by the new data.

Let $\mathcal{X}_i \in \Re^{m_i \times n}$, $\forall i \in \{1, 2, \cdots, \mathcal{M}\}$ be $\mathcal{M}$ multidimensional data sets with m_i variables and n number of samples. Suppose the optimal solution $\mathcal{S}_{\text{opt}}^{\mathcal{M}}$ is the set of first p principal eigenvectors of $\mathcal{Y}^{\mathcal{M}}$, where $\mathcal{Y}^{\mathcal{M}} = \sum_{i=1}^{\mathcal{M}}\left(\mathcal{X}_i{}^{\dagger}\mathcal{X}_i - \gamma\mathcal{L}_{\mathcal{G}_i}\right)$. Let $\Lambda^{\mathcal{M}}$ be the diagonal matrix, whose diagonal elements are the eigenvalues of $\mathcal{Y}^{\mathcal{M}}$. So,

$$\mathcal{Y}^{\mathcal{M}} = \mathcal{S}_{\text{opt}}^{\mathcal{M}}\Lambda^{\mathcal{M}}\left[\mathcal{S}_{\text{opt}}^{\mathcal{M}}\right]^T. \quad (7.18)$$

Each basis vector $\mathcal{W}_i^{\mathcal{M}}$ of $\mathcal{X}_i$ can be computed with the help of (7.2). Here, superscript $\mathcal{M}$ of $\mathcal{S}_{\text{opt}}^{\mathcal{M}}$, $\mathcal{Y}^{\mathcal{M}}$, $\Lambda^{\mathcal{M}}$, and $\mathcal{W}_i^{\mathcal{M}}$ denotes that all $\mathcal{M}$ multidimensional variables are considered to determine the solution. When a new view $\mathcal{X}_{(\mathcal{M}+1)}$ is added with the existing views, the basis vector $\mathcal{W}_{(\mathcal{M}+1)}^{\mathcal{M}+1}$ associated with $\mathcal{X}_{(\mathcal{M}+1)}$ has to be determined and the previous basis vectors $\mathcal{W}_i^{\mathcal{M}}$ of the preceding multidimensional variables $\mathcal{X}_i$ have to be updated by using $\mathcal{S}_{\text{opt}}^{\mathcal{M}}$ and $\Lambda^{\mathcal{M}}$. Let the updated basis vectors of the preceding multidimensional variables $\mathcal{X}_i$ be $\mathcal{W}_i^{\mathcal{M}+1}$. As $\mathcal{X}_{(\mathcal{M}+1)}$ is added, $\mathcal{Y}^{\mathcal{M}+1}$ will be as follows:

$$\mathcal{Y}^{\mathcal{M}+1} = \sum_{i=1}^{\mathcal{M}+1} \left(\mathcal{X}_i{}^{\dagger}\mathcal{X}_i - \gamma\mathcal{L}_{\mathcal{G}_i}\right) = \mathcal{Y}^{\mathcal{M}} + \mathcal{X}_{(\mathcal{M}+1)}^{\dagger}\mathcal{X}_{(\mathcal{M}+1)} - \gamma\mathcal{L}_{\mathcal{G}_{(\mathcal{M}+1)}}$$

$$= \mathcal{S}_{\text{opt}}^{\mathcal{M}}\Lambda^{\mathcal{M}}\left[\mathcal{S}_{\text{opt}}^{\mathcal{M}}\right]^T + \mathcal{X}_{(\mathcal{M}+1)}^{\dagger}\mathcal{X}_{(\mathcal{M}+1)} - \gamma\mathcal{L}_{\mathcal{G}_{(\mathcal{M}+1)}}. \tag{7.19}$$

Now, the optimal solution $\mathcal{S}_{\text{opt}}^{\mathcal{M}+1}$ is the first p principal eigenvectors of $\mathcal{Y}^{\mathcal{M}+1}$, where $p = \min\{m_i, n\}$, $\forall i \in \{1, 2, \cdots, (\mathcal{M}+1)\}$.

To get rid of the singularity issue associated with covariance matrix $\mathcal{C}_{ii}$, the ridge regression optimization scheme is used by adding regularization parameter $\mathfrak{r}_i$ to the diagonals of the matrix $\mathcal{C}_{ii}$. As the views are added sequentially, the optimal combination of the regularization parameter set corresponding to the preceding multidimensional variables are already evaluated. The optimal regularization parameter associated with the new view has to be determined only. Let the regularization parameter $\mathfrak{r}_{(\mathcal{M}+1)}$, corresponding to $\mathcal{X}_{(\mathcal{M}+1)}$, be varied within a specified range $[\mathfrak{r}_{min}, \mathfrak{r}_{max}]$. Let $d_{(\mathcal{M}+1)}$ be the common difference for regularization parameter $\mathfrak{r}_{(\mathcal{M}+1)}$, while the parameter $\mathfrak{t}_{(\mathcal{M}+1)}$ indicates the number of possible values of $\mathfrak{r}_{(\mathcal{M}+1)}$. Hence, the matrix $\mathcal{X}_{(\mathcal{M}+1)k_{(\mathcal{M}+1)}}^{\dagger}$, $\forall k_{(\mathcal{M}+1)} \in \{0, 1, \cdots, (\mathfrak{t}_{(\mathcal{M}+1)} - 1)\}$ can be computed using (7.12). Now, combining (7.12) and (7.3.3), the matrix $\mathcal{Y}^{\mathcal{M}+1}$ has to be determined $\mathfrak{t}_{(\mathcal{M}+1)}$ times, which is given by

$$\mathcal{Y}_r^{\mathcal{M}+1} = \mathcal{S}_{\text{opt}}^{\mathcal{M}}\Lambda^{\mathcal{M}}\left[\mathcal{S}_{\text{opt}}^{\mathcal{M}}\right]^T + \left(\mathcal{X}_{(\mathcal{M}+1)}^{\dagger} - \mathcal{B}_{(\mathcal{M}+1)r}\right)\mathcal{X}_{(\mathcal{M}+1)} - \gamma\mathcal{L}_{\mathcal{G}_{(\mathcal{M}+1)}}$$

$$= \mathcal{S}_{\text{opt}}^{\mathcal{M}}\Lambda^{\mathcal{M}}\left[\mathcal{S}_{\text{opt}}^{\mathcal{M}}\right]^T + \mathcal{X}_{(\mathcal{M}+1)}^{\dagger}\mathcal{X}_{(\mathcal{M}+1)} - \mathcal{B}_{(\mathcal{M}+1)r}\mathcal{X}_{(\mathcal{M}+1)} - \gamma\mathcal{L}_{\mathcal{G}_{(\mathcal{M}+1)}}; \tag{7.20}$$

$\forall r \in \{1, 2, \cdots, \mathfrak{t}_{(\mathcal{M}+1)}\}$, where $\mathcal{B}_{(\mathcal{M}+1)r}$ can be computed by using (7.13). Let the t-th eigenvalue and corresponding eigenvector of $\mathcal{Y}_r^{\mathcal{M}+1}$ be $\rho(t)^{\mathcal{M}+1}$ and $\mathcal{S}_{\text{opt}_r}^{\mathcal{M}+1}(t)$, respectively, associated with the r-th regularization parameter, which together represent the dominant eigenvalue-eigenvector pair of the matrix $\mathcal{Y}_r(t)^{\mathcal{M}+1}$, where $t < p$. Now, using the Deflation method, the $(t+1)$-th eigenvalue-eigenvector pair of the matrix $\mathcal{Y}_r^{\mathcal{M}+1}$ will be the dominant eigenvalue and corresponding eigenvector of the matrix $\mathcal{Y}_r(t+1)^{\mathcal{M}+1}$, where

$$\mathcal{Y}_r(t+1)^{\mathcal{M}+1} = \mathcal{Y}_r(1)^{\mathcal{M}+1} - \sum_{\ell=1}^{t} \rho(\ell)^{\mathcal{M}+1}\mathcal{S}_{\text{opt}_r}^{\mathcal{M}+1}(\ell)\left[\mathcal{S}_{\text{opt}_r}^{\mathcal{M}+1}(\ell)\right]^T$$

$$= \mathcal{S}_{\text{opt}}^{\mathcal{M}}\Lambda^{\mathcal{M}}\left[\mathcal{S}_{\text{opt}}^{\mathcal{M}}\right]^T + \mathcal{X}_{(\mathcal{M}+1)}^{\dagger}\mathcal{X}_{(\mathcal{M}+1)} - \mathcal{B}_{(\mathcal{M}+1)k_{(\mathcal{M}+1)}}\mathcal{X}_{(\mathcal{M}+1)}$$

$$-\gamma\mathcal{L}_{\mathcal{G}_{(\mathcal{M}+1)}} - \sum_{\ell=1}^{t} \rho(\ell)^{\mathcal{M}+1}\mathcal{S}_{\text{opt}_r}^{\mathcal{M}+1}(\ell)\left[\mathcal{S}_{\text{opt}_r}^{\mathcal{M}+1}(\ell)\right]^T. \tag{7.21}$$

From (7.3.3), it is clear that every $(t+1)$-th eigenvalue-eigenvector pair of the matrix $\mathcal{Y}_r^{\mathcal{M}+1}$ corresponding to each r-th regularization parameter can be determined by the help of all t number of previously computed eigenvalue-eigenvector pair of the matrix $\mathcal{Y}_r^{\mathcal{M}+1}$. The basis vectors of the preceding multidimensional variables can be updated as

$$\mathcal{W}_{ir}^{\mathcal{M}+1}(t) = \left[\mathcal{X}_{i\tilde{r}_i}^{\dagger}\right]^T \mathcal{S}_{\text{opt}_r}^{\mathcal{M}+1}(t); \tag{7.22}$$

where $\tilde{r}_i$ denotes the optimal regularization parameter selected for the i-th preceding multidimensional variable, $\forall i \in \{1, 2, \cdots, \mathcal{M}\}$. On the other hand, the basis vector of the $(\mathcal{M}+1)$-th multidimensional variable can be determined as

$$\mathcal{W}_{(\mathcal{M}+1)r}^{\mathcal{M}+1}(t) = \left[\mathcal{X}_{(\mathcal{M}+1)r}^{\dagger}\right]^T \mathcal{S}_{\text{opt}_r}^{\mathcal{M}+1}(t). \tag{7.23}$$

The canonical variables can be computed as

$$\mathcal{U}_{ir}^{\mathcal{M}+1}(t) = \left[\mathcal{W}_{ir}^{\mathcal{M}+1}(t)\right]^T \mathcal{X}_i; \tag{7.24}$$

$\forall i \in \{1, 2, \cdots, (\mathcal{M}+1)\}$, for each r-th regularization parameter. Finally, the t-th feature is generated as

$$\mathcal{F}_r^{\mathcal{M}+1}(t) = \sum_{i=1}^{\mathcal{M}+1} \mathcal{U}_{ir}^{\mathcal{M}+1}(t). \tag{7.25}$$

Thus, all the features are extracted sequentially without repeating the same steps with the original data augmented by the new view. The algorithm to compute the basis vector $\mathcal{W}_{(i+1)r}^{i+1}(t)$ of newly added view $\mathcal{X}_{(i+1)}$ for each r-th regularization parameter and to update the basis vectors of all preceding multidimensional variables is presented in Algorithm 7.1.

The proposed algorithm is designed in such a way that it considers the relevance of multidimensional variables while adding them sequentially. The largest eigenvalue of the covariance matrix of a view represents its relevance value, as the dominant eigenvalue indicates the variance and the corresponding eigenvector denotes the direction of the largest spread of the data. So, if the data is projected towards the direction of the largest spread or first principal axis, then the eigenvalue of the covariance matrix represents the variance of the data in that direction. As both eigenvalues and eigenvectors of each covariance matrix have to be computed to calculate the inverse of the covariance matrix for each multidimensional variable, the multidimensional variables are arranged, in descending order, according to their largest eigenvalues of covariance matrices. Let $\{\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_{(\mathcal{M}+1)}\}$ be the ordered list of $(\mathcal{M}+1)$ multidimensional data sets. Let us assume that $\mathbb{S}^i$ be the set of $\mathcal{D}(\leqslant p)$ selected features where all i multidimensional variables are considered, $\forall i \in \{1, 2, \cdots, (\mathcal{M}+1)\}$ and initially $\mathbb{S}^i \leftarrow \emptyset$. Let us consider that the set $\mathbb{C}^i$ contains the t-th extracted features which are computed by using all r-th combinations of regularization parameters, $\forall t \in \{1, 2, \cdots, p\}$ where all i multidimensional variables are considered. The relevance of the feature $\mathcal{F}_r^i(t)$ with respect to the sample categories $\mathbb{D}$ is denoted by $\gamma_{\mathcal{F}_r^i(t)}(\mathbb{D})$. Let $\sigma_{\{\mathcal{F}_r^i(t), \mathcal{F}_\ell^i\}}(\mathbb{D}, \mathcal{F}_r^i(t))$ denote the significance of the feature $\mathcal{F}_r^i(t)$ with respect to already-selected feature $\mathcal{F}_\ell^i \in \mathbb{S}^i$. For $t = 1$, the most relevant feature is selected from the set $\mathbb{C}^i$ and is included to $\mathbb{S}^i$, that is,

$$\mathcal{F}^i(t) = \underset{\mathcal{F}^i_r(t) \in \mathbb{C}^i}{\arg\max} \left\{ \gamma_{\mathcal{F}^i_r(t)}(\mathbb{D}) \right\}; \tag{7.26}$$

while for $t > 1$, the feature which has maximum relevance among the features of $\mathbb{C}^i$ and significance with respect to the features of $\mathbb{S}^i$ is selected as follows:

$$\mathcal{F}^i(t) = \underset{\mathcal{F}^i_r(t) \in \mathbb{C}^i}{\arg\max} \left\{ \gamma_{\mathcal{F}^i_r(t)}(\mathbb{D}) + \frac{1}{t-1} \sum_{\mathcal{F}^i_\ell \in \mathbb{S}^i} \sigma_{\{\mathcal{F}^i_r(t), \mathcal{F}^i_\ell\}}(\mathbb{D}, \mathcal{F}^i_r(t)) \right\}. \tag{7.27}$$

The proposed algorithm starts with first three multidimensional variables $\mathcal{X}_1$,$\mathcal{X}_2$, and $\mathcal{X}_3$ from the ordered list, and produces a feature set $\mathbb{S}^3$. Then, other views come sequentially one after another. If the t-th feature of $\mathbb{S}^i$ has higher relevance and significance value than that of $\mathbb{S}^{i+1}$, $\forall i \in \{3, 4, \cdots, \mathcal{M}\}$, then the t-th feature of $\mathbb{S}^i$ is considered instead of $\mathbb{S}^{i+1}$. So, if a multidimensional variable $\mathcal{X}_{(i+1)}$ is relevant in extracting the t-th feature, then only it is considered, otherwise, the multidimensional variables $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_i$ are integrated to extract the t-th feature. So, each feature is extracted by integrating different number of multidimensional variables. The problem of generating a set of most significant and relevant feature set $\mathbb{S}^{\mathcal{M}+1}$ from the selected multi-view data sets is addressed by a greedy algorithm which is reported in Algorithm 7.2. In the current research work, both significance and relevance of an extracted feature are computed by using the concept of rough hypercuboid approach [172], which is described in Section 3.3.3 of Chapter 3. It helps to optimize the regularization parameters.

Algorithm 7.1 Proposed Dynamic Fusion Algorithm.

Input: The $(i+1)$-th multidimensional variable $\mathcal{X}_{(i+1)}$ and the optimal solution $\{\mathcal{S}^i_{\text{opt}}, \Lambda^i\}$ of preceding i multidimensional variables.

Output: The t-th basis vectors $\mathcal{W}^{i+1}_{kr}(t)$ of all $(i+1)$-th multidimensional variables corresponding to r-th regularization parameter, $\forall k \in \{1, 2, \cdots, (i+1)\}$.

```
for each r-th regularization parameter, where $\forall r \in \{1, 2, \cdots, \mathtt{t}_{(i+1)}\}$ for each t-th extracted feature do
    (i) Calculate $\mathcal{Y}_r(t)^{i+1}$ using (7.3.3) if $t = 1$, otherwise using (7.3.3).
    (ii) Compute the dominant eigenvalue $\rho(t)^{i+1}$ and corresponding eigenvector $\mathcal{S}^{i+1}_{\text{opt}_r}(t)$ of $\mathcal{Y}_r(t)^{i+1}$.
    (iii) Update the t-th basis vector $\mathcal{W}^{i+1}_{kr}(t)$ of all the $k$ preceding multidimensional variables using (7.22), where $\forall k \in \{1, 2, \cdots, i\}$.
    (iv) Determine the t-th basis vector $\mathcal{W}^{i+1}_{(i+1)r}(t)$ of $(i+1)$-th multidimensional variable using (7.23).
end for
```

Algorithm 7.2 GraDiM: Graph Discriminant Multiset CCA

```
Input: $(\mathcal{M}+1)$ multidimensional variables $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_{(\mathcal{M}+1)}$.
Output: A set $\mathbb{S}^{\mathcal{M}+1}$ of $\mathcal{D}$ selected features.
Compute the Laplacian matrix $\mathcal{L}_{\mathcal{G}_i} \in \Re^{n \times n}$ of $\mathcal{X}_i$, $\forall i \in \{1, 2, \cdots, (\mathcal{M}+1)\}$.
Calculate the covariance matrix $\mathcal{C}_{ii} \in \Re^{m_i \times m_i}$ of $\mathcal{X}_i$, $\forall i \in \{1, 2, \cdots, (\mathcal{M}+1)\}$.
Determine the eigenvalues $\delta_{i\ell}$ of $\mathcal{C}_{ii}$, along with the corresponding eigenvectors $\omega_{i\ell}$,
    $\forall i \in \{1, 2, \cdots, (\mathcal{M}+1)\}$, $\forall \ell \in \{1, 2, \cdots, m_i\}$.
Construct the diagonal matrix $\Delta_i \in \Re^{m_i \times m_i}$, whose diagonal elements are $\delta_{i\ell}$, and the
    square matrix $\Omega_i \in \Re^{m_i \times m_i}$, whose each column is $\omega_{i\ell}$.
Rearrange the multidimensional variables according their largest eigenvalues of covariance matrices. Let $\{\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_{(\mathcal{M}+1)}\}$ be the order list.
For each $i = 3, \cdots, (\mathcal{M}+1)$ do
    (I) Initialize $\mathbb{S}^i = \varnothing$ and $t = 1$.
    (II) For each $j \leqslant p$ where $p = \min\{m_l, n\}$, $\forall \ell \in \{1, \cdots, i\}$ do
        (i) Initialize $\mathbb{C}^i = \varnothing$.
        (ii) For each $r$-th combinations of regularization parameters, where $\forall r \in \{1, 2, \cdots, \prod_{\ell=1}^{i} \mathfrak{t}_\ell\}$ if $i = 3$; otherwise, for each $r$-th regularization parameter, where $\forall r \in \{1, 2, \cdots, \mathfrak{t}_i\}$ do
            (a) When $i = 3$, compute $\mathcal{Y}_r(t)^i$ using (7.3.1) if $t = 1$; otherwise, using (7.17).
            (b) Determine the $t$-th basis vector $\mathcal{W}^i_{kr}(t)$ of $k$-th multidimensional variable using (7.23) if $i = 3$; otherwise, call the Algorithm 7.1 to compute $\mathcal{W}^i_{kr}(t)$, $\forall k \in \{1, \cdots, i\}$.
            (c) Calculate the $t$-th canonical variable $\mathcal{U}^i_{kr}(t)$; $\forall k \in \{1, 2, \cdots, i\}$ using (7.24).
            (d) Extract the $t$-th feature $\mathcal{F}^i_r(t)$ using (7.25).
            (e) Compute the relevance $\gamma_{\mathcal{F}^i_r(t)}(\mathbb{D})$ of the feature $\mathcal{F}^i_r(t)$.
            (f) If $t > 1$, determine the significance $\sigma_{\{\mathcal{F}^i_r(t), \mathcal{F}^i_\ell\}}(\mathbb{D}, \mathcal{F}^i_r(t))$ of the extracted feature $\mathcal{F}^i_r(t)$.
            (g) Add $\mathcal{F}^i_r(t)$ to $\mathbb{C}^i$ if its significance is non-zero with respect to all the selected features of $\mathbb{S}^i$. In effect, $\mathbb{C}^i = \mathbb{C}^i \bigcup \mathcal{F}^i_r(t)$.
        (iii) end for
        (iv) If $\mathbb{C}^i \neq \varnothing$, select a feature as $t$-th feature $\mathcal{F}^i_r(t)$ from all the features of $\mathbb{C}^i$, which maximizes the condition (7.26) when $t = 1$, otherwise (7.27). As a result of that, $\mathbb{S}^i = \mathbb{S}^i \bigcup \mathcal{F}^i_r(t)$.
        (v) For $i > 3$, if the value of objective function ((7.26) for $t = 1$ and (7.27) otherwise) of the $t$-th feature of $\mathbb{S}^{i-1}$ is greater than that of $\mathbb{S}^i$, then $\mathbb{S}^i = \mathbb{S}^i \bigcup \mathcal{F}^{i-1}_r(t)$.
        (vi) Set $t = t + 1$.
    (III) end for
end for
Stop.
```

7.3.4 Validation of Proposed Algorithm

When a new view $\mathcal{X}_{(\mathcal{M}+1)}$ is available for analysis, the proposed GraDiM algorithm does not repeat the same steps with the previous views $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_{\mathcal{M}}$ augmented by the new view $\mathcal{X}_{(\mathcal{M}+1)}$. Rather, it starts with the results obtained using the previous set of views and generates the new solution. In fact, if the initial set of views and the new view come together, then GraDiM produces the same set of basis vectors. Without loss of generality, let us consider that initially $\mathcal{M}$ multidimensional variables are present. The optimal solution $\mathcal{S}_{\text{opt}}^{\mathcal{M}}$ is the eigenvector of $\mathcal{Y}^{\mathcal{M}}$, where eigenvalues are placed in the diagonal positions of the diagonal matrix $\Lambda^{\mathcal{M}}$. Thus, $\mathcal{Y}^{\mathcal{M}}$ can be expressed as (7.18).

Suppose, a new multidimensional variable $\mathcal{X}_{(\mathcal{M}+1)}$ has come. According to GraDiM, the results of the previous $\mathcal{M}$ views, that is, $\Lambda^{\mathcal{M}}$ and $\mathcal{S}_{\text{opt}}^{\mathcal{M}}$ have to be used to generate the new results. The optimal solution $\mathcal{S}_{\text{opt}}^{\mathcal{M}+1}$ is the eigenvectors of $\mathcal{Y}^{\mathcal{M}+1}$, where

$$\mathcal{Y}^{\mathcal{M}+1} = \sum_{i=1}^{\mathcal{M}+1} \left({\mathcal{X}_i}^{\dagger} \mathcal{X}_i - \gamma \mathcal{L}_{\mathcal{G}_i} \right). \tag{7.28}$$

Let the t-th eigenvalue and corresponding eigenvector of $\mathcal{Y}^{\mathcal{M}+1}$ be $\rho(t)^{\mathcal{M}+1}$ and $\mathcal{S}_{\text{opt}}^{\mathcal{M}+1}(t)$, respectively, where $t < \min\{m_i, n\}$, $\forall i \in \{1, 2, \cdots, (\mathcal{M}+1)\}$. Hence, let the dominant eigenvalue-eigenvector pair of $\mathcal{Y}^{\mathcal{M}+1}$ be $\rho(1)^{\mathcal{M}+1}$ and $\mathcal{S}_{\text{opt}}^{\mathcal{M}+1}(1)$, respectively,

$$\mathcal{Y}^{\mathcal{M}+1} = \rho(1)^{\mathcal{M}+1} \mathcal{S}_{\text{opt}}^{\mathcal{M}+1}(1) \left[\mathcal{S}_{\text{opt}}^{\mathcal{M}+1}(1) \right]^T. \tag{7.29}$$

On the other hand, using GraDiM, the eigenvector of the matrix $\widehat{\mathcal{Y}}^{\mathcal{M}+1}$ is the optimal solution $\widehat{\mathcal{S}}_{\text{opt}}^{\mathcal{M}+1}$, where

$$\widehat{\mathcal{Y}}^{\mathcal{M}+1} = \mathcal{S}_{\text{opt}}^{\mathcal{M}} \Lambda^{\mathcal{M}} \left[\mathcal{S}_{\text{opt}}^{\mathcal{M}} \right]^T + \mathcal{X}_{(\mathcal{M}+1)}^{\dagger} \mathcal{X}_{(\mathcal{M}+1)} - \gamma \mathcal{L}_{\mathcal{G}(\mathcal{M}+1)}. \tag{7.30}$$

Let the t-th eigenvalue-eigenvector pair of $\widehat{\mathcal{Y}}^{\mathcal{M}+1}$ be $\left(\widehat{\rho}(t)^{\mathcal{M}+1}, \widehat{\mathcal{S}}_{\text{opt}}^{\mathcal{M}+1}(t) \right)$. Thus, the dominant eigenvalue and corresponding eigenvector of $\widehat{\mathcal{Y}}^{\mathcal{M}+1}$ be $\widehat{\rho}(1)^{\mathcal{M}+1}$ and $\widehat{\mathcal{S}}_{\text{opt}}^{\mathcal{M}+1}(1)$, respectively,

$$\widehat{\mathcal{Y}}^{\mathcal{M}+1} = \widehat{\rho}(1)^{\mathcal{M}+1} \widehat{\mathcal{S}}_{\text{opt}}^{\mathcal{M}+1}(1) \left[\widehat{\mathcal{S}}_{\text{opt}}^{\mathcal{M}+1}(1) \right]^T. \tag{7.31}$$

To establish the characteristics, we need to show that the eigenvalue-eigenvector pairs $\left(\rho(t)^{\mathcal{M}+1}, \mathcal{S}_{\text{opt}}^{\mathcal{M}+1}(t) \right)$ and $\left(\widehat{\rho}(t)^{\mathcal{M}+1}, \widehat{\mathcal{S}}_{\text{opt}}^{\mathcal{M}+1}(t) \right)$ of $\mathcal{Y}^{\mathcal{M}+1}$ and $\widehat{\mathcal{Y}}^{\mathcal{M}+1}$, where $t < \min\{m_i, n\}$, $\forall i \in \{1, 2, \cdots, (\mathcal{M}+1)\}$ are the same. Mathematical induction is used to prove this affirmation.

Base step: Comparing the value of $\mathcal{Y}^{\mathcal{M}+1}$ of (7.3.3) and (7.29) and using the value of $\widehat{\mathcal{Y}}^{\mathcal{M}+1}$ of (7.31), we get,

$$\mathcal{S}_{\text{opt}}^{\mathcal{M}} \Lambda^{\mathcal{M}} \left[\mathcal{S}_{\text{opt}}^{\mathcal{M}} \right]^T + \mathcal{X}_{(\mathcal{M}+1)}^{\dagger} \mathcal{X}_{(\mathcal{M}+1)} - \gamma \mathcal{L}_{\mathcal{G}(\mathcal{M}+1)} = \rho(1)^{\mathcal{M}+1} \mathcal{S}_{\text{opt}}^{\mathcal{M}+1}(1) \left[\mathcal{S}_{\text{opt}}^{\mathcal{M}+1}(1) \right]^T$$

$$\Rightarrow \widehat{\mathcal{Y}}^{\mathcal{M}+1} = \rho(1)^{\mathcal{M}+1}\mathcal{S}_{\text{opt}}^{\mathcal{M}+1}(1)\left[\mathcal{S}_{\text{opt}}^{\mathcal{M}+1}(1)\right]^T. \tag{7.32}$$

From (7.3.4), it is clear that, $\rho(1)^{\mathcal{M}+1}$ and $\mathcal{S}_{\text{opt}}^{\mathcal{M}+1}(1)$ are the dominant eigenvalue and corresponding eigenvector of $\widehat{\mathcal{Y}}^{\mathcal{M}+1}$. As the dominant eigenvalue-eigenvector pair has to be unique, $\left(\rho(1)^{\mathcal{M}+1}, \mathcal{S}_{\text{opt}}^{\mathcal{M}+1}(1)\right)$ and $\left(\widehat{\rho}(1)^{\mathcal{M}+1}, \widehat{\mathcal{S}}_{\text{opt}}^{\mathcal{M}+1}(1)\right)$ are the same.

Inductive step: Let all t eigenvalues and the corresponding eigenvector pairs $\left(\rho(t)^{M+1}, \mathcal{S}_{\text{opt}}^{\mathcal{M}+1}(t)\right)$ and $\left(\widehat{\rho}(t)^{\mathcal{M}+1}, \widehat{\mathcal{S}}_{\text{opt}}^{\mathcal{M}+1}(t)\right)$ of $\mathcal{Y}^{\mathcal{M}+1}$ and $\widehat{\mathcal{Y}}^{\mathcal{M}+1}$, are same. We have to prove that the $(t+1)$-th eigenvalue-eigenvector pair $\left(\rho(t+1)^{\mathcal{M}+1}, \mathcal{S}_{\text{opt}}^{\mathcal{M}+1}(t+1)\right)$ and $\left(\widehat{\rho}(t+1)^{\mathcal{M}+1}, \widehat{\mathcal{S}}_{\text{opt}}^{\mathcal{M}+1}(t+1)\right)$ of $\mathcal{Y}^{\mathcal{M}+1}$ and $\widehat{\mathcal{Y}}^{\mathcal{M}+1}$ are same, where $(t+1) \leqslant \min\{m_i, n\}, \forall i \in \{1, 2, \cdots, (\mathcal{M}+1)\}$. Now, using Deflation method,

$$\mathcal{Y}(t+1)^{\mathcal{M}+1} = \mathcal{Y}(1)^{\mathcal{M}+1} - \sum_{\ell=1}^{t} \rho(\ell)^{\mathcal{M}+1}\mathcal{S}_{\text{opt}}^{\mathcal{M}+1}(\ell)\left[\mathcal{S}_{\text{opt}}^{\mathcal{M}+1}(\ell)\right]^T; \tag{7.33}$$

and

$$\begin{aligned}\widehat{\mathcal{Y}}(t+1)^{\mathcal{M}+1} &= \widehat{\mathcal{Y}}(1)^{\mathcal{M}+1} - \sum_{\ell=1}^{t} \widehat{\rho}(\ell)^{\mathcal{M}+1}\widehat{\mathcal{S}}_{\text{opt}}^{\mathcal{M}+1}(\ell)\left[\widehat{\mathcal{S}}_{\text{opt}}^{\mathcal{M}+1}(\ell)\right]^T \\ &= \widehat{\mathcal{Y}}(1)^{\mathcal{M}+1} - \sum_{\ell=1}^{t} \rho(\ell)^{\mathcal{M}+1}\mathcal{S}_{\text{opt}}^{\mathcal{M}+1}(\ell)\left[\mathcal{S}_{\text{opt}}^{\mathcal{M}+1}(\ell)\right]^T. \end{aligned} \tag{7.34}$$

As $\mathcal{Y}(1)^{\mathcal{M}+1}$ and $\widehat{\mathcal{Y}}(1)^{\mathcal{M}+1}$ are nothing but $\mathcal{Y}^{\mathcal{M}+1}$ and $\widehat{\mathcal{Y}}^{\mathcal{M}+1}$, respectively, (7.33) and (7.3.4) can be redefined by using (7.28) and (7.30), respectively, which are expressed as follows:

$$\mathcal{Y}(t+1)^{\mathcal{M}+1} = \sum_{i=1}^{\mathcal{M}+1}\left(\mathcal{X}_i{}^{\dagger}\mathcal{X}_i - \gamma\mathcal{L}_{\mathcal{G}_i}\right) - \sum_{\ell=1}^{t} \rho(\ell)^{\mathcal{M}+1}\mathcal{S}_{\text{opt}}^{\mathcal{M}+1}(\ell)\left[\mathcal{S}_{\text{opt}}^{\mathcal{M}+1}(\ell)\right]^T; \tag{7.35}$$

$$\widehat{\mathcal{Y}}(t+1)^{\mathcal{M}+1} = \mathcal{S}_{\text{opt}}^{\mathcal{M}}\Lambda^{\mathcal{M}}\left[\mathcal{S}_{\text{opt}}^{\mathcal{M}}\right]^T + \mathcal{X}_{(\mathcal{M}+1)}^{\dagger}\mathcal{X}_{(\mathcal{M}+1)} - \gamma\mathcal{L}_{\mathcal{G}_{(\mathcal{M}+1)}} - \sum_{\ell=1}^{t} \rho(\ell)^{\mathcal{M}+1}\mathcal{S}_{\text{opt}}^{\mathcal{M}+1}(\ell)\left[\mathcal{S}_{\text{opt}}^{\mathcal{M}+1}(\ell)\right]^T. \tag{7.36}$$

Now, from (7.35), we get

$$\begin{aligned}\mathcal{Y}(t+1)^{\mathcal{M}+1} &= \sum_{i=1}^{\mathcal{M}}\mathcal{X}_i{}^{\dagger}\mathcal{X}_i + \mathcal{X}_{(\mathcal{M}+1)}^{\dagger}\mathcal{X}_{(\mathcal{M}+1)} - \sum_{i=1}^{\mathcal{M}}\gamma\mathcal{L}_{\mathcal{G}_i} - \gamma\mathcal{L}_{\mathcal{G}_{(\mathcal{M}+1)}} - \sum_{\ell=1}^{t} \rho(\ell)^{\mathcal{M}+1}\mathcal{S}_{\text{opt}}^{\mathcal{M}+1}(\ell)\left[\mathcal{S}_{\text{opt}}^{\mathcal{M}+1}(\ell)\right]^T \\ &= \mathcal{Y}^{\mathcal{M}} + \mathcal{X}_{(\mathcal{M}+1)}^{\dagger}\mathcal{X}_{(\mathcal{M}+1)} - \gamma\mathcal{L}_{\mathcal{G}_{(\mathcal{M}+1)}} - \sum_{\ell=1}^{t} \rho(\ell)^{\mathcal{M}+1}\mathcal{S}_{\text{opt}}^{\mathcal{M}+1}(\ell)\left[\mathcal{S}_{\text{opt}}^{\mathcal{M}+1}(\ell)\right]^T \\ &= \mathcal{S}_{\text{opt}}^{\mathcal{M}}\Lambda^{\mathcal{M}}\left[\mathcal{S}_{\text{opt}}^{\mathcal{M}}\right]^T + \mathcal{X}_{(\mathcal{M}+1)}^{\dagger}\mathcal{X}_{(\mathcal{M}+1)} - \gamma\mathcal{L}_{\mathcal{G}_{(\mathcal{M}+1)}} - \sum_{\ell=1}^{t} \rho(\ell)^{\mathcal{M}+1}\mathcal{S}_{\text{opt}}^{\mathcal{M}+1}(\ell)\left[\mathcal{S}_{\text{opt}}^{\mathcal{M}+1}(\ell)\right]^T. \end{aligned} \tag{7.37}$$

Comparing (7.36) and (7.3.4), it is clear that

$$\mathcal{Y}(t+1)^{\mathcal{M}+1} = \widehat{\mathcal{Y}}(t+1)^{\mathcal{M}+1}. \tag{7.38}$$

Hence, the $(t+1)$-th eigenvalue-eigenvector pairs $\left(\rho(t+1)^{\mathcal{M}+1}, \mathcal{S}_{\text{opt}}^{\mathcal{M}+1}(t+1)\right)$ and $\left(\widehat{\rho}(t+1)^{\mathcal{M}+1}, \widehat{\mathcal{S}}_{\text{opt}}^{\mathcal{M}+1}(t+1)\right)$ of $\mathcal{Y}^{\mathcal{M}+1}$ and $\widehat{\mathcal{Y}}^{\mathcal{M}+1}$, respectively, are same. Thus, it is proved that the proposed GraDiM algorithm produces the same set of solutions if all $(\mathcal{M}+1)$ views are considered simultaneously.

7.3.5 Complexity Analysis

This section presents the time complexity of the proposed GraDiM algorithm. Let $\{\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_{(\mathcal{M}+1)}\}$ be the $(\mathcal{M}+1)$ views, with c classes and n samples, where each $\mathcal{X}_i \in \Re^{m_i \times n}$ and m_i represents the number of features in $\mathcal{X}_i$. Let us assume that the regularization parameter $\mathfrak{r}_i$ has $\mathfrak{t}_i$ possible values. Let $\{\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_{(\mathcal{M}+1)}\}$ is the order list, which is rearranged according to their largest eigenvalues of covariance matrices. Let $q = \max\{m_1, m_2, \cdots, m_{(\mathcal{M}+1)}\}$, $p = \min\{m_1, m_2, \cdots, m_{(\mathcal{M}+1)}\}$, where the number of extracted features $\mathcal{D} << p$. Let $\tau = \mathfrak{t}_1\mathfrak{t}_2\mathfrak{t}_3 + \sum\limits_{\ell=4}^{(\mathcal{M}+1)} \mathfrak{t}_\ell$.

The adjacency matrix, the degree matrix, and the Laplacian matrix of graph for each view can be computed with time complexity $\mathcal{O}(n^2)$, $\mathcal{O}(n)$, and $\mathcal{O}(n^2)$, respectively. Hence, the total time complexity to compute the Laplacian matrix of graph for all views is $\mathcal{O}((\mathcal{M}+1)n^2)$. The covariance matrices $\{\mathcal{C}_{ii}\}$ can be computed with a complexity $\mathcal{O}(\sum\limits_i m_i^2 n) \approx \mathcal{O}(q^2 n)$, $\forall i \in \{1, 2, \cdots, (\mathcal{M}+1)\}$. All the eigenvalues $\delta_{i\ell}$, along with corresponding eigenvectors $\omega_{i\ell}$, are computed with computational complexity $\mathcal{O}(\sum\limits_i m_i^3) \approx \mathcal{O}(q^3)$; $\forall \ell \in \{1, 2, \cdots, m_i\}$, in step 3. On the other hand, step 4 and step 5 have constant time complexity of $\mathcal{O}(1)$. Thus, the total computational complexity of these five steps is $\mathcal{O}((\mathcal{M}+1)n^2 + q^2 n + q^3) \approx \mathcal{O}(q^3)$ as $n << q$ and $\mathcal{M} << q$.

In step 6, there is a loop which is executed $(\mathcal{M}-2)$ times. The first step of this loop has constant time complexity, which is given as $\mathcal{O}(1)$ and the next step has another loop, which is executed $\mathcal{D}$ times. Again, the first step of this loop has constant time complexity, which is given by $\mathcal{O}(1)$ and the next step has another loop, which is executed $\prod\limits_{\ell=1}^{i} \mathfrak{t}_\ell$ times if $i = 3$; otherwise, $\mathfrak{t}_i$ times, $\forall i \in \{4, 5, \cdots, (\mathcal{M}+1)\}$. The complexity to compute $\mathcal{Y}_r(t)^i$ is $\mathcal{O}(\tau(q^3 + qn^2 + n^2)) \approx \mathcal{O}(\tau q^3)$. The eigenvector of the matrix $\mathcal{Y}_r(t)^i$ can be calculated with computational complexity $\mathcal{O}(q^2)$. The t-th basis vector $\mathcal{W}^i_{\mathcal{K}r}(t)$ of i-th view can be computed with time complexity $\mathcal{O}(pqn)$. The t-th canonical variable $\mathcal{U}^i_{\mathcal{K}r}(t)$ can be computed with time complexity $\mathcal{O}(qn)$. Hence, a feature $\mathcal{F}^i_r(t)$ can be extracted with computational complexity $\mathcal{O}(n)$. The step to compute both significance and relevance of a feature has the same time complexity, which is given by $\mathcal{O}(cn)$. The selection of a feature from $\mathfrak{t}_1\mathfrak{t}_2\mathfrak{t}_3$ candidate features, for $i = 3$, otherwise $(\tau - \mathfrak{t}_1\mathfrak{t}_2\mathfrak{t}_3)$ candidate features, by maximizing both relevance and significance, has complexity $\mathcal{O}(\tau)$. The last step of the loop has constant time complexity of $\mathcal{O}(1)$. So, the total complexity to execute the loop $\mathcal{D}$

times is $\mathcal{O}(\mathcal{D}(\tau q^3 + q^2 + pqn + qn + n + cn + \tau)) \approx \mathcal{O}(\mathcal{D}\tau q^3)$. Hence, the proposed GDiM algorithm has computational complexity of $\mathcal{O}(q^3 + \mathcal{D}\tau q^3) \approx \mathcal{O}(\mathcal{D}\tau q^3)$.

7.4 Performance Analysis

The performance of the proposed feature extraction algorithm, termed as GraDiM, is extensively studied and compared with that of several existing multimodal data integration algorithms. To evaluate the performance of different algorithms, support vector machine with linear kernels is used. Each regularization parameter is varied in between 0.0 and 1.0, with a difference of 0.1. Five benchmark data sets, namely, CiteSeer, Handwritten, NUS-WIDE-OBJECT (NW-OBJECT), Reuters, and Caltech; and five cancer data sets, namely, glioblastoma multiforme (GBM), lung (LUNG), kidney (KIDNEY), lower grade glioma (LGG), and ovarian serous cystadenocarcinoma (OV), are used in the current research work. All the data sets are summarized in Table 5.1 and Table 5.2 of Chapter 5, and briefly described in Appendix A. The proposed algorithm is implemented in C language and run in Ubuntu 14.04 LTS having machine configuration Intel(R) Core(TM) i7-4790 CPU @ 3.60GHz×8 and 32 GB RAM.

The randomly selected 50% samples from each class are used for training and the rest are used for testing purposes for each of the data sets. The 10-fold cross-validation (CV) is also performed on each data set to assess the performance of the proposed algorithm statistically. To analyze the statistical significance of the derived results, paired-t test (one-tailed), Wilcoxon signed rank test (one-tailed), and Friedman test (one-tailed), with 95% confidence level, are used to compute the p-values. For each data set, 25 top-ranked correlated features are selected for the analysis.

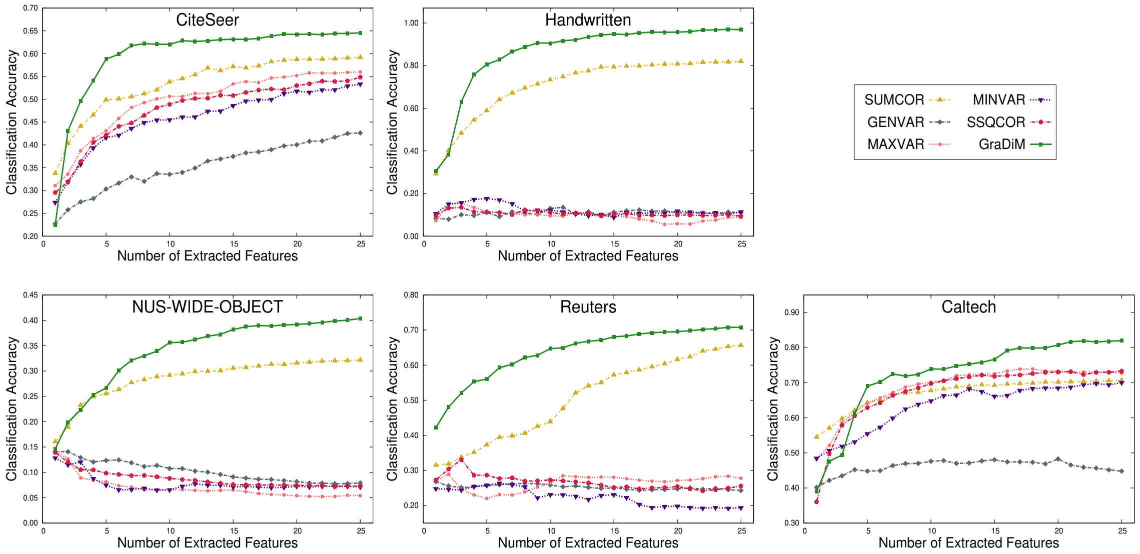


Figure 7.1: Variation of classification accuracy with respect to number of extracted features for different criteria of the MCCA and proposed (GraDiM) algorithm on benchmark data sets using 10-fold CV.

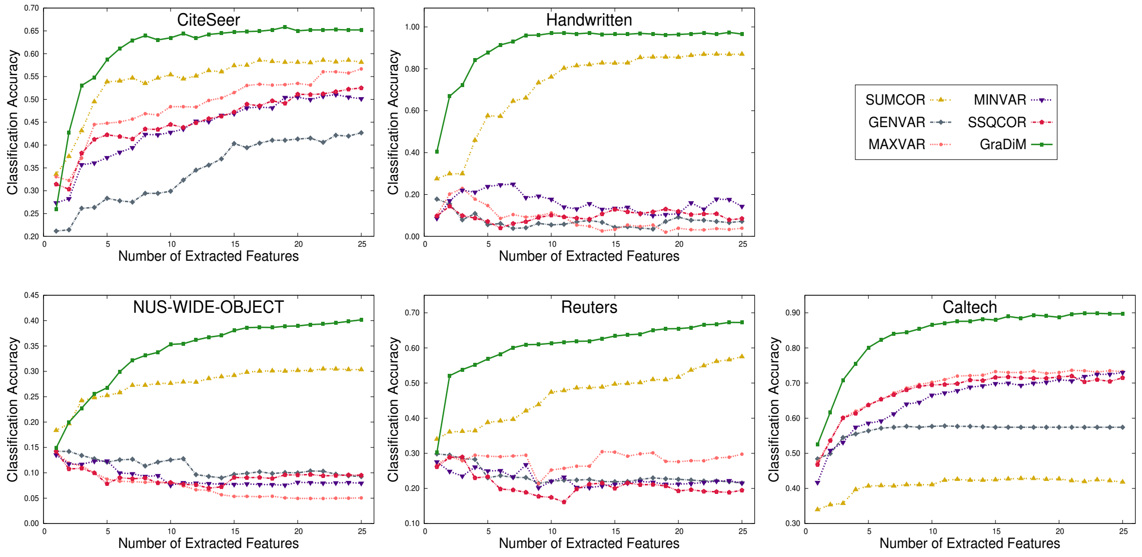


Figure 7.2: Variation of classification accuracy with respect to number of extracted features for different criteria of the MCCA and proposed (GraDiM) algorithm on benchmark data sets using training-testing.

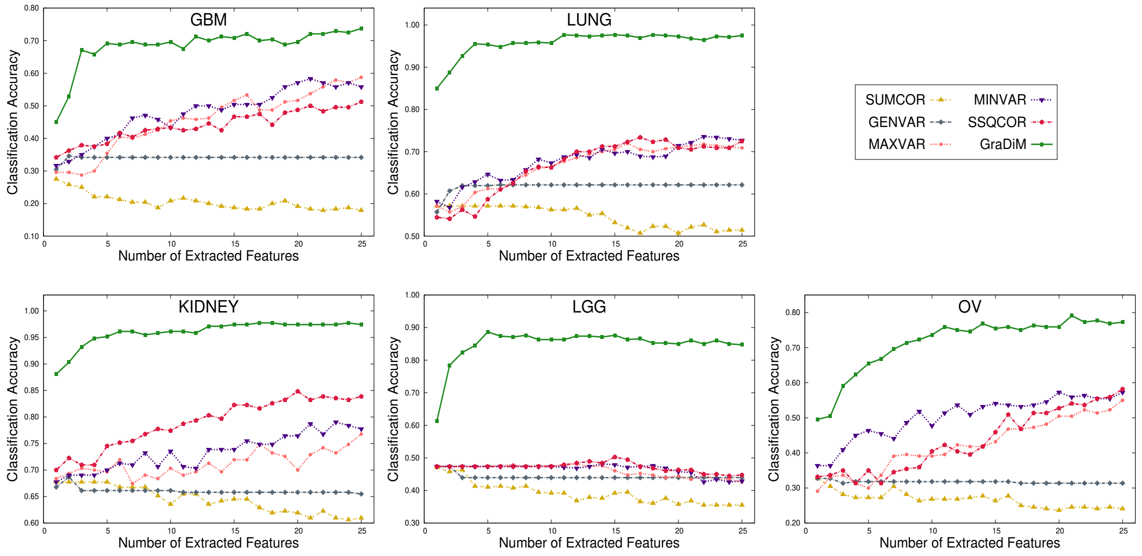


Figure 7.3: Variation of classification accuracy with respect to number of extracted features for different criteria of the MCCA and proposed (GraDiM) algorithm on omics data sets using 10-fold CV.

7.4.1 Importance of Various Criteria of MCCA

Figure 7.1, Figure 7.2, Figure 7.3, and Figure 7.4 compare the performance of the proposed algorithm with that of various criteria of the MCCA, namely, SUMCOR, MAXVAR, generalized variance (GENVAR), minimum variance (MINVAR), and sum of squared cor-

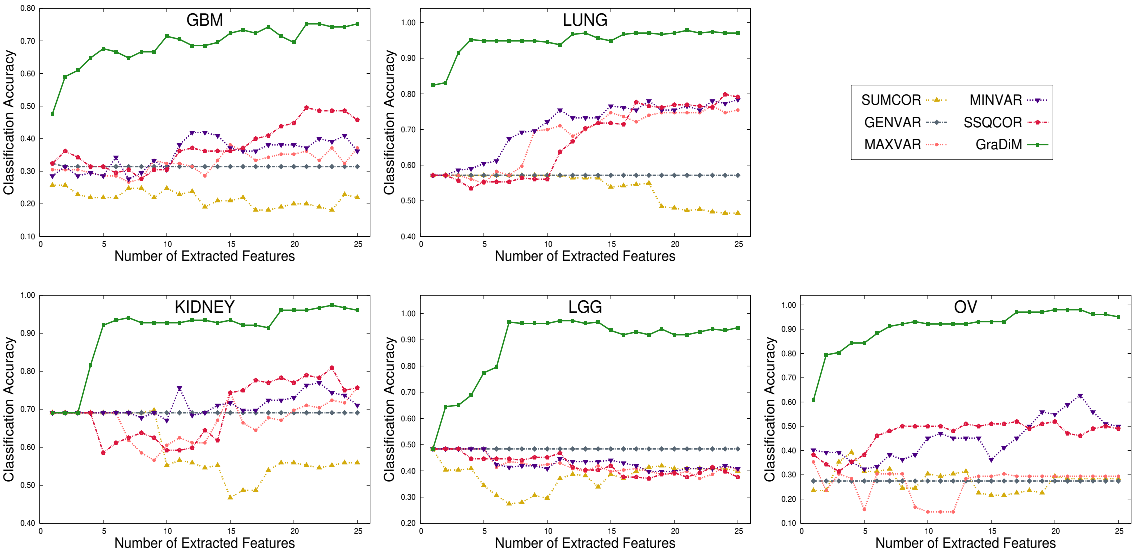


Figure 7.4: Variation of classification accuracy with respect to number of extracted features for different criteria of the MCCA and proposed (GraDiM) algorithm on omics data sets using training-testing.

relations (SSQCOR) [135]. Table 7.1 presents the classification accuracy obtained using the proposed algorithm on each data set in case of training-testing. The mean, median, and standard deviation of 10-fold CV are also reported in Table 7.1 for each data set. To perform the statistical significance analysis, the p-values computed using different tests are reported in Table 7.2, Table 7.3, and Table 7.4.

7.4.1.1 Performance on Benchmark Data

The results presented in Figure 7.1 and Figure 7.2 convey that the MAXVAR provides the highest accuracy irrespective of the features among different criteria of the MCCA on Caltech data sets. However, the performance of the proposed algorithm is significantly higher as compared to that of various criteria of the MCCA, irrespective of the generated features and data sets used. The results reported in Table 5.3, Table 5.4, and Table 5.5 of Chapter 5, and Table 7.1 demonstrate that the MAXVAR criterion provides the highest classification accuracy of 0.733 on Caltech data set. On the other hand, the SUMCOR attains the highest accuracy of 0.581, 0.870, 0.303, and 0.575 on CiteSeer, Handwritten, NW-OBJECT, and Reuters data sets, respectively, among five criteria of the MCCA. However, the proposed algorithm achieves the highest classification accuracy on all five benchmark data sets. All the results reported in Table 7.2, Table 7.3, and Table 7.4 depict that the proposed GraDiM algorithm achieves significantly better (marked in bold) p-values than different criteria of the MCCA in all 75 cases. Figure 7.5 shows the scatter plots using the first two extracted features, along with the class separability index, of the proposed algorithm on five benchmark and five omics data sets, while the corresponding results of different criteria of the MCCA are reported in the first five columns of Figure 5.5 of Chapter 5. All the results reported in Figure 5.5 of Chapter 5 and Figure 7.5 establish

Table 7.1: Classification Accuracy and Execution Time of Proposed GraDiM Algorithm

Different		Accuracy	Accuracy for 10-Fold CV			Time
Data Sets		(Train-Test)	Mean	Median	StdDev	(in sec.)
Benchmark	CiteSeer	0.652	0.646	0.647	0.027	90.5
	Handwritten	0.966	0.969	0.970	0.016	103.8
	NW-OBJECT	0.402	0.404	0.405	0.006	11659.2
	Reuters	0.672	0.708	0.712	0.013	26924.3
	Caltech	0.897	0.820	0.821	0.015	1552.7
Omics	GBM	0.752	0.738	0.696	0.079	223.1
	LUNG	0.971	0.975	0.970	0.029	447.2
	KIDNEY	0.961	0.974	0.971	0.025	281.8
	LGG	0.946	0.847	0.863	0.152	332.4
	OV	0.951	0.773	0.750	0.096	220.6

Table 7.2: Statistical Significance Analysis of Different Algorithms on CiteSeer and GBM Data Sets

Different		Data	p-values for 10-Fold CV			Data	p-values for 10-Fold CV		
Algorithms		Sets	Paired-t	Wilcoxon	Friedman	Sets	Paired-t	Wilcoxon	Friedman
MCCA	SUMCOR	CiteSeer	**2.15E-05**	**2.52E-03**	**1.57E-03**	GBM	**1.39E-06**	**2.40E-03**	**1.57E-03**
	GENVAR		**2.60E-05**	**2.53E-03**	**1.57E-03**		**3.18E-04**	**3.33E-03**	**2.70E-03**
	MAXVAR		**6.21E-06**	**2.52E-03**	**1.57E-03**		**5.40E-06**	**2.45E-03**	**1.57E-03**
	MINVAR		**2.01E-06**	**2.50E-03**	**1.57E-03**		**1.24E-02**	**1.94E-02**	*2.06E-01*
	SSQCOR		**5.37E-05**	**3.46E-03**	**1.14E-02**		**4.18E-03**	**7.23E-03**	**1.96E-02**
RGCCA			**3.47E-10**	**2.53E-03**	**1.57E-03**		**2.28E-03**	**7.03E-03**	**1.14E-02**
GMCCA			**2.66E-09**	**2.53E-03**	**1.57E-03**		**2.00E-04**	**3.74E-03**	**2.70E-03**
GMKCCA			**9.52E-11**	**2.53E-03**	**1.57E-03**		**1.87E-06**	**2.46E-03**	**1.57E-03**
LasCCA			**1.25E-10**	**2.52E-03**	**1.57E-03**		**4.66E-02**	**4.42E-02**	*1.57E-01*
DisCCA			**1.22E-11**	**2.53E-03**	**1.57E-03**		**1.86E-05**	**2.46E-03**	**1.57E-03**
BsMCCA			**1.53E-10**	**2.53E-03**	**1.57E-03**		**4.43E-02**	**3.81E-02**	*5.78E-02*
ReDMiCA			**7.48E-03**	**9.78E-03**	*9.56E-02*		6.92E-01	7.06E-01	6.55E-01
MvDA			**9.19E-09**	**2.53E-03**	**1.57E-03**		7.78E-01	7.54E-01	*2.57E-01*
MvDA-VC			**1.78E-07**	**2.53E-03**	**1.57E-03**		**2.64E-02**	**2.68E-02**	*5.88E-02*
LiveGCANO			**3.70E-09**	**2.45E-03**	**1.57E-03**		**8.06E-03**	**1.20E-02**	**3.39E-02**
OPID			**1.52E-09**	**2.52E-03**	**1.57E-03**		*3.36E-01*	*3.66E-01*	7.06E-01
SAC			**2.50E-10**	**2.53E-03**	**1.57E-03**		**3.93E-02**	*6.01E-02*	*9.56E-02*
SeFGeIM			*1.40E-01*	*1.21E-01*	*4.80E-01*		*4.60E-01*	7.03E-01	*4.14E-01*

the superiority of the proposed algorithm over different criteria of the MCCA.

7.4.1.2 Performance on Omics Data

All the results reported in Figure 7.3 and Figure 7.4 demonstrate that the classification accuracy of the proposed algorithm is significantly higher as compared to that of various criteria of the MCCA, irrespective of the generated features, data sets, and experimental setup used. All the results reported in Table 5.3, Table 5.4, and Table 5.5 of Chapter 5 and Table 7.1 confirm that the proposed algorithm attains the highest mean and median accuracy, irrespective of all five omics data sets. From the results reported in Table 7.2,

Table 7.3: Statistical Significance Analysis of Different Algorithms on Handwritten, NW-OBJECT, LUNG, and KIDNEY Data Sets

Different		Data	p-values for 10-Fold CV			Data	p-values for 10-Fold CV		
Algorithms		Sets	Paired-t	Wilcoxon	Friedman	Sets	Paired-t	Wilcoxon	Friedman
MCCA	SUMCOR	Handwritten	**1.93E-08**	**2.50E-03**	**1.57E-03**	LUNG	**3.98E-11**	**2.50E-03**	**1.57E-03**
	GENVAR		**3.91E-11**	**2.52E-03**	**1.57E-03**		**5.59E-05**	**2.49E-03**	**1.57E-03**
	MAXVAR		**2.22E-13**	**2.52E-03**	**1.57E-03**		**1.60E-06**	**2.50E-03**	**1.57E-03**
	MINVAR		**1.11E-12**	**2.50E-03**	**1.57E-03**		**8.81E-06**	**2.53E-03**	**1.57E-03**
	SSQCOR		**1.26E-14**	**2.52E-03**	**1.57E-03**		**9.41E-06**	**2.52E-03**	**1.57E-03**
RGCCA			**1.56E-05**	**3.82E-03**	**2.70E-03**		**4.45E-06**	**2.50E-03**	**1.57E-03**
GMCCA			**8.74E-15**	**2.47E-03**	**1.57E-03**		**5.16E-08**	**2.49E-03**	**1.57E-03**
GMKCCA			**9.23E-13**	**2.53E-03**	**1.57E-03**		**1.64E-03**	**3.98E-03**	**1.14E-02**
LasCCA			**1.12E-15**	**2.46E-03**	**1.57E-03**		**5.89E-05**	**2.50E-03**	**1.57E-03**
DisCCA			**3.03E-12**	**2.50E-03**	**1.57E-03**		**4.69E-09**	**2.49E-03**	**1.57E-03**
BsMCCA			**7.94E-12**	**2.52E-03**	**1.57E-03**		**1.32E-02**	**1.03E-02**	**3.39E-02**
ReDMiCA			1.00E+00	1.00E+00	1.00E+00		**2.61E-02**	**2.90E-02**	*9.56E-02*
MvDA			**1.34E-03**	**2.42E-03**	**1.57E-03**		*5.98E-02*	*5.58E-02*	*2.06E-01*
MvDA-VC			**2.47E-02**	**2.30E-02**	*5.78E-02*		*8.88E-02*	*1.17E-01*	*2.06E-01*
LiveGCANO			**5.24E-13**	**2.52E-03**	**1.57E-03**		**2.96E-10**	**2.42E-03**	**1.57E-03**
OPID			**8.08E-03**	**8.02E-03**	**1.14E-02**		**1.06E-02**	*1.39E-02*	**3.39E-02**
SAC			**2.20E-03**	**5.81E-03**	**4.68E-03**		*6.08E-02*	*7.00E-02*	*4.80E-01*
SeFGeIM			5.00E-01	*4.66E-01*	7.06E-01		*1.29E-01*	*1.49E-01*	*3.17E-01*
MCCA	SUMCOR	NW-OBJECT	**1.74E-12**	**2.53E-03**	**1.57E-03**	KIDNEY	**9.65E-09**	**2.47E-03**	**1.57E-03**
	GENVAR		**3.98E-14**	**2.53E-03**	**1.57E-03**		**1.47E-02**	**1.97E-02**	**3.39E-02**
	MAXVAR		**2.20E-16**	**2.53E-03**	**1.57E-03**		**1.38E-06**	**2.46E-03**	**1.57E-03**
	MINVAR		**5.04E-15**	**2.52E-03**	**1.57E-03**		**2.12E-06**	**2.50E-03**	**1.57E-03**
	SSQCOR		**2.20E-16**	**2.52E-03**	**1.57E-03**		**3.61E-07**	**2.40E-03**	**1.57E-03**
RGCCA			**3.17E-16**	**2.52E-03**	**1.57E-03**		**1.48E-03**	**3.90E-03**	**1.14E-02**
GMCCA			**3.27E-16**	**2.53E-03**	**1.57E-03**		**5.10E-06**	**2.52E-03**	**1.57E-03**
GMKCCA			**5.18E-13**	**2.53E-03**	**1.57E-03**		**7.27E-05**	**2.39E-03**	**1.57E-03**
LasCCA			**2.20E-16**	**2.53E-03**	**1.57E-03**		**2.15E-04**	**2.50E-03**	**1.57E-03**
DisCCA			**5.51E-14**	**2.53E-03**	**1.57E-03**		**1.72E-07**	**2.39E-03**	**1.57E-03**
BsMCCA			**1.20E-15**	**2.53E-03**	**1.57E-03**		**1.98E-03**	**3.32E-03**	**1.14E-02**
ReDMiCA			**2.46E-06**	**2.53E-03**	**1.57E-03**		*3.63E-01*	*3.53E-01*	*3.17E-01*
MvDA			**1.95E-12**	**2.53E-03**	**1.57E-03**		**3.73E-04**	**4.69E-03**	**4.68E-03**
MvDA-VC			**5.27E-12**	**2.53E-03**	**1.57E-03**		**7.48E-03**	**1.05E-02**	**1.96E-02**
LiveGCANO			**1.75E-11**	**2.53E-03**	**1.57E-03**		**2.83E-06**	**2.47E-03**	**1.57E-03**
OPID			**8.67E-14**	**2.53E-03**	**1.57E-03**		*1.72E-01*	*1.59E-01*	*4.14E-01*
SAC			**4.65E-12**	**2.53E-03**	**1.57E-03**		**5.35E-03**	**1.16E-02**	**1.43E-02**
SeFGeIM			**3.26E-03**	**8.30E-03**	*5.78E-02*		*3.39E-01*	*3.27E-01*	6.55E-01

Table 7.3, and Table 7.4, it is evident that out of a total of 75 cases, the proposed algorithm achieves significantly better (marked in bold) p-values than different criteria of the MCCA in 73 cases. On the other hand, the proposed algorithm provides better but not significant (marked in italics) p-values in only 2 cases, for the SSQCOR and the MINVAR using the Friedman test on OV and GBM data sets, respectively. Comparing the results reported in the first five columns of Figure 5.6 of Chapter 5 and the bottom row of Figure 7.5, it is clear that the scatter plots of the first two extracted features of the proposed algorithm is able to separate different classes on omics data sets more precisely than the different

Table 7.4: Statistical Significance Analysis of Different Algorithms on Reuters, Caltech, LGG, and OV Data Sets

Different Algorithms		Data Sets	p-values for 10-Fold CV			Data Sets	p-values for 10-Fold CV		
			Paired-t	Wilcoxon	Friedman		Paired-t	Wilcoxon	Friedman
MCCA	SUMCOR	Reuters	**5.15E-08**	**2.53E-03**	**1.57E-03**	LGG	**1.83E-05**	**3.46E-03**	**1.14E-02**
	GENVAR		**4.73E-13**	**2.53E-03**	**1.57E-03**		**1.50E-04**	**3.46E-03**	**1.14E-02**
	MAXVAR		**2.99E-13**	**2.53E-03**	**1.57E-03**		**5.38E-06**	**2.49E-03**	**1.57E-03**
	MINVAR		**9.16E-16**	**2.53E-03**	**1.57E-03**		**2.36E-05**	**3.44E-03**	**1.14E-02**
	SSQCOR		**9.42E-12**	**2.53E-03**	**1.57E-03**		**7.52E-06**	**2.50E-03**	**1.57E-03**
RGCCA			**1.47E-13**	**2.53E-03**	**1.57E-03**		**3.62E-05**	**3.42E-03**	**1.14E-02**
GMCCA			**5.19E-13**	**2.53E-03**	**1.57E-03**		**8.87E-06**	**2.50E-03**	**1.57E-03**
GMKCCA			**2.64E-13**	**2.53E-03**	**1.57E-03**		**9.13E-07**	**2.53E-03**	**1.57E-03**
LasCCA			**8.20E-12**	**2.53E-03**	**1.57E-03**		**2.16E-06**	**2.50E-03**	**1.57E-03**
DisCCA			**3.48E-08**	**2.53E-03**	**1.57E-03**		**6.17E-06**	**2.50E-03**	**1.57E-03**
BsMCCA			**1.42E-11**	**2.53E-03**	**1.57E-03**		**6.03E-03**	**1.42E-02**	**1.14E-02**
ReDMiCA			**5.76E-04**	**3.44E-03**	**1.14E-02**		5.21E-01	*2.20E-01*	*3.17E-01*
MvDA			**4.87E-11**	**2.53E-03**	**1.57E-03**		**2.99E-02**	**2.88E-02**	**1.14E-02**
MvDA-VC			**2.11E-11**	**2.53E-03**	**1.57E-03**		*2.80E-01*	*7.81E-02*	**3.39E-02**
LiveGCANO			**8.77E-11**	**2.50E-03**	**1.57E-03**		**5.16E-06**	**2.52E-03**	**1.57E-03**
OPID			**2.67E-09**	**2.53E-03**	**1.57E-03**		5.00E-01	*3.60E-01*	7.39E-01
SAC			**1.43E-08**	**2.53E-03**	**1.57E-03**		*4.59E-01*	*1.06E-01*	*9.56E-02*
SeFGeIM			*2.97E-01*	*3.80E-01*	5.27E-01		*4.78E-01*	*2.42E-01*	*1.57E-01*
MCCA	SUMCOR	Caltech	**4.92E-07**	**2.47E-03**	**1.57E-03**	OV	**4.54E-08**	**2.50E-03**	**1.57E-03**
	GENVAR		**1.29E-08**	**2.52E-03**	**1.57E-03**		**9.13E-09**	**2.46E-03**	**1.57E-03**
	MAXVAR		**3.74E-08**	**2.50E-03**	**1.57E-03**		**5.55E-06**	**2.52E-03**	**1.57E-03**
	MINVAR		**5.74E-09**	**2.52E-03**	**1.57E-03**		**2.16E-03**	**8.37E-03**	**3.39E-02**
	SSQCOR		**2.55E-09**	**2.47E-03**	**1.57E-03**		**8.32E-03**	**1.42E-02**	*5.78E-02*
RGCCA			**2.12E-15**	**2.52E-03**	**1.57E-03**		**2.69E-03**	**5.40E-03**	**1.96E-02**
GMCCA			**3.87E-16**	**2.53E-03**	**1.57E-03**		**2.59E-06**	**2.52E-03**	**1.57E-03**
GMKCCA			**1.35E-15**	**2.52E-03**	**1.57E-03**		**1.23E-06**	**2.52E-03**	**1.57E-03**
LasCCA			**2.33E-15**	**2.52E-03**	**1.57E-03**		**5.70E-08**	**2.50E-03**	**1.57E-03**
DisCCA			**1.23E-10**	**2.53E-03**	**1.57E-03**		**2.84E-08**	**2.52E-03**	**1.57E-03**
BsMCCA			**7.66E-04**	**3.44E-03**	**1.14E-02**		**2.42E-02**	**3.44E-02**	**3.39E-02**
ReDMiCA			**5.70E-03**	**1.08E-02**	*5.78E-02*		**4.71E-02**	**4.25E-02**	**1.96E-02**
MvDA			**6.04E-04**	**3.46E-03**	**1.14E-02**		**2.87E-02**	**3.27E-02**	*3.17E-01*
MvDA-VC			**1.49E-03**	**3.74E-03**	**2.70E-03**		**8.66E-05**	**3.74E-03**	**2.70E-03**
LiveGCANO			**2.20E-16**	**2.42E-03**	**1.57E-03**		**5.83E-07**	**2.52E-03**	**1.57E-03**
OPID			**1.04E-05**	**2.47E-03**	**1.57E-03**		**2.75E-03**	**9.41E-03**	**1.14E-02**
SAC			**2.24E-04**	**2.50E-03**	**1.57E-03**		**3.93E-03**	**3.71E-03**	**2.70E-03**
SeFGeIM			**7.14E-04**	**3.79E-03**	**2.70E-03**		*3.13E-01*	*3.35E-01*	7.06E-01

criteria of the MCCA do.

7.4.2 Comparative Performance Analysis

Finally, Figure 7.6, Figure 7.7, Figure 7.8, and Figure 7.9 along with Table 7.1, Table 7.2, Table 7.3, Table 7.4, and Table 7.5 analyze the performance of the proposed multimodal data integration algorithm, termed as GraDiM, with that of various state-of-the-art MCCA based methods, namely, RGCCA [262], GMCCA [43], GMKCCA [43], large-scale generalized CCA (LasCCA) [84], distributed generalized CCA (DisCCA) [84], block sparse MCCA

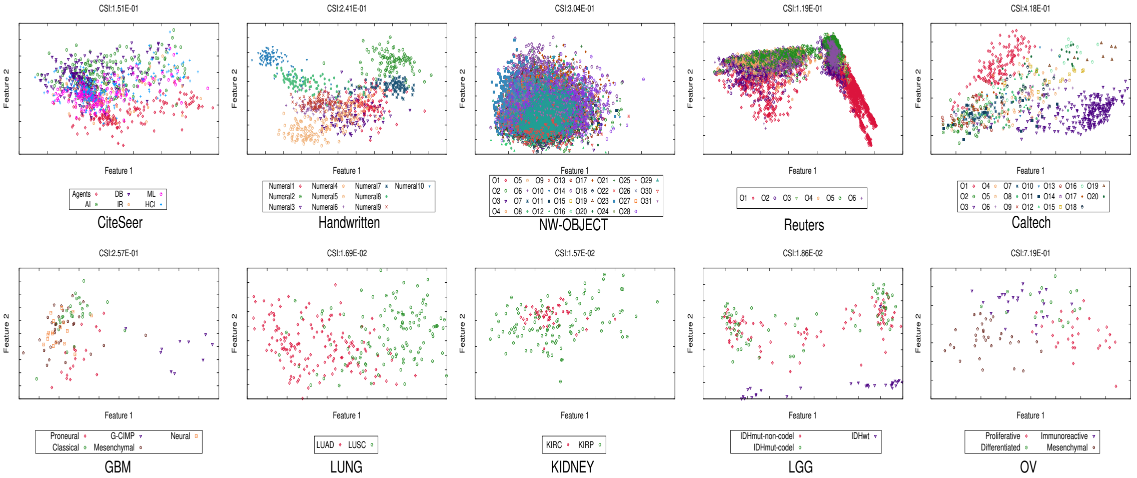


Figure 7.5: Scatter plots for the proposed GraDiM algorithm, along with class separability index, each Oi denotes the i-th object class.

(BsMCCA) [235], and ReDMiCA [185] presented in Chapter 5; two popular multidimensional data integration algorithms, namely, multi-view discriminant analysis (MvDA) [128] and MvDA with view-consistency (MvDA-VC) [129]; three multi-view incremental algorithms, namely, live generalized canonical correlation analysis (LiveGCANO) [187], one-pass learning with incremental and decremental features (OPID) [114], safe classification with augmented features (SAC) [113], and SeFGeIM [184] presented in Chapter 6; and three deep learning-based algorithms, namely, deep MCCA (dMCCA) [244], deep multi-view learning via task-optimal CCA (TOCCA) [55], and multimodal deep Boltzmann machines (MDBM) [247]. On the other hand, Figure 7.5 shows the scatter plots using the first two extracted features of the proposed algorithm on each data set, while the corresponding plots of the aforementioned algorithms are reported in Figure 5.11 and Figure 5.12 of Chapter 5 and Figure 6.6, and Figure 6.7 of Chapter 6.

7.4.2.1 MCCA Based Methods

Figure 7.6, Figure 7.7, Figure 7.8, and Figure 7.9 along with Table 5.3, Table 5.4, and Table 5.5 of Chapter 5 and Table 7.1 demonstrate that the accuracy of the proposed multi-view data integration algorithm is significantly higher as compared to that of existing MCCA based methods on both omics and benchmark data sets. All the results reported in Table 5.3, Table 5.4, and Table 5.5, of Chapter 5 and Table 7.1 confirm that the proposed algorithm attains the highest mean and median accuracy, in most of the data sets. From the results reported in Table 7.2, Table 7.3, and Table 7.4, it is seen that out of total 180 cases, the proposed algorithm attains significantly better (marked in bold) p-values than existing MCCA based methods in 178 cases, and better but not significant (marked in italics) p-values in 2 cases. From the first four columns of Figure 5.11 and the first two columns of Figure 5.12 of Chapter 5, and Figure 7.5, it can be seen that the separation among various classes using the first two extracted features of the proposed algorithm is significantly better than that of the existing algorithms on Handwritten and Caltech data sets. It shows that the proposed algorithm can separate different classes of LUNG data set using the first two

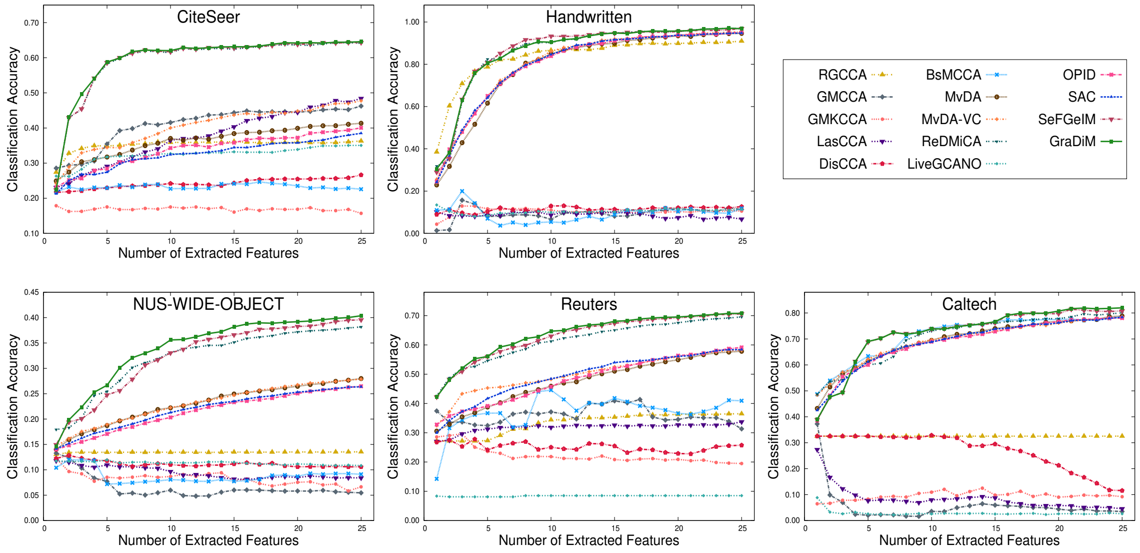


Figure 7.6: Variation of classification accuracy with respect to number of extracted features for different existing algorithms and proposed (GraDiM) algorithm on benchmark data sets using 10-fold CV.

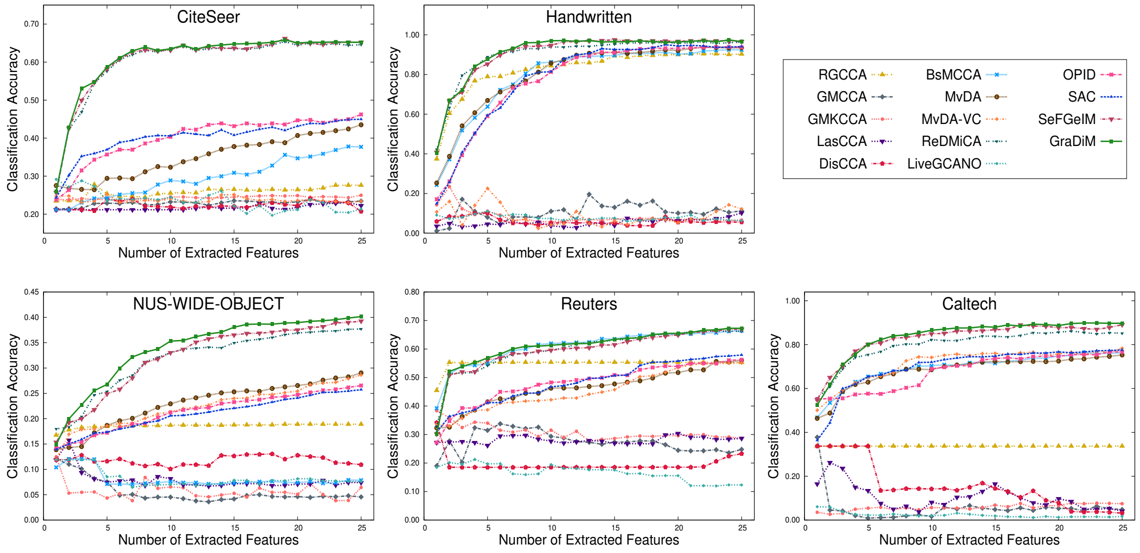


Figure 7.7: Variation of classification accuracy with respect to number of extracted features for different existing algorithms and proposed (GraDiM) algorithm on benchmark data sets using training-testing.

extracted features only. For the LGG data set, the proposed algorithm isolates almost all the samples of IDHwt class properly, but there is an overlap between the samples of the other two classes. On the other hand, for the OV data set, most of the samples of Proliferative and Mesenchymal classes are well segregated, though Immunoreactive and

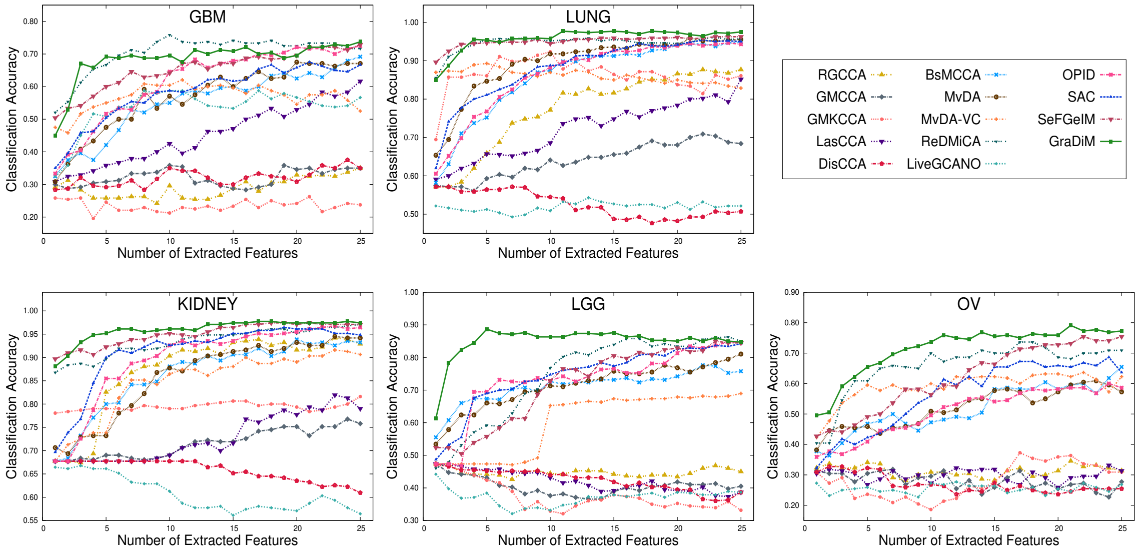


Figure 7.8: Variation of classification accuracy with respect to number of extracted features for different existing algorithms and proposed (GraDiM) algorithm on omics data sets using 10-fold CV.

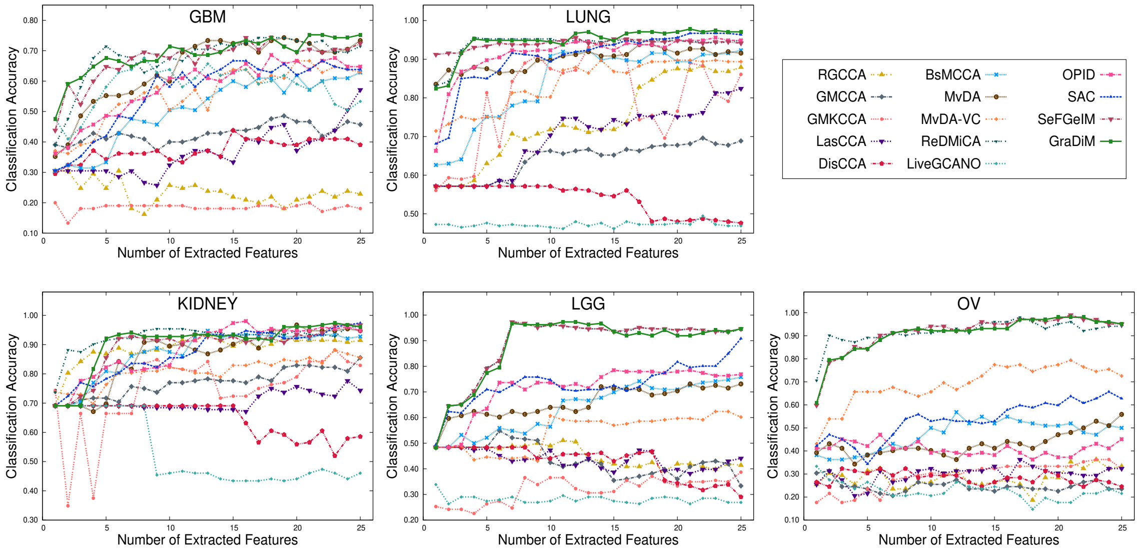


Figure 7.9: Variation of classification accuracy with respect to number of extracted features for different existing algorithms and proposed (GraDiM) algorithm on omics data sets using training-testing.

Differentiated classes do not have a linear boundary between them. For the GBM data set, most of the samples of Proneural and G-CIMP classes are disconnected, but a few samples of Classical, Mesenchymal, and Neural classes are not classified properly.

7.4.2.2 Multi-View Learning Algorithms

From Figure 7.6, Figure 7.7, Figure 7.8, and Figure 7.9, it is seen that the classification accuracy of the proposed algorithm is significantly higher, irrespective of the number of extracted features, as compared to that of both MvDA and MvDA-VC on each of the data sets. As shown in Table 7.2, Table 7.3, and Table 7.4, out of total 60 cases, the proposed algorithm attains significantly better (marked in bold) p-values than other two data integration methods in 46 cases, and better but not significant (marked in italics) p-values in 12 cases. The proposed algorithm is not significantly better than MvDA according to paired-t test and Wilcoxon signed rank test on the GBM data set. Comparing the results reported in the 3rd and 4th columns of Figure 5.12 of Chapter 5 and Figure 7.5, it is evident that different classes are remarkably separable using the first two extracted features of the proposed algorithm than these two existing multi-view learning algorithms on omics as well as benchmark data sets. Moreover, the CSI of the proposed algorithm is higher compared to that of MvDA and MvDA-VC, which indicates the better separation of the classes.

7.4.2.3 Multi-View Incremental Learning Algorithms

The results reported in Figure 7.6, Figure 7.7, Figure 7.8, and Figure 7.9 clearly establish the fact that the proposed GraDiM algorithm provides better performance than three incremental multi-view data integration methods, namely, LiveGCANO, OPID, and SAC, in most of the cases. However, for the GBM data set, SAC attains a little overlap with GraDiM. From the results presented in Table 6.2 of Chapter 6 and Table 7.1, it can be seen that the proposed algorithm attains the higher classification accuracy of training-testing in 9 cases, while the SAC algorithm achieves it only for the KIDNEY data set. The results corresponding to 10-fold CV indicate that the proposed algorithm attains the higher mean accuracy in 10 cases and higher median accuracy in 8 cases, out of total 10 cases each. Moreover, the proposed GraDiM algorithm attains significantly better p-values (marked in bold) than the three incremental methods in 73 cases, out of the total 90 cases, and better but not significant p-values (marked in italics) in 14 cases, considering 95% confidence level. The proposed algorithm is not significantly better than the OPID according to paired-t test and Friedman test on the LGG data set and Friedman test on the GBM data set. Finally, the comparative analysis of the scatter plots presented in the first three columns of Figure 6.6 and Figure 6.7 of Chapter 6 and Figure 7.5 confirm that the proposed algorithm can separate different classes better than the existing approaches.

7.4.2.4 Deep Learning Based Methods

Finally, the performance of the proposed GraDiM algorithm is compared with that of several deep learning-based algorithms, namely, dMCCA [244], TOCCA [55], and MDBM [247]. The results presented in Table 5.9 and Table 5.10 of Chapter 5 and Table 7.1 demonstrate that the classification accuracy of the proposed algorithm is significantly higher as compared to that of various deep learning-based methods in all the cases. The TOCCA algorithm performs well on benchmark data sets other than CiteSeer, but it fails to achieve judicious results on most of the omics data sets. The MDBM and dMCCA obtain 87.9% and 86.2% accuracy on LUNG and KIDNEY data sets, respectively, whereas both of them perform moderately on the GBM and LGG data sets. On the other hand, none of the deep learning-based methods performs better on the OV data set. Both MDBM and dMCCA provide poor performance on Handwritten, Caltech, and NW-OBJECT data sets due to the over training of these models. Table 7.5 presents the statistical significance analysis on five omics data sets. The results reported in Table 7.5 establish that the proposed algorithm attains significantly better p-values than the three deep learning-based methods, irrespective of the significance analysis and omics data sets used.

Table 7.5: Statistical Significance Analysis of Different Deep Learning Algorithms on Omics Data Sets

Different Data Sets	Different Algorithms	p-values for 10-Fold CV		
		Paired-t	Wilcoxon	Friedman
GBM	dMCCA	**1.76E-05**	**2.53E-03**	**1.57E-03**
	TOCCA	**1.95E-06**	**2.52E-03**	**1.57E-03**
	MDBM	**6.94E-05**	**2.53E-03**	**1.57E-03**
LUNG	dMCCA	**1.24E-10**	**2.50E-03**	**1.57E-03**
	TOCCA	**4.57E-12**	**2.36E-03**	**1.57E-03**
	MDBM	**5.25E-04**	**2.46E-03**	**1.57E-03**
KIDNEY	dMCCA	**1.52E-09**	**2.32E-03**	**1.57E-03**
	TOCCA	**1.50E-07**	**2.25E-03**	**1.57E-03**
	MDBM	**5.53E-06**	**3.45E-03**	**2.70E-03**
LGG	dMCCA	**7.89E-05**	**3.42E-03**	**1.14E-02**
	TOCCA	**4.06E-05**	**3.42E-03**	**1.14E-02**
	MDBM	**3.49E-06**	**2.52E-03**	**1.57E-03**
OV	dMCCA	**1.25E-06**	**2.52E-03**	**1.57E-03**
	TOCCA	**4.47E-05**	**3.46E-03**	**1.14E-02**
	MDBM	**1.96E-05**	**2.53E-03**	**1.57E-03**

All the results, reported here establish the effectiveness of the proposed incremental multi-view data integration algorithm over state-of-the-art data integration approaches. In real-life data analysis, all the modalities may not be required to extract different features. Considering this fact, the GraDiM is developed in such a way that only relevant modalities are integrated to extract features. This property of the proposed method helps to perform significantly better than existing methods.

7.5 Conclusion

A novel supervised sequential feature extraction algorithm, termed as GraDiM, has been proposed in this chapter. It integrates multi-view data sets by using the MAXVAR criterion and the knowledge of the graph. It can update the solutions adaptively wherever a new view is available for the analysis, without repeating the same procedure with the original data augmented by the new view. Moreover, the algorithm is designed in such a way that if all the views are available at the beginning of the analysis, the algorithm starts with the three most relevant views, and the remaining views are added sequentially according to their relevance. The proposed GraDiM model deals with the "curse of dimensionality" problem due to 'large p and small n' characteristics of real-life multi-view data sets, by using the ridge regression optimization technique. The quality of the extracted features depends on the supervised information of sample categories. Analytical formulation facilitates the generation of relevant and significant features from multi-view dynamic data sets with significantly lower computational costs. The effectiveness of the proposed GraDiM algorithm, along with a comparison with other algorithms, has been demonstrated on several benchmarks and real-life cancer data sets.

Chapter 8

Conclusion and Future Directions

The major contributions of the research presented in different chapters of this thesis are summarized in this chapter. The possible extensions and applications of the proposed research work are also discussed in this chapter.

8.1 Major Contributions

The thesis presents some novel approaches for multi-view data integration. The main challenge associated with multi-view data analysis is five-fold, namely, (i) curse of dimensionality problem of each view, (ii) unavailability of all the modalities at the beginning of the analysis, (iii) integration of the most informative and relevant views while discarding the redundant and insignificant views, (iv) sequential extraction of relevant features based on the supervised information of sample categories, and (v) modeling the structural information associated with geometrical knowledge of individual views. All these aforementioned issues have been addressed in this thesis. A brief summary, highlighting the main features of the proposed approaches, is discussed next.

Chapter 3 presents a novel supervised regularized canonical correlation analysis (CCA), termed as CuRSaR, to extract relevant and significant features from two multidimensional data sets. An analytical formulation, based on spectral decomposition, has been introduced to establish the relationship between the covariance matrices of different regularization parameters. It makes the computational complexity of the proposed algorithm significantly lower than that of the existing methods. The algorithm proposed in Chapter 3 extracts a set of features simultaneously from two multidimensional data sets by maximizing the relevance of extracted features with respect to sample categories and significance among them. However, instead of producing all canonical variables simultaneously, if each variable is computed sequentially, the quality of each generated feature can be evaluated independently, and eventually, a reduced set of features can be selected based on their quality. In this regard, a fast and robust feature extraction algorithm, termed as FaRoC, has been presented in Chapter 4, which extracts new features sequentially from two multidimensional data sets by maximizing their relevance with respect to the class label and significance with respect to the already-extracted features. To generate canonical variables sequentially, an analytical formulation has been introduced to establish the relation between regularization parameters and CCA. The formulation enables FaRoC to extract a required number of correlated features sequentially with lesser computational cost as compared to existing algorithms. The efficacy of both CuRSaR and FaRoC, along with a comparison with other existing methods, has been extensively established on several real-life cancer data sets.

Both CuRSaR and FaRoC can only account for two sets of variables. In this regard, a new multi-view data integration algorithm, termed as ReDMiCA, has been presented in Chapter 5. It integrates multimodal multidimensional data sets by solving the maximal correlation problem of multiset CCA (MCCA). A new block matrix representation has been introduced to determine the basis vectors of the MCCA. The proposed algorithm has addressed the high-dimension low-sample size issue of real-world multi-view data sets by using the ridge regression optimization technique. A theoretical analysis has been presented to generate the desired number of correlated features sequentially, without producing the complete set of possible features. An important result of this analysis is that the proposed algorithm computes the canonical variable for a single modality having the lowest dimension with the initial regularization parameter, and this canonical variable can be used to compute the canonical variables of all other modalities at different combinations of regularization parameters. The optimum values of regularization parameters have been estimated by computing the relevance and significance of the corresponding feature. The performance of the proposed multiblock data integration algorithm has been compared with that of different existing integrative methods on several real-life cancer as well as benchmark data sets from varying application domains.

The algorithms proposed in Chapter 3 to Chapter 5 are applicable for multiblock static data analysis. In multiblock dynamic data, all the modalities may not be available initially. The databases are generally updated incrementally by the new modalities. In this regard, both Chapter 6 and Chapter 7 present two models which are applicable for multi-view dynamic data analysis. In Chapter 6, a new MCCA, termed as incremental MCCA (IMCCA) has been presented, which can update its solutions adaptively wherever a new modality is available for the analysis. It deals with the "curse of dimensionality" problem due to high-dimension low-sample size characteristics of real-life multimodal data sets, by using the ridge regression optimization technique with shrinkage estimation. Using the proposed IMCCA model, a new feature extraction algorithm, termed as SeFGeIM has been introduced, which considers a new modality for the analysis if it has relevant and significant information with respect to existing modalities. The quality of the extracted features depends on the supervised information of sample categories. Analytical formulation facilitates the generation of relevant and significant features from multiblock dynamic data sets with significantly lower computational costs. Both the algorithms presented in Chapter 5 and Chapter 6 are based on the sum of correlations (SUMCOR) criterion of the MCCA, which has higher cost as compared to the maximum variance (MAXVAR) criterion of the MCCA. Moreover, they do not consider the geometry of the multi-view data. In this regard, a new supervised feature extraction algorithm, termed as GraDiM, has been presented in Chapter 7, which integrates dynamic multi-view data sets by using the MAXVAR criterion and the knowledge of the graph. It can update the solutions adaptively wherever a new view is available for the analysis, without repeating the same procedure with the original data augmented by the new view. Moreover, the algorithm is designed in such a way that if all the views are available at the beginning of the analysis, the algorithm starts with the three most relevant views, and the remaining views are added sequentially according to their relevance. The proposed GraDiM algorithm addresses the singularity issue of the covariance matrices by using the ridge regression optimization technique. The optimum regularization parameters for the proposed algorithm are estimated based on the supervised information of sample categories. Analytical formulation facilitates the generation of relevant and significant features from multi-view dynamic data sets with significantly lower computational costs. The effectiveness of the proposed SeFGeIM algorithm of Chapter 6 and GraDiM algorithm of Chapter 7, along with a comparison with other algorithms, has been demonstrated on several multi-omics cancer and benchmark data sets.

In brief, the concept of incremental MCCA proposed in this thesis is unique.

8.2 Future Directions

There are various key characteristics of the research presented in this thesis that can be extended further for the progress of multi-view data analysis. Some improvements and future directions are reported next with which the research can be continued.

- **Incomplete views**: All the proposed multi-view data integration algorithms presented in the thesis assume that all the views have the same set of common samples. But, in practical application, it may happen that the data set has incomplete views due to pre-processing and measurement errors. For example, in Web analysis, some websites may contain texts, pictures, and videos, but others may contain one or two types only, which produces data with missing views. Multi-view data integration algorithms are supposed to work with incomplete views as well. Hence, all the multi-view data integration algorithms presented in the thesis can be extended such that by establishing a connection between the views the missing sample can be restored with the help of the complete views [299] without discarding the missing sample from all views.

- **Non-linear or kernel learning**: All the algorithms presented in the thesis are based on the linear relationship among different views. A kernel function generally transforms data points into a high (possibly infinite) dimensional space and returns the inner-product between two points in a standard feature dimension [265]. All the proposed algorithms can be extended in such a way that non-linear correlation may be achieved during the integration process. It could be done by mapping the data sets into a very high-dimensional Hilbert space using a non-linear transformation and then correlated subspaces are obtained.

- **Manifold learning based optimization**: Many real-life data sets have meaningful structures that lie on a low-dimensional manifold embedded in a higher-dimensional Euclidean space [236]. As each view can have a separate manifold, the multi-view data set can be considered as a mixture of manifolds. Based on this hypothesis, all the multi-view data integration algorithms presented in the thesis can be extended where correlated subspaces are constructed from a low-dimensional manifold.

- **Deep network based optimization**: The thesis focuses on spectral decomposition-based analysis and shallow optimization solutions. On the other hand, deep learning architectures such as the multimodal deep Boltzmann machines [247] can learn non-linear transformations from multiple modalities. As the extensions of the proposed algorithms, an attempt will be made to integrate multi-view data sets using a deep learning framework to learn maximally correlated subspaces, where the spectral decomposition-based solutions can help to initialize the deep optimization model.

- **Supervised information during learning**: Both CCA and MCCA are unsupervised methods. All the multi-view data integration algorithms presented in this thesis consider the class label information during the evaluation of each extracted feature. Instead of considering the supervised knowledge in the feature selection process, the class information can be taken during the learning process also. It may be done by computing the within-class covariance matrix.

- **Views observed in heterogeneous measurement spaces**: All the multi-view data integration algorithms presented in the thesis assume that each of the views is observed in a real-valued Euclidean space, and hence is provided in feature space-based representation. However, in practical application, it may be possible that some of the views may not be embedded in real-valued space. They may consist of textual, integer count, or categorical information. The proposed multi-view algorithms presented in this thesis should be modified such that heterogeneous multi-view data where different views are embedded in different measurement spaces can be integrated.

- **Tensor spectral analysis**: Tensor is the extension of matrix factorization in multi-view data analysis. It is used to capture higher-order correlations among multiple views [47, 297, 298]. Hence, generalization of CCA for more than two views can be done by tensor [169]. As none of the proposed algorithms consider tensors to analyze multi-view data set, it can be evident extension where tensor spectral is used to compute maximally correlated canonical variables.

- **Regression optimization**: All the multi-view data integration algorithms presented in the thesis are based on the assumption that noise in the data sets is Gaussian, independent, and identically distributed. For this reason, there is a non-zero variance for each feature and the covariance is 0. Thus, ridge regression optimization can address this issue. It also takes care of the invertibility problem of the covariance matrix of each high-dimensional variable. However, sometimes real-world data may fail to satisfy the mixture of Gaussian assumptions. Hence, other optimization techniques may be used to overcome this situation. The singularity issue of the covariance matrices can also be addressed in different ways.

- **Sparse data integration**: The algorithms presented in this thesis do not take care of the sparsity of the data sets directly. However, sparse features can cause problems like overfitting and suboptimal results in learning models. Hence, the data integration algorithms should be extended in such a way that sparse multi-view data sets can be integrated efficiently.

- **View ranking**: Although both SeFGeIM and GraDiM, presented in Chapter 6 and Chapter 7, respectively, integrate views according to their relevance, a normality-based measure of relevance of an individual modality and an orthogonality-based measure of shared information or dependency between two modalities can be considered to rank the views [136, 137]. As an extension of SeFGeIM or GraDiM, an attempt will be made to integrate the multi-view data sets according to the ranking of the views.

Appendix A

Description of Data Sets

The appendix presents a brief description of the multi-view benchmark and multi-omics cancer data sets used in the thesis for comparative analysis of the proposed and the existing multi-view clustering algorithms. Five benchmark data sets, namely, CiteSeer, Handwritten, NUS-WIDE-OBJECT (NW-OBJECT), Reuters, and Caltech; and five cancer data sets, namely, glioblastoma multiforme (GBM), lung (LUNG), kidney (KIDNEY), lower grade glioma (LGG), and ovarian serous cystadenocarcinoma (OV), are used in the current work.

A.1 Benchmark Data Sets

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

1. **CiteSeer**: The CiteSeer database is obtained from http://networkrepository.com. The set is generated by sampling scientific documents from CiteSeer digital library. The publications are classified into one of the six classes, namely, Agents, AI, DB, IR, ML, and HCI. There are 3312 papers in the data set. The papers are selected in a way such that in the final set every paper cites or is cited by at least one other paper. After stemming and removing the stopwords, a vocabulary of size 3703 unique words is obtained. All the words with document frequency less than 10 are removed. Each publication in the database is described by a 0 or 1 valued word vector indicating the absence or presence of the corresponding word in the document.

2. **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, created by Lab for Media Search in National University of Singapore, is intended for object recognition tasks. It consists of 30000 images categorized into 31 different classes. The 30000 images of the database are represented in terms of the five feature sets.

3. **Reuters**: This multilingual data has been downloaded from http://archive.ics.uci.edu/ml/machine-learning-databases/00259/. The collection contains feature characteristics of documents originally written in the English language and

Table A.1: Description of Benchmark Data Sets

Data Sets	# Classes	# Samples	Different Modalities	# Features
CiteSeer	6	3312	1st View	3703
			2nd View	3312
			3rd View	3309
			4th View	3312
NUS-WIDE-OBJECT	31	30000	Color histogram	64
			Block-wise color moments	225
			Color correlogram	144
			Edge direction histogram	73
			Wavelet texture	128
Reuters	6	18758	Original English documents	21531
			French documents translated to English	24892
			German documents translated to English	34251
			Italian documents translated to English	15506
			Spanish documents translated to English	11547
Handwritten	10	2000	Pixel averages in 2 x 3 windows	240
			Fourier coefficients of the character shapes	76
			Profile correlations	216
			Zernike moment	47
			Karhunen-Love coefficients	64
			Morphological features	6
Caltech	20	2386	Gabor feature	48
			Wavelet moments	40
			CENTRIST feature	254
			Histogram of oriented gradients feature	1984
			GIST feature	512
			Local binary patterns feature	928

the corresponding translations in French, German, Spanish, and Italian languages over a common set of 6 categories. This collection can be used for multilingual categorization and multi-view learning research. Documents have been translated, preprocessed, and are made available as feature characteristics in a "bag of words" format. 18758 documents are partitioned into 6 categories which include CCAT, C15, ECAT, E21, GCAT, and M11; and represented in terms of the following five feature sets.

4. **Handwritten**: This data has been downloaded from `https://archive.ics.uci.edu/ml/datasets/Multiple+Features`. This dataset consists of features of handwritten numerals ('0'–'9') extracted from a collection of Dutch utility maps. 200 patterns per class (for a total of 2,000 patterns) have been digitized in binary images. These digits are represented in terms of the six feature sets. Though Handwritten data set has six modalities, one of them has only six features. As the MCCA based methods can generate $\min(m_1, m_2, \cdots, m_{\mathcal{M}})$ features at most, so the modality with six features, is not considered in all the MCCA based methods including the proposed algorithm. Both MvDA and MvDA-VC have the same property. Hence, that modality is also excluded in the integration process of Handwritten data set using MvDA and MvDA-VC. On the other hand, deep learning based methods do not hold this characteristic. Thus, in deep learning based methods, all six modalities are considered to extract features from Handwritten data set.

5. **Caltech**: This data has been downloaded from `http://www.vision.caltech.edu/Image_Datasets/Caltech101/`. Caltech-101 consists of pictures of objects belonging to 101 categories. There are 40 to 800 images per category. Most categories have about 50 images, collected in September 2003 by Fei-Fei Li, Marco Andreetto, and Marc Aurelio Ranzato. The size of each image is roughly 300× 200 pixels. Caltech-20 is a subset of Caltech-101, which contains only 20 classes. In the current research work, Caltech-20 is used to analyze the performance of the proposed algorithm.

The summary of each data set, in terms of the sample size, dimension of different views and the number of classes, is provided in Table A.1.

A.2 Omics Data Sets

This section presents a brief description of the five multi-view omics data sets of The Cancer Genome Atlas (TCGA) [2]. All the data sets have been downloaded from the Genomic Data Commons (GDC) Data Portal [1]. A brief description of the four cancer data sets used in this work is presented next.

1. **Glioblastoma multiforme (GBM):** It is the most common and malignant form of brain cancer and has four subtypes identified in the study by Veerhak *et al.* [72]. The subtypes are proneural, neural, classical, and mesenchymal. The updated 2016 World Health Organization (WHO) classification of tumors of the central nervous system reflects a refinement of tumor diagnostics by integrating the genotypic and phenotypic features, thereby narrowing the defined subgroups. The new classification

Table A.2: Description of Cancer Data Sets

Data Sets	Different Classes	# Samples	Different Modalities	# Features
GBM	Proneural Classical G-CIMP Mesenchymal Neural	39 52 21 64 37	miRNA Expression Gene Expression DNA Methylation Copy Number Segmentation	534 12042 21422 4070
LUNG	Lung Adenocarcinoma Lung Squamous Cell Carcinoma	312 234	Protein Expression miRNA Sequence RNA Sequence DNA Methylation Copy Number Segmentation	180 216 20502 294668 49230
KIDNEY	Kidney Renal Clear Cell Carcinoma Kidney Renal Papillary Cell Carcinoma	95 210	Protein Expression miRNA Sequence RNA Sequence DNA Methylation Copy Number Segmentation	174 209 20502 300451 9059
LGG	IDHmut-non-codel IDHmut-codel IDHwt	180 129 65	Protein Expression miRNA Sequence RNA Sequence DNA Methylation Copy Number Segmentation	181 139 11973 293965 6261
OV	Proliferative Differentiated Immunoreactive Mesenchymal	60 35 47 64	Protein Expression miRNA Sequence Gene Expression DNA Methylation Copy Number Segmentation	195 129 12042 20311 4332

recommends molecular diagnosis of isocitrate dehydrogenase (IDH) mutational status in gliomas. Using TCGA data, Noushmehr *et al.* [71] identified a subset of GBM tumors with characteristic promoter DNA methylation alterations, referred to as a glioma cytosine-phosphate-guanine (CpG) island methylator phenotype (G-CIMP). G-CIMP tumors have distinct molecular features, including a high frequency of IDH1 mutation and characteristic copy-number alterations. Patients with G-CIMP tumors are younger at diagnosis and display improved survival times. The molecular alterations in G-CIMP tumors define a distinct subset of human gliomas with specific clinical features. G-CIMP tumors belong to the proneural subgroup. In order to obtain an integrated view of the relationships of G-CIMP status and gene expression differences, Noushmehr *et al.* performed pairwise comparisons between members of different molecular subgroups. In this paper Noushmehr *et al.* calculated the mean Euclidean distance in both DNA methylation and expression for each possible pairwise combination of the five different subtypes: G-CIMP-positive proneural, G-CIMP-negative proneural, classical, mesenchymal, and neural tumors. Hence, in the current research work, G-CIMP is also used as a class label along with proneural, neural, classical, and mesenchymal; where G-CIMP-positive proneural and G-CIMP-negative proneural are considered as G-CIMP and proneural, respectively. 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.

2. **Lung Carcinoma (LUNG):** There are two subtypes, namely, lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD) are present in the current research work, based on the same primary site of origin. According to the 2015 WHO lung cancer classification [272], these had been the two major subtypes. The total number of samples in LUSC and LUAD are 234 and 312, respectively.

3. **Kidney Carcinoma (KIDNEY):** The kidney cancer data set has two histological subtypes, namely, renal clear cell carcinoma (KIRC) and renal papillary cell carcinoma (KIRP). These subtypes were included in the 2004 WHO classification of adult renal tumors [217]. The KIDNEY data set consists of 305 samples of kidney cancer with 95 samples of KIRC and 210 samples of KIRP.

4. **Lower Grade Glioma (LGG):** According to World Health Organization, lower-grade glioma is grades II and III, which is made up of diffuse low-grade and intermediate-grade gliomas. As LGG has highly variable clinical behavior, it is very difficult to predict LGG based on histologic class [202]. Some are indolent; others quickly progress to glioblastoma. In the current research work, 374 LGG samples are used to analyze the performance of each algorithm. The first subtype exhibits IDH mutation with no 1p/19q codeletion and has 180 samples. The second subtype has 129 samples that exhibit both IDH mutation and 1p/19q codeletion. The wild-type IDH subtype is the third subtype, which has 65 samples.

5. **Ovarian Serous Cystadenocarcinoma (OV):** Ovarian serous cystadenocarcinoma is the malignant form of ovarian serous tumor, which is the most common type of ovarian epithelial tumor. It is the eighth-most commonly occurring cancer in women. According to [32], there were nearly 300,000 new cases in 2018. In [201], four

subtypes are identified, which are used in the current research work. These subtypes are proliferative, differentiated, mesenchymal, and immunoreactive, which consist of 60, 35, 64, and 47 samples, respectively.

These subtypes are clinically relevant and provide a roadmap for patient stratification and trials of targeted therapies. The information of DNA methylation (mDNA) has been utilized in all omics data sets. On the other hand, reverse phase protein array expression (RPPA) is common for LUNG, KIDNEY, LGG, and OV data sets. The aforementioned data sets make use of microRNA (miRNA) in sequence form, but GBM takes the knowledge of miRNA in expression form. The details of the gene (RNA) have been taken from RNA sequences in LUNG, KIDNEY, and LGG data sets, while gene expression provides gene-related information in the GBM and OV data sets. The information of Copy number segmentation (CNS) is used in GBM, LUNG, KIDNEY, LGG, and OV data sets.

The thesis addresses the problem associated with the high-dimension low-sample issue of real-life data sets. Thus, all five omics data sets have this kind of property, where the number of features is large enough, and the number of samples is very small. On the other hand, different benchmark data sets are used to study the other possible situations. For example, CiteSeer, Handwritten, and Caltech data sets have a moderate number of features as well as samples. NUS-WIDE-OBJECT (NW-OBJECT) data set addresses the situation where the number of samples is huge, but the number of features is undersized; while in case of Reuters both the samples and features are large. According to the experimental analysis, the proposed algorithms work well for not only the high-dimension low-sample case, but also the other situations.

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 [193] 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 [193], 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. The summary of each data set, in terms of the dimension of different modalities, the number of classes, and the sample size is provided in Table A.2.

Appendix B

More Results Using F1 Score

This section provides a brief description of F1 score that are used to validate the performance of the proposed algorithms presented in Chapter 3, Chapter 4, Chapter 5, Chapter 6, and Chapter 7. In case of classification, the F1 score is the harmonic mean of the precision and recall [232], that is,

$$\text{F1 score} = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}; \tag{B.1}$$

where, the precision is the number of true positive results divided by the number of all positive results, including those not identified correctly, and the recall is the number of true positive results divided by the number of all samples that should have been identified as positive [265] and can be defined as:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \text{and} \quad \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}; \tag{B.2}$$

where, TP, FP, and FN are known as true positive, false positive, and false negative, respectively. TP, FP, and FN denote a test result that correctly indicates the presence of a condition or characteristic, a test result which wrongly indicates that a particular condition or attribute is present, and a test result which wrongly indicates that a particular condition or attribute is absent, respectively. Precision is also known as positive predictive value, and recall is also known as sensitivity value. True negative rate is also known as specificity and can be defined as follows:

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}; \tag{B.3}$$

where, TN is known as true negative and denotes a test result that correctly indicates the absence of a condition or characteristic.

Table B.1, Table B.2, Table B.3, Table B.4, and Table B.5 present the F1 score of the five benchmark data sets, namely, CiteSeer, Handwritten, NUS-WIDE-OBJECT (NW-OBJECT), Reuters, and Caltech and five cancer data sets, namely, glioblastoma multiforme (GBM), lung (LUNG), kidney (KIDNEY), lower grade glioma (LGG) and ovarian serous

Table B.1: F1 Score of Different Algorithms on CiteSeer and GBM Data Sets

Different Algorithms		Data Sets	F1 Score (Train-Test)	F1 Score and Significance Analysis for 10-Fold CV					
				Mean	Median	StdDev	Paired-t:p	Wilcoxon:p	Friedman:p
MCCA	SUMCOR	CiteSeer	0.527	0.549	0.560	0.026	**1.48E-04**	**2.53E-03**	**1.57E-03**
	GENVAR		0.366	0.359	0.401	0.102	**3.28E-05**	**2.53E-03**	**1.57E-03**
	MAXVAR		0.533	0.519	0.391	0.022	**1.35E-04**	**2.53E-03**	**1.57E-03**
	MINVAR		0.470	0.482	0.514	0.023	**6.41E-07**	**2.53E-03**	**1.57E-03**
	SSQCOR		0.487	0.501	0.480	0.032	**6.37E-05**	**2.53E-03**	**1.57E-03**
	RGCCA		0.252	0.401	0.503	0.025	**2.87E-08**	**2.53E-03**	**1.57E-03**
	GMCCA		0.154	0.423	0.422	0.031	**5.14E-09**	**2.53E-03**	**1.57E-03**
	GMKCCA		0.165	0.150	0.141	0.041	**1.97E-10**	**2.53E-03**	**1.57E-03**
	LasCCA		0.209	0.456	0.450	0.023	**1.09E-09**	**2.53E-03**	**1.57E-03**
	DisCCA		0.175	0.211	0.208	0.017	**1.69E-11**	**2.53E-03**	**1.57E-03**
	BsMCCA		0.203	0.050	0.193	0.000	**1.53E-10**	**2.53E-03**	**1.57E-03**
	MvDA		0.316	0.364	0.364	0.018	**1.52E-09**	**2.53E-03**	**1.57E-03**
	MvDA-VC		0.374	0.414	0.410	0.026	**1.09E-07**	**2.53E-03**	**1.57E-03**
	LiveGCANO		0.183	0.050	0.372	0.000	**1.89E-07**	**2.53E-03**	**1.57E-03**
	OPID		0.424	0.343	0.334	0.032	**6.91E-11**	**2.53E-03**	**1.57E-03**
	SAC		0.420	0.326	0.321	0.042	**6.78E-12**	**2.53E-03**	**1.57E-03**
	ReDMiCA		0.591	0.593	0.591	0.030	*5.31E-02*	*5.71E-02*	*2.06E-01*
	SeFGeIM		**0.600**	0.595	**0.595**	0.032	*9.46E-02*	*6.97E-02*	5.27E-01
	GraDiM		0.597	**0.598**	0.592	0.030	-	-	-
MCCA	SUMCOR	GBM	0.169	0.150	0.116	0.098	**3.70E-08**	**2.53E-03**	**1.57E-03**
	GENVAR		0.162	0.238	0.238	0.079	**4.89E-08**	**2.53E-03**	**1.57E-03**
	MAXVAR		0.394	0.610	0.455	0.109	**1.66E-03**	**4.67E-03**	**1.14E-02**
	MINVAR		0.377	0.554	0.496	0.120	**5.39E-04**	**3.46E-03**	**1.14E-02**
	SSQCOR		0.452	0.497	0.420	0.115	**1.06E-04**	**2.53E-03**	**1.57E-03**
	RGCCA		0.223	0.351	0.257	0.126	**6.75E-07**	**2.53E-03**	**1.57E-03**
	GMCCA		0.408	0.321	0.269	0.112	**2.50E-07**	**2.53E-03**	**1.57E-03**
	GMKCCA		0.240	0.194	0.183	0.096	**9.26E-08**	**2.53E-03**	**1.57E-03**
	LasCCA		0.559	0.631	0.405	0.094	**6.04E-03**	**4.67E-03**	**1.14E-02**
	DisCCA		0.360	0.335	0.258	0.138	**5.47E-06**	**2.53E-03**	**1.57E-03**
	BsMCCA		0.667	0.566	0.274	0.210	**1.09E-02**	**1.09E-02**	*5.78E-02*
	MvDA		0.644	0.718	0.614	0.078	*2.19E-01*	*2.54E-01*	1.00E+00
	MvDA-VC		0.744	0.692	0.625	0.086	**4.46E-02**	**4.63E-02**	*2.06E-01*
	LiveGCANO		0.533	0.566	0.275	0.105	**3.82E-04**	**2.53E-03**	**1.57E-03**
	OPID		0.670	0.744	0.736	0.054	*4.13E-01*	*4.80E-01*	5.27E-01
	SAC		0.663	0.685	0.655	0.097	**3.80E-02**	**4.63E-02**	*5.78E-02*
	ReDMiCA		0.742	0.729	**0.742**	0.045	*1.65E-01*	*1.42E-01*	*2.06E-01*
	SeFGeIM		0.712	0.745	0.681	0.162	*4.66E-01*	5.61E-01	1.00E+00
	GraDiM		**0.770**	**0.750**	0.704	0.070	-	-	-

cystadenocarcinoma (OV). All the data sets are briefly described in Appendix A. The randomly selected 50% samples from each class are used for training and the rest are used for testing purposes for each of the data sets. The 10-fold cross-validation is also performed on each data set to assess the performance of the proposed algorithm statistically. To analyze the statistical significance of the derived results, paired-t test (one-tailed), Wilcoxon signed rank test (one-tailed) and Friedman test (one-tailed), with a 95% confidence level, are used to compute the p-values. For each data set, 25 top-ranked correlated features are

Table B.2: F1 Score of Different Algorithms on Handwritten and LUNG Data Sets

Different Algorithms		Data Sets	F1 Score (Train-Test)	F1 Score and Significance Analysis for 10-Fold CV					
				Mean	Median	StdDev	Paired-t:p	Wilcoxon:p	Friedman:p
MCCA	SUMCOR	Handwritten	0.870	0.821	0.825	0.023	**1.41E-08**	**2.53E-03**	**1.57E-03**
	GENVAR		0.077	0.095	0.907	0.072	**1.92E-11**	**2.53E-03**	**1.57E-03**
	MAXVAR		0.048	0.066	0.065	0.047	**2.97E-13**	**2.53E-03**	**1.57E-03**
	MINVAR		0.129	0.103	0.047	0.044	**7.37E-13**	**2.53E-03**	**1.57E-03**
	SSQCOR		0.089	0.084	0.117	0.040	**1.59E-14**	**2.53E-03**	**1.57E-03**
RGCCA			0.904	0.911	0.093	0.013	**2.17E-05**	**3.46E-03**	**1.14E-02**
GMCCA			0.096	0.101	0.096	0.029	**5.22E-14**	**2.53E-03**	**1.57E-03**
GMKCCA			0.061	0.091	0.102	0.046	**4.01E-13**	**2.53E-03**	**1.57E-03**
LasCCA			0.098	0.069	0.063	0.020	**6.16E-16**	**2.53E-03**	**1.57E-03**
DisCCA			0.057	0.134	0.126	0.063	**1.92E-11**	**2.53E-03**	**1.57E-03**
BsMCCA			0.114	0.116	0.100	0.061	**3.96E-12**	**2.53E-03**	**1.57E-03**
MvDA			0.925	0.947	0.954	0.021	**1.27E-03**	**2.53E-03**	**1.57E-03**
MvDA-VC			0.935	0.955	0.953	0.013	**2.34E-02**	**2.34E-02**	*5.78E-02*
LiveGCANO			0.084	0.099	0.094	0.040	**1.44E-13**	**2.53E-03**	**1.57E-03**
OPID			0.937	0.953	0.958	0.021	**7.43E-03**	**1.09E-02**	**1.14E-02**
SAC			0.941	0.949	0.949	0.016	**2.22E-03**	**3.46E-03**	**1.14E-02**
ReDMiCA			0.964	**0.969**	0.970	0.016	1.00E+00	1.00E+00	1.00E+00
SeFGeIM			**0.966**	**0.969**	**0.971**	0.009	*4.88E-01*	6.01E-01	5.27E-01
GraDiM			**0.966**	**0.969**	0.970	0.016	-	-	-
MCCA	SUMCOR	LUNG	0.484	0.498	0.492	0.035	**1.95E-10**	**2.53E-03**	**1.57E-03**
	GENVAR		0.364	0.556	0.864	0.160	**1.85E-05**	**2.52E-03**	**1.57E-03**
	MAXVAR		0.752	0.686	0.575	0.143	**8.13E-05**	**2.53E-03**	**1.57E-03**
	MINVAR		0.781	0.704	0.689	0.139	**1.29E-04**	**2.53E-03**	**1.57E-03**
	SSQCOR		0.786	0.723	0.716	0.066	**3.35E-06**	**2.53E-03**	**1.57E-03**
RGCCA			0.876	0.875	0.730	0.042	**5.16E-05**	**2.53E-03**	**1.57E-03**
GMCCA			0.677	0.684	0.697	0.069	**2.99E-08**	**2.53E-03**	**1.57E-03**
GMKCCA			0.861	0.861	0.875	0.083	**1.55E-03**	**3.46E-03**	**1.14E-02**
LasCCA			0.821	0.853	0.855	0.064	**5.49E-05**	**2.53E-03**	**1.57E-03**
DisCCA			0.510	0.508	0.512	0.098	**1.80E-08**	**2.53E-03**	**1.57E-03**
BsMCCA			0.891	0.827	0.891	0.161	**1.31E-02**	**1.04E-02**	**3.39E-02**
MvDA			0.921	0.947	0.964	0.042	*5.68E-02*	**3.71E-02**	*2.06E-01*
MvDA-VC			0.914	0.955	0.956	0.034	*8.60E-02*	*1.42E-01*	*2.06E-01*
LiveGCANO			0.502	0.529	0.526	0.056	**4.49E-10**	**2.53E-03**	**1.57E-03**
OPID			0.948	0.942	0.947	0.040	**1.07E-02**	**7.58E-03**	**1.96E-02**
SAC			0.963	0.949	0.946	0.037	*5.76E-02*	*5.49E-02*	*3.17E-01*
ReDMiCA			0.948	0.957	0.955	0.032	**2.35E-02**	**2.07E-02**	*5.78E-02*
SeFGeIM			0.940	0.962	0.955	0.025	*1.24E-01*	*9.59E-02*	*3.17E-01*
GraDiM			**0.970**	**0.975**	**0.969**	0.030	-	-	-

selected for the analysis.

The proposed algorithm ReDMiCA, SeFGeIM, and GraDiM presented in Chapter 5, Chapter 6, and Chapter 7, respectively are compared with (i) different criteria of the MCCA, namely, SUMCOR, MAXVAR, generalized variance (GENVAR), minimum variance (MINVAR), and sum of squared correlations (SSQCOR) [135]; (ii) various state-of-the-art MCCA based methods, namely, RGCCA [262], GMCCA [43], GMKCCA [43], large-scale generalized CCA (LasCCA) [84], distributed generalized CCA (DisCCA) [84],

Table B.3: F1 Score of Different Algorithms on NW-OBJECT and KIDNEY Data Sets

Different Algorithms		Data Sets	F1 Score (Train-Test)	F1 Score and Significance Analysis for 10-Fold CV					
				Mean	Median	StdDev	Paired-t:p	Wilcoxon:p	Friedman:p
MCCA	SUMCOR	NW-OBJECT	0.135	0.152	0.153	0.008	**5.11E-09**	**2.53E-03**	**1.57E-03**
	GENVAR		0.033	0.030	0.015	0.005	**8.01E-13**	**2.53E-03**	**1.57E-03**
	MAXVAR		0.028	0.031	0.029	0.008	**1.60E-12**	**2.53E-03**	**1.57E-03**
	MINVAR		0.032	0.030	0.032	0.005	**2.99E-13**	**2.53E-03**	**1.57E-03**
	SSQCOR		0.032	0.032	0.029	0.005	**4.82E-14**	**2.53E-03**	**1.57E-03**
RGCCA			0.053	0.015	0.032	0.001	**3.80E-13**	**2.53E-03**	**1.57E-03**
GMCCA			0.027	0.031	0.032	0.004	**7.73E-13**	**2.53E-03**	**1.57E-03**
GMKCCA			0.018	0.023	0.023	0.005	**4.92E-13**	**2.53E-03**	**1.57E-03**
LasCCA			0.039	0.048	0.048	0.006	**1.08E-12**	**2.53E-03**	**1.57E-03**
DisCCA			0.028	0.024	0.023	0.007	**8.29E-13**	**2.53E-03**	**1.57E-03**
BsMCCA			0.047	0.040	0.041	0.003	**1.30E-12**	**2.53E-03**	**1.57E-03**
MvDA			0.145	0.125	0.124	0.012	**1.35E-09**	**2.53E-03**	**1.57E-03**
MvDA-VC			0.135	0.121	0.119	0.013	**3.39E-10**	**2.53E-03**	**1.57E-03**
LiveGCANO			0.019	0.023	0.020	0.008	**7.04E-13**	**2.53E-03**	**1.57E-03**
OPID			0.125	0.116	0.115	0.009	**3.00E-10**	**2.53E-03**	**1.57E-03**
SAC			0.119	0.114	0.114	0.007	**1.63E-10**	**2.53E-03**	**1.57E-03**
ReDMiCA			0.241	0.241	0.242	0.011	**5.92E-03**	**1.42E-02**	*2.06E-01*
SeFGeIM			0.260	0.256	0.257	0.016	*2.96E-01*	*2.88E-01*	5.27E-01
GraDiM			**0.274**	**0.260**	**0.260**	0.013	-	-	-
MCCA	SUMCOR	KIDNEY	0.487	0.539	0.551	0.063	**1.10E-08**	**2.53E-03**	**1.57E-03**
	GENVAR		0.409	0.541	0.910	0.131	**9.76E-07**	**2.53E-03**	**1.57E-03**
	MAXVAR		0.728	0.749	0.531	0.060	**7.15E-07**	**2.53E-03**	**1.57E-03**
	MINVAR		0.691	0.752	0.742	0.059	**1.98E-07**	**2.53E-03**	**1.57E-03**
	SSQCOR		0.728	0.821	0.759	0.098	**1.15E-03**	**3.44E-03**	**1.14E-02**
RGCCA			0.898	0.920	0.837	0.054	**1.32E-02**	**1.77E-02**	**3.39E-02**
GMCCA			0.826	0.706	0.727	0.088	**8.85E-06**	**2.53E-03**	**1.57E-03**
GMKCCA			0.834	0.822	0.846	0.074	**6.06E-05**	**2.53E-03**	**1.57E-03**
LasCCA			0.686	0.791	0.790	0.094	**8.80E-05**	**2.53E-03**	**1.57E-03**
DisCCA			0.518	0.500	0.508	0.090	**1.75E-08**	**2.53E-03**	**1.57E-03**
BsMCCA			0.825	0.895	0.890	0.055	**2.01E-03**	**3.42E-03**	**1.14E-02**
MvDA			0.917	0.925	0.927	0.027	**2.46E-04**	**3.82E-03**	**2.70E-03**
MvDA-VC			0.938	0.935	0.931	0.029	**8.53E-03**	**2.53E-02**	**1.96E-02**
LiveGCANO			0.396	0.562	0.550	0.127	**1.41E-06**	**2.53E-03**	**1.57E-03**
OPID			0.941	0.960	0.963	0.026	*1.50E-01*	*5.79E-02*	*4.14E-01*
SAC			**0.969**	0.942	0.930	0.032	**6.45E-03**	**1.37E-02**	*5.88E-02*
ReDMiCA			0.954	0.968	0.964	0.031	*3.60E-01*	*3.58E-01*	*3.17E-01*
SeFGeIM			0.961	0.967	0.949	0.021	*3.01E-01*	*1.11E-01*	6.55E-01
GraDiM			0.954	**0.972**	**0.968**	0.028	-	-	-

and block sparse MCCA (BsMCCA) [235]; (iii) two popular multidimensional data integration algorithms, namely, multi-view discriminant analysis (MvDA) [128] and MvDA with view-consistency (MvDA-VC) [129]; (iv) three multi-view incremental algorithms, namely, live generalized canonical correlation analysis (LiveGCANO) [187], one-pass learning with incremental and decremental features (OPID) [114], and safe classification with augmented features (SAC) [113].

All the results reported in Table B.1, Table B.2, Table B.3, Table B.4, and Table B.5

Table B.4: F1 Score of Different Algorithms on Reuters and LGG Data Sets

Different Algorithms		Data Sets	F1 Score (Train-Test)	F1 Score and Significance Analysis for 10-Fold CV					
				Mean	Median	StdDev	Paired-t:p	Wilcoxon:p	Friedman:p
MCCA	SUMCOR	Reuters	0.517	0.605	0.609	0.015	**2.89E-06**	**2.53E-03**	**1.57E-03**
	GENVAR		0.129	0.158	0.165	0.027	**8.93E-12**	**2.53E-03**	**1.57E-03**
	MAXVAR		0.219	0.187	0.149	0.046	**3.71E-10**	**2.53E-03**	**1.57E-03**
	MINVAR		0.171	0.168	0.173	0.020	**3.17E-12**	**2.53E-03**	**1.57E-03**
	SSQCOR		0.142	0.189	0.171	0.025	**9.76E-12**	**2.53E-03**	**1.57E-03**
RGCCA			0.316	0.165	0.187	0.004	**3.92E-14**	**2.53E-03**	**1.57E-03**
GMCCA			0.203	0.252	0.253	0.033	**1.99E-10**	**2.53E-03**	**1.57E-03**
GMKCCA			0.252	0.203	0.197	0.019	**4.83E-13**	**2.53E-03**	**1.57E-03**
LasCCA			0.237	0.254	0.269	0.038	**5.07E-11**	**2.53E-03**	**1.57E-03**
DisCCA			0.132	0.147	0.136	0.050	**1.90E-10**	**2.53E-03**	**1.57E-03**
BsMCCA			0.607	0.346	0.340	0.039	**2.38E-09**	**2.53E-03**	**1.57E-03**
MvDA			0.498	0.518	0.525	0.021	**8.89E-09**	**2.53E-03**	**1.57E-03**
MvDA-VC			0.491	0.514	0.516	0.024	**8.25E-09**	**2.53E-03**	**1.57E-03**
LiveGCANO			0.174	0.025	0.020	0.016	**1.92E-14**	**2.53E-03**	**1.57E-03**
OPID			0.499	0.539	0.534	0.020	**1.45E-08**	**2.53E-03**	**1.57E-03**
SAC			0.498	0.527	0.524	0.023	**1.99E-07**	**2.53E-03**	**1.57E-03**
ReDMiCA			**0.614**	0.647	0.645	0.010	**1.44E-02**	**1.42E-02**	*5.78E-02*
SeFGeIM			**0.614**	**0.660**	**0.661**	0.014	*4.87E-01*	6.01E-01	5.27E-01
GraDiM			0.608	**0.660**	**0.661**	0.019	-	-	-
MCCA	SUMCOR	LGG	0.305	0.341	0.315	0.101	**1.45E-05**	**2.53E-03**	**1.57E-03**
	GENVAR		0.217	0.225	0.465	0.025	**2.76E-07**	**2.53E-03**	**1.57E-03**
	MAXVAR		0.376	0.285	0.214	0.057	**1.99E-06**	**2.53E-03**	**1.57E-03**
	MINVAR		0.372	0.285	0.298	0.056	**2.86E-07**	**2.53E-03**	**1.57E-03**
	SSQCOR		0.380	0.318	0.270	0.054	**2.00E-06**	**2.53E-03**	**1.57E-03**
RGCCA			0.361	0.428	0.326	0.123	**6.88E-05**	**2.53E-03**	**1.57E-03**
GMCCA			0.356	0.345	0.309	0.102	**6.68E-06**	**2.53E-03**	**1.57E-03**
GMKCCA			0.330	0.307	0.302	0.047	**3.78E-07**	**2.53E-03**	**1.57E-03**
LasCCA			0.367	0.325	0.315	0.078	**6.01E-07**	**2.53E-03**	**1.57E-03**
DisCCA			0.280	0.306	0.288	0.050	**9.75E-07**	**2.53E-03**	**1.57E-03**
BsMCCA			0.653	0.729	0.726	0.055	**2.24E-02**	**3.72E-02**	**1.14E-02**
MvDA			0.785	0.793	0.795	0.069	*1.16E-01*	**3.72E-02**	**1.14E-02**
MvDA-VC			0.770	0.837	0.822	0.071	*4.46E-01*	*1.01E-01*	*2.06E-01*
LiveGCANO			0.297	0.412	0.379	0.071	**1.83E-05**	**3.46E-03**	**1.14E-02**
OPID			0.801	**0.870**	0.857	0.046	7.16E-01	6.39E-01	1.00E+00
SAC			0.914	0.865	**0.889**	0.052	6.51E-01	*1.93E-01*	*2.06E-01*
ReDMiCA			**0.953**	0.852	0.839	0.043	5.55E-01	*2.22E-01*	5.27E-01
SeFGeIM			**0.953**	0.844	0.774	0.056	*4.93E-01*	*1.66E-01*	*2.06E-01*
GraDiM			0.950	0.845	0.861	0.152	-	-	-

confirm that ReDMiCA, SeFGeIM, and GraDiM attain the highest F1 score of training-testing in 2, 5, and 6 cases, respectively, out of total 10 casses each. The results corresponding to 10-fold CV indicate that the proposed ReDMiCA, SeFGeIM, and GraDiM algorithm attains the higher mean F1 score in 1, 2, and 9 cases and higher median F1 score in 1, 3, and 6 cases, respectively, out of total 10 cases each. Out of the total of 330 cases, the proposed GraDiM algorithm attains significantly better p-values (marked in bold) than existing MCCA based methods including existing incremental algoritm in

Table B.5: F1 Score of Different Algorithms on Caltech and OV Data Sets

Different		Data	F1 Score	F1 Score and Significance Analysis for 10-Fold CV					
Algorithms		Sets	(Train-Test)	Mean	Median	StdDev	Paired-t:p	Wilcoxon:p	Friedman:p
MCCA	SUMCOR	Caltech	0.189	0.413	0.404	0.052	**9.36E-09**	**2.53E-03**	**1.57E-03**
	GENVAR		0.199	0.178	0.025	0.033	**4.68E-12**	**2.53E-03**	**1.57E-03**
	MAXVAR		0.520	0.522	0.170	0.045	**1.62E-07**	**2.53E-03**	**1.57E-03**
	MINVAR		0.494	0.498	0.541	0.038	**1.16E-09**	**2.53E-03**	**1.57E-03**
	SSQCOR		0.499	0.532	0.494	0.042	**5.26E-08**	**2.53E-03**	**1.57E-03**
RGCCA			0.025	0.025	0.540	0.000	**5.18E-15**	**2.53E-03**	**1.57E-03**
GMCCA			0.028	0.025	0.026	0.010	**4.56E-15**	**2.53E-03**	**1.57E-03**
GMKCCA			0.082	0.095	0.094	0.030	**2.59E-13**	**2.53E-03**	**1.57E-03**
LasCCA			0.026	0.042	0.035	0.021	**3.20E-13**	**2.53E-03**	**1.57E-03**
DisCCA			0.041	0.051	0.050	0.015	**2.11E-13**	**2.53E-03**	**1.57E-03**
BsMCCA			0.575	0.610	0.598	0.057	**7.76E-04**	**3.46E-03**	**1.14E-02**
MvDA			0.522	0.619	0.609	0.039	**4.88E-04**	**2.53E-03**	**1.57E-03**
MvDA-VC			0.517	0.613	0.611	0.040	**3.41E-04**	**6.26E-03**	*5.78E-02*
LiveGCANO			0.026	0.034	0.034	0.014	**4.41E-15**	**2.53E-03**	**1.57E-03**
OPID			0.521	0.605	0.601	0.035	**7.72E-06**	**2.53E-03**	**1.57E-03**
SAC			0.556	0.613	0.623	0.041	**5.15E-04**	**2.53E-03**	**1.57E-03**
ReDMiCA			0.670	0.651	0.639	0.052	**1.48E-02**	**1.42E-02**	*5.78E-02*
SeFGeIM			0.752	0.668	0.662	0.021	**3.56E-03**	**4.67E-03**	**1.14E-02**
GraDiM			**0.764**	**0.683**	**0.689**	0.023	-	-	-
MCCA	SUMCOR	OV	0.165	0.201	0.183	0.101	**2.91E-07**	**2.53E-03**	**1.57E-03**
	GENVAR		0.172	0.247	0.305	0.075	**4.58E-07**	**2.53E-03**	**1.57E-03**
	MAXVAR		0.189	0.545	0.255	0.169	**4.00E-03**	**1.42E-02**	*2.06E-01*
	MINVAR		0.474	0.567	0.474	0.180	**1.14E-02**	**1.83E-02**	*5.78E-02*
	SSQCOR		0.462	0.563	0.517	0.120	**1.92E-03**	**4.67E-03**	**1.14E-02**
RGCCA			0.337	0.305	0.603	0.073	**2.38E-08**	**2.53E-03**	**1.57E-03**
GMCCA			0.236	0.267	0.223	0.131	**7.38E-06**	**2.53E-03**	**1.57E-03**
GMKCCA			0.389	0.287	0.279	0.151	**9.18E-07**	**2.53E-03**	**1.57E-03**
LasCCA			0.337	0.297	0.298	0.080	**2.15E-08**	**2.53E-03**	**1.57E-03**
DisCCA			0.238	0.256	0.272	0.105	**2.79E-07**	**2.53E-03**	**1.57E-03**
BsMCCA			0.720	0.634	0.671	0.181	**4.50E-02**	**4.63E-02**	*2.06E-01*
MvDA			0.473	0.641	0.666	0.112	**2.33E-02**	**4.63E-02**	5.27E-01
MvDA-VC			0.547	0.557	0.584	0.140	**1.51E-04**	**2.53E-03**	**1.57E-03**
LiveGCANO			0.229	0.266	0.243	0.076	**8.23E-07**	**2.53E-03**	**1.57E-03**
OPID			0.438	0.567	0.553	0.133	**3.14E-03**	**1.09E-02**	**1.14E-02**
SAC			0.613	0.618	0.591	0.132	**7.44E-03**	**6.26E-03**	**1.14E-02**
ReDMiCA			0.941	0.710	0.722	0.092	*8.80E-02*	**2.97E-02**	**1.14E-02**
SeFGeIM			**0.951**	0.738	0.616	0.102	*2.58E-01*	*2.88E-01*	5.27E-01
GraDiM			**0.951**	**0.765**	**0.749**	0.105	-	-	-

297 cases, considering 95% confidence level. On the other hand, the proposed GraDiM algorithm provides better but not significant p-values (marked in italics) in 28 cases.

The ReDMiCA extracts the most relevant and significant features sequentially and can handels the high-dimension low-sample issue of real-life data sets. On the other hand, SeFGeIM integrates only the relevant views for the analysis, while GraDiM incorporates the geometrical knowledge along with the categorical information of the data set. Thus, ReDMiCA, SeFGeIM, and GraDiM perform much better than existing algorithms.

List of Related Publications

International Journal Papers

- **Ankita Mandal** and Pradipta Maji. Adaptive Generalized Multi-View Canonical Correlation Analysis for Incrementally Update Multiblock Data. ***IEEE Transactions on Knowledge and Data Engineering***, pages 1-14, 2022. DOI:10.1109/TKDE.2022.3185399.

- **Ankita Mandal** and Pradipta Maji. Multiview Regularized Discriminant Canonical Correlation Analysis: Sequential Extraction of Relevant Features from Multiblock Data. ***IEEE Transactions on Cybernetics***, pages 1-13, 2022. DOI:10.1109/TCYB.2022.3155875.

- **Ankita Mandal** and Pradipta Maji. FaRoC: Fast and Robust Supervised Canonical Correlation Analysis for Multimodal Omics Data. ***IEEE Transactions on Cybernetics***, 48(4):1229-1241, April 2018. DOI:10.1109/TCYB.2017.2685625.

- Pradipta Maji and **Ankita Mandal**. Multimodal Omics Data Integration Using Max Relevance-Max Significance Criterion. ***IEEE Transactions on Biomedical Engineering***, 64(8):1841-1851, August 2017. DOI:10.1109/TBME.2016.2624823.

International Conference Papers

- **Ankita Mandal** and Pradipta Maji. Regularization and Shrinkage in Rough Set Based Canonical Correlation Analysis. ***Proceedings of International Joint Conference on Rough Sets (IJCRS2017)***, Olsztyn, Poland, pages 432-446, July 2017.

- **Ankita Mandal** and Pradipta Maji. A New Method to Address Singularity Problem in Multimodal Data Analysis. ***Proceedings of 7th International Conference on Pattern Recognition and Machine Intelligence (PReMI2017)***, Kolkata, India, pages 43-51, December 2017.

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] C. C. Aggarwal, A. Hinneburg, and D. A. Keim. On the Surprising Behavior of Distance Metrics in High Dimensional Spaces. In *Proceedings of the 8th International Conference on Database Theory*, ICDT '01, pages 420–434, Berlin, Heidelberg, 2001.

[6] P. Agius, Y. Ying, and C. Campbell. Bayesian Unsupervised Learning With Multiple Data Types. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–27, 2009.

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

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

[9] A. S. Al-Waisy, R. Qahwaji, S. S. Ipson, and S. Al-Fahdawi. A Multimodal Deep Learning Framework Using Local Feature Representations For Face Recognition. *Machince Vision and Applications*, 29:35–54, 2018.

[10] 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.

[11] M. Alioscha-Perez, M. C. Oveneke, and H. Sahli. SVRG-MKL: A Fast and Scalable Multiple Kernel Learning Solution for Features Combination in Multi-Class Classification Problems. *IEEE Transactions on Neural Networks and Learning Systems*, 31(5):1710–1723, 2020.

[12] 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.

[13] L. An, S. Yang, and B. Bhanu. Person Re-Identification by Robust Canonical Correlation Analysis. *IEEE Signal Processing Letters*, 22(8):1103–1107, 2015.

[14] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep Canonical Correlation Analysis. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 1247–1255, 2013.

[15] O. Arandjelovic. Discriminative Extended Canonical Correlation Analysis for Pattern Set Matching. *Machine Learning*, 94(3), 2014.

[16] R. Arora and K. Livescu. Kernel CCA for Multi-View Learning of Acoustic Features Using Articulatory Measurements. In *Proceedings of the Machine Learning in Speech and Language Processing*, pages 34–37, 2012.

[17] R. Arora and K. Livescu. Multi-View Learning with Supervision for Transformed Bottleneck Features. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2499–2503, 2014.

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

[19] F. R. Bach and M. I. Jordan. A Probabilistic Interpretation of Canonical Correlation Analysis. Technical report, Department of Statistics, University of California, Berkeley, USA, 2005.

[20] 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.

[21] 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.

[22] M. Belkin and P. Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, 15(6):1373–1396, 2003.

[23] R. Bellman. *Adaptive Control Processes.* Princeton University Press, Princeton, New Jersey, 1961.

[24] A. Benton, H. Khayrallah, B. Gujral, D. A. Reisinger, S. Zhang, and R. Arora. Deep Generalized Canonical Correlation Analysis. *arXiv:1702.02519*, 2017.

[25] A. Benton, H. Khayrallah, B. Gujral, D. A. Reisinger, S. Zhang, and R. Arora. Deep Generalized Canonical Correlation Analysis. In *Proceedings of the 4th Workshop on Representation Learning for NLP*, pages 1–6, 2019.

[26] S. Bickel and T. Scheffer. Multi-view Clustering. In *Proceedings of the Fourth IEEE International Conference on Data Mining*, pages 19–26, 2004.

[27] T. D. Bie and B. D. Moor. On the Regularization of Canonical Correlation Analysis. In *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation*, pages 785–790, 2003.

[28] 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.

[29] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, New York, 2006. ISBN: 978-0-387-31073-2.

[30] M. B. Blaschko, C. H. Lampert, and A. Gretton. Semi-Supervised Laplacian Regularization of Kernel Canonical Correlation Analysis. In *Machine Learning and Knowledge Discovery in Databases*, pages 133–145, 2008.

[31] 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.

[32] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal. Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, 68(6):394–424, 2018.

[33] M. W. Browne. The Maximum-Likelihood Solution in Inter-Battery Factor Analysis. *British Journal of Mathematical and Statistical Psychology*, 32(1):75–86, 1979.

[34] J. Cai and H. W. Sun. Convergence Rate of Kernel Canonical Correlation Analysis. *Science China Mathematics*, 54(10):2161–2170, 2011.

[35] K. A. L. Cao, I. Gonzalez, and S. Dejean. integrOmics: An R Package to Unravel Relationships Between Two Omics Datasets. *Bioinformatics*, 25(21):2855–2856, 2009.

[36] K. A. L. Cao, P. G. P. Martin, C. R. Granie, and P. Besse. Sparse Canonical Methods for Biological Data Integration: Application to a Cross-Platform Study. *BMC Bioinformatics*, 10(1):1–17, 2009.

[37] J. D. Carroll. Generalization of Canonical Correlation Analysis to Three or More Sets of Variables. In *Proceedings of 76th Annual Convention of the American Psychological Association*, pages 227–228, 1968.

[38] S. Chandar, M. M. Khapra, H. Larochelle, and B. Ravindran. Correlational Neural Networks. *Neural Computation*, 28(2):257–285, 2016.

[39] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan. Multi-view Clustering Via Canonical Correlation Analysis. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 129–136, 2009.

[40] H. Chen, Z. Chen, Z. Chai, B. Jiang, and B. Huang. A Single-Side Neural Network-Aided Canonical Correlation Analysis With Applications to Fault Diagnosis. *IEEE Transactions on Cybernetics*, pages 1–13, 2021.

[41] H. Chen, Z. Song, M. Qian, C. Bai, and X. Wang. Selection of Disease-specific Biomarkers by Integrating Inflammatory Mediators with Clinical Informatics in AE-COPD Patients: A Preliminary Study. *Journal of Cellular and Molecular Medicine*, 16(6):1286–1297, 2012.

[42] J. Chen and I. D. Schizas. Online Distributed Sparsity-Aware Canonical Correlation Analysis. *IEEE Transactions on Signal Processing*, 64(3):688–703, 2016.

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

[44] J. Chen, G. Wang, Y. Shen, and G. B. Giannakis. Canonical Correlation Analysis of Datasets with a Common Source Graph. *IEEE Transactions on Signal Processing*, 66(16):4398–4408, 2018.

[45] N. Chen, J. Zhu, and E. Xing. Predictive Subspace Learning for Multi-view Data: A Large Margin Approach. In *Advances in Neural Information Processing Systems*, volume 23, 2010.

[46] X. Chen, L. Han, and J. Carbonell. Structured Sparse Canonical Correlation Analysis. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22, pages 199–207, 2012.

[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] C. M. Christoudias, R. Urtasun, and T. Darrell. Multi-View Learning in the Presence of View Disagreement. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, UAI'08, pages 88–96, Arlington, Virginia, USA, 2008.

[49] D. Chu, L.-Z. Liao, M. K. Ng, and X. Zhang. Sparse Canonical Correlation Analysis: New Formulation and Algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):3050–3065, 2013.

[50] M. Chu and J. Watterson. On a Multivariate Eigenvalue Problem, Part I: Algebraic Theory and a Power Method. *SIAM Journal on Scientific Computing*, 14(5):1089–1106, 1993.

[51] H. Chun and S. 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.

[52] D. Cordes, M. Jin, T. Curran, and R. Nandy. Optimizing the Performance of Local Canonical Correlation Analysis in fMRI Using Spatial Constraints. *Human Brain Mapping*, 33(11):2611–2626, 2012.

[53] D. Cordes, M. Jin, T. Curran, and R. Nandy. The Smoothing Artifactof Spatially Constrained Canonical Correlation Analysis in FunctionalMRI. *International Journal of Biomedical Imaging*, pages 1–11, 2012.

[54] C. Cortes, M. Mohri, and A. Rostamizadeh. Algorithms for Learning Kernels Based on Centered Alignment. *Journal of Machine Learning Research*, 13(1):795–828, 2012.

[55] H. D. Couture, R. Kwitt, J. S. Marron, M. Troester, C. M. Perou, and M. Niethammer. Deep Multi-View Learning via Task-Optimal CCA. *arXiv preprint arXiv:1907.07739*, 2019.

[56] R. Cruz-Cano and M.-L. T. Lee. Fast Regularized Canonical Correlation Analysis. *Computational Statistics and Data Analysis*, 70:88–100, 2014.

[57] A. Damianou, C. Ek, M. Titsias, and N. Lawrence. Manifold Relevance Determination. In *Proceedings of International Conference on Machine Learning*, pages 145–152, 2012.

[58] A. Damianou, N. D. Lawrence, and C. H. Ek. Multi-view Learning as a Nonparametric Nonlinear Inter-Battery Factor Analysis. *Journal of Machine Learning Research*, 22:1–51, 2021.

[59] H. Dashtestani, R. Zaragoza, H. Pirsiavash, K. M. Knutson, R. Kermanian, J. Cui, J. D. Jr. Harrison, M. Halem, and A. Gandjbakhche. Canonical Correlation Analysis of Brain Prefrontal Activity Measured by Functional Near Infra-red Spectroscopy (fNIRS) During A Moral Judgment Task. *Behavioural Brain Research*, 359:73–80, 2019.

[60] G. P de Bruin. An Interbattery Factor Analysis of the Comrey Personality Scales and the 16 Personality Factor Questionnaire. *Journal of Industrial Psychology*, 26(3):4–7, 2000.

[61] J. P. Van de Geer. Linear Relations Among k Sets of Variables. *Psychometrika*, 49(1):79–94, 1984.

[62] 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.

[63] N. Desai, A.-K. Seghouane, and M. Palaniswami. Algorithms for Two Dimensional Multi Set Canonical Correlation Analysis. *Pattern Recognition Letters*, 111:101–108, 2018.

[64] T. Diethe, D. R. Hardoon, and J. Shawe-Taylor. Multiview Fisher Discriminant Analysis. In *NIPS 2008 workshop on Learning from Multiple Sources*, 2008.

[65] J. Donahue, P. Krahenbuhl, and T. Darrell. Adversarial Feature Learning. In *Proceedings of the 5th International Conference on Learning Representations*, 2017.

[66] L. Dong, Y. Zhang, R. Zhang, X. Zhang, D. Gong, P. A. Valdes-Sosa, P. Xu, C. Luo, and D. Yao. Characterizing Nonlinear Relationships in Functional Imaging Data Using Eigenspace Maximal Information Canonical Correlation Analysis (emiCCA). *NeuroImage*, 109:388–401, 2015.

[67] 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.

[68] M. L. Eaton and M. D. Perlman. The Non-Singularity of Generalized Sample Covariance Matrices. *The Annals of Statistics*, 1(4):710–717, 1973.

[69] N. E. D. Elmadany, Y. He, and L. Guan. Information Fusion for Human Action Recognition via Biset/Multiset Globality Locality Preserving Canonical Correlation Analysis. *IEEE Transactions on Image Processing*, 27(11):5275–5287, 2018.

[70] N. El Din Elmadany, Y. He, and L. Guan. Multiview Learning Via Deep Discriminative Canonical Correlation Analysis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2409–2413, 2016.

[71] H. Noushmehr et al. Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma. *Cancer Cell*, 17(5):510–522, 2010.

[72] 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.

[73] Q. Fan, Z. Wang, H. Zha, and D. Gao. MREKLM: A Fast Multiple Empirical Kernel Learning Machine. *Pattern Recognition*, 61:197–209, 2017.

[74] S. Fan, X. Wang, C. Shi, E. Lu, K. Lin, and B. Wang. One2Multi Graph Autoencoder for Multi-View Graph Clustering. In *Proceedings of the Web Conference 2020*, pages 3070–3076, 2020.

[75] Z. Fan and H. Lian. Minimax Convergence Rates for Kernel CCA. *Journal of Multivariate Analysis*, 150:183–190, 2016.

[76] F. Feng, X. Wang, and R. Li. Cross-Modal Retrieval with Correspondence Autoencoder. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 7–16, 2014.

[77] Y. Feng, Z. Zhang, X. Zhao, R. Ji, and Y. Gao. GVCNN: Group-View Convolutional Neural Networks for 3D Shape Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 264–272, 2018.

[78] R. A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[79] P. Flach. *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press, New York, 2012. ISBN: 978-1-107-09639-4.

[80] A. L. N. Fred and A. K. Jain. Combining Multiple Clusterings Using Evidence Accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):835–850, 2005.

[81] Y. Freund and R. E. Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[82] O. Friman, M. Borga, P. Lundberg, and H. Knutsson. Adaptive Analysis of fMRI Data. *NeuroImage*, 19(3):837–845, 2003.

[83] X. Fu, K. Huang, M. Hong, N. D. Sidiropoulos, and A. M. So. Scalable and Flexible Multiview MAX-VAR Canonical Correlation Analysis. *IEEE Transactions on Signal Processing*, 65(16):4150–4165, 2017.

[84] X. Fu, K. Huang, E. E. Papalexakis, H. Song, P. P. Talukdar, N. D. Sidiropoulos, C. Faloutsos, and T. Mitchell. Efficient and Distributed Algorithms for Large-Scale Generalized Canonical Correlations Analysis. In *Proceedings of the IEEE 16th International Conference on Data Mining*, pages 871–876, 2016.

[85] X. Fu, K. Huang, E. E. Papalexakis, H. A. Song, P. Talukdar, N. D. Sidiropoulos, C. Faloutsos, and T. Mitchell. Efficient and Distributed Generalized Canonical Correlation Analysis for Big Multiview Data. *IEEE Transactions on Knowledge and Data Engineering*, 31(12):2304–2318, 2019.

[86] K. Fukumizu, F. Bach, and A. Gretton. Statistical Consistency of Kernel Canonical Correlation Analysis. *Journal of Machine Learning Research*, 8:361–383, 2007.

[87] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, 2nd edition, 1990. ISBN: 0-12-269851-7.

[88] X. Gao, S. Niu, and Q. Sun. Two-Directional Two-Dimensional Kernel Canonical Correlation Analysis. *IEEE Signal Processing Letters*, 26(11):1578–1582, 2019.

[89] G. M. L. Gladwell. On Isospectral Spring — Mass Systems. *Inverse Problems*, 11(3):591–602, 1995.

[90] G. H. Golub and C. F. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, Baltimore and London, 3 edition, 1996.

[91] 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.

[92] M. Gonen and E. Alpaydin. Multiple Kernel Learning Algorithms. *Journal of Machine Learning Research*, 12:2211–2268, 2011.

[93] I. Gonzalez, S. Dejean, P. G. P. Martin, and A. Baccini. CCA: An R Package to Extend Canonical Correlation Analysis. *Journal of Statistical Software*, 23(12):1–14, 2008.

[94] I. Gonzalez, S. Dejean, P. G. P. Martin, O. Goncalves, P. Besse, and A. Baccini. Highlighting Relationships Between Heterogeneous Biological Data Through Graphical Displays Based on Regularized Canonical Correlation Analysis. *Journal of Biological Systems*, 17(2):173–199, 2009.

[95] L. Grosenick, T. C. Shi, F. M. Gunning, M. J. Dubin, J. Downar, and C. Liston. Functional and Optogenetic Approaches to Discovering Stable Subtype-Specific Circuit Mechanisms in Depression. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 4(6):554–566, 2019.

[96] Y. Guo, X. Ding, C. Liu, and J. Xue. Sufficient Canonical Correlation Analysis. *IEEE Transactions on Image Processing*, 25(6):2610–2619, 2016.

[97] Y. Guo, T. Hastie, and R. Tibshirani. Regularized Linear Discriminant Analysis and Its Application in Microarrays. *Biostatistics*, 8(1):86–100, 2007.

[98] 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.

[99] M. Hanafi. PLS Path Modelling: Computation of Latent Variables with the Estimation Mode B. *Computational Statistics*, 22(2):275–292, 2007.

[100] 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.

[101] D. R. Hardoon and J. Shawe-Taylor. Convergence Analysis of Kernel Canonical Correlation Analysis: Theory and Practice. *Machine Learning*, 74(1):23–38, 2009.

[102] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical Correlation Analysis: An Overview with Application to Learning Methods. *Neural Computation*, 16(12):2639–2664, 2004.

[103] K. Hassani and A. H. Khasahmadi. Contrastive Multi-View Representation Learning on Graphs. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[104] X. He and P. Niyogi. Locality Preserving Projections. In *Advances in Neural Information Processing Systems*, volume 16, pages 585–591. MIT Press, 2003.

[105] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang. Face Recognition using Laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):328–340, 2005.

[106] G. E. Hinton, S. Osindero, and Y.-W. Teh. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7):1527–1554, 2006.

[107] A. E. Hoerl and R. W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67, 1970.

[108] 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.

[109] P. Horst. Generalized Canonical Correlations and Their Applications to Experimental Data. *Journal of Clinical Psychology*, 17(4):331–347, 1961.

[110] P. Horst. Relations Among M Sets of Measures. *Psychometrika*, 26(2):129–149, 1961.

[111] H. Hotelling. Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, 24(6):417–441, 1933.

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

[113] C. Hou, L.-L. Zeng, and D. Hu. Safe Classification with Augmented Features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2176–2192, 2019.

[114] C. Hou and Z.-H. Zhou. One-Pass Learning with Incremental and Decremental Features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2776–2792, 2018.

[115] C. Hovine and A. Bertrand. Distributed MAXVAR: Identifying Common Signal Components across the Nodes of a Sensor Network. In *Proceedings of the European Signal Processing Conference*, 2021.

[116] 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.

[117] R. Huang, S. Zhang, T. Li, and R. He. Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2458–2467, 2017.

[118] S. Y. Huang, M. H. Lee, and C. K. Hsiao. Nonlinear Measures of Association With Kernel Canonical Correlation Analysis and Applications. *Journal of Statistical Planning and Inference*, 139(7):2162–2174, 2009.

[119] X. Huang, B. Zhang, H. Qiao, and X. Nie. Local Discriminant Canonical Correlation Analysis for Supervised PolSAR Image Classification. *IEEE Geoscience and Remote Sensing Letters*, 14(11):2102–2106, 2017.

[120] 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.

[121] I. Huopaniemi, T. Suvitaival, J. Nikkila, M. Oresic, and S. Kaski. Multivariate Multi-Way Analysis of Multi-Source Data. *Bioinformatics*, 26(12):i391–i398, 2010.

[122] E. J. Ientilucci. Using the Singular Value Decomposition. Technical report, Chester F. Carlson Center for Imaging Science, Rochester Institute of Technology, 2003.

[123] Z. Ji, Y. Yu, Y. Pang, L. Chen, and Z. Zhang. Zero-Shot Learning with Multi-Battery Factor Analysis. *Signal Processing*, 138(C):265–272, 2017.

[124] Y. Jia, M. Salzmann, and T. Darrell. Factorized Latent Spaces with Structured Sparsity. In *Advances in Neural Information Processing Systems*, volume 23, pages 982–990, 2010.

[125] B. Jiang, C. Ding, B. Luo, and J. Tang. Graph-Laplacian PCA: Closed-Form Solution and Robustness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3492–3498, 2013.

[126] S. Jung and J. S. Marron. PCA Consistency in High Dimension, Low Sample Size Context. *The Annals of Statistics*, 37(6B):4104–4130, 2009.

[127] S. M. Kakade and D. P. Foster. Multi-view Regression Via Canonical Correlation Analysis. In *Learning Theory*, pages 82–96, 2007.

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

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

[130] 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.

[131] C. I. Kanatsoulis, X. Fu, N. D. Sidiropoulos, and M. Hong. Structured SUMCOR Multiview Canonical Correlation Analysis for Large-Scale Data. *IEEE Transactions on Signal Processing*, 67(2):306–319, 2019.

[132] G. Kang, K. Liu, B. Hou, and N. Zhang. 3D Multi-View Convolutional Neural Networks for Lung Nodule Classification. *Nature*, 12(11), 2017.

[133] M. Kang, S. Li, D. Kim, C. Liu, B. Zhang, X. Wu, and J. Gao. eQTL Mapping Study via Regularized Sparse Canonical Correlation Analysis. In *Proceedings of the 12th International Conference on Machine Learning and Applications*, pages 129–134, 2013.

[134] H. Kato, T. Nishimura, N. Ikeda, T. Yamada, T. Kondo, N. Saijo, K. Nishio, J. Fujimoto, M. Nomura, Y. Oda, B. Lindmark, J. Maniwa, H. Hibino, M. Unno, T. Ito, Y. Sawa, H. Tojo, S. Egawa, G. Edula, M. Lopez, M. Wigmore, N. Inase, Y. Yoshizawa, F. Nomura, and G. Marko-Varga. Developments for a Growing Japanese Patient Population: Facilitating New Technologies for Future Health Care. *Journal of Proteomics*, 74(6):759–764, 2011.

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

[136] A. Khan and P. Maji. Low-Rank Joint Subspace Construction for Cancer Subtype Discovery. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 17(4):1290–1302, 2020.

[137] A. Khan and P. Maji. Selective Update of Relevant Eigenspaces for Integrative Clustering of Multimodal Data. *IEEE Transactions on Cybernetics*, 52(2):947–959, 2022.

[138] 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.

[139] 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.

[140] T.-K. Kim, J. Kittler, and R. Cipolla. Discriminative Learning and Recognition of Image Set Classes Using Canonical Correlations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1005–1018, 2007.

[141] A. Klami and S. Kaski. Local Dependent Components. In *Proceedings of the 24th International Conference on Machine Learning*, pages 425–432, 2007.

[142] A. Klami and S. Kaski. Probabilistic Approach to Detecting Dependencies Between Data Sets. *Neurocomputing*, 72(1):39–46, 2008.

[143] A. Klami, S. Virtanen, and S. Kaski. Bayesian Canonical Correlation Analysis. *Journal of Machine Learning Research*, 14(1):965–1003, 2013.

[144] 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.

[145] M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien. Lp-Norm Multiple Kernel Learning. *Journal of Machine Learning Research*, 12:953–997, 2011.

[146] 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.

[147] 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.

[148] A. Kumar, P. Rai, and H. Daume III. Co-regularized Spectral Clustering with Multiple Kernels. 2010.

[149] A. Kumar, P. Rai, and H. Daume III. Co-regularized Multi-view Spectral Clustering. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

[150] G. R. Lanckriet, M. Deng, N. Cristianini, M. I. Jordan, and W. S. Noble. Kernel-based Data Fusion and Its Application to Protein Function Prediction in Yeast. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 300–311, 2004.

[151] G. R. G. Lanckriet, T. D. Bie, N. Cristianini, M. I. Jordan, and W. S. Noble. A Statistical Framework for Genomic Data Fusion. *Bioinformatics*, 20(16):2626–2635, 2004.

[152] 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.

[153] Y. LeCun, Y. Bengio, and G. E. Hinton. Deep Learning. *Nature*, 521(7553):436–444, 2015.

[154] G. Lee, S. Doyle, J. Monaco, A. Madabhushi, M. D. Feldman, S. R. Master, and J. E. Tomaszewski. A Knowledge Representation Framework for Integration, Classification of Multi-scale Imaging and Non-imaging Data: Preliminary Results in Predicting Prostate Cancer Recurrence by Fusing Mass Spectrometry and Histology. In *Proceedings of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 77–80, 2009.

[155] G. Lee, A. Singanamalli, H. Wang, M. D. Feldman, S. R. Master, N. N. C. Shih, E. Spangler, T. Rebbeck, J. E. Tomaszewski, and A. Madabhushi. 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, 2015.

[156] S. E. Leurgans, R. A. Moyeed, and B. W. Silverman. Canonical Correlation Analysis when the Data are Curves. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(3):725–740, 1993.

[157] 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.

[158] 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.

[159] D. Lin, V. D. Calhoun, and Y.-P. Wang. Correspondence Between fMRI and SNP Data by Group Sparse Canonical Correlation Analysis. *Medical Image Analysis*, 18(6):891–902, 2014.

[160] D. Lin, J. Zhang, J. Li, V. D. Calhoun, H. W. Deng, and Y. P. Wang. Group Sparse Canonical Correlation Analysis for Genomic Data Integration. *BMC Bioinformatics*, 14(245), 2013.

[161] W. Lin, D. Lv, Z. Han, J. Dong, and L. Yang. Major Depressive Disorder Identification by Referenced Multiset Canonical Correlation Analysis with Clinical Scores. *Medical Image Analysis*, 60:101600, 2020.

[162] G. Lisanti, S. Karaman, and I. Masi. Multichannel-Kernel Canonical Correlation Analysis for Cross-View Person Reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 13(2), 2017.

[163] G. Lisanti, I. Masi, and A. Del Bimbo. Matching People across Camera Views Using Kernel Canonical Correlation Analysis. In *Proceedings of the International Conference on Distributed Smart Cameras*, 2014.

[164] Q. Liu, Y. Jiao, Y. Miao, C. Zuo, X. Wang, A. Cichocki, and J. Jin. Efficient Representations of EEG Signals for SSVEP Frequency Recognition Based on Deep Multiset CCA. *Neurocomputing*, 378:36–44, 2020.

[165] X. Liu, L. Jiao, L. Li, L. Cheng, F. Liu, S. Yang, and B. Hou. Deep Multiview Union Learning Network for Multisource Image Classification. *IEEE Transactions on Cybernetics*, pages 1–13, 2020.

[166] X. Liu, L. Wang, J. Yin, E. Zhu, and J. Zhang. An Efficient Approach to Integrating Radius Information into Multiple Kernel Learning. *IEEE Transactions on Cybernetics*, 43(2):557–569, 2013.

[167] 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.

[168] A. Lu, W. Wang, M. Bansal, K. Gimpel, and K. Livescu. Deep Multilingual Correlation for Improved Word Embeddings. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 250–256, 2015.

[169] 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.

[170] P. Maji. f-Information Measures for Efficient Selection of Discriminative Genes from Microarray Data. *IEEE Transactions on Biomedical Engineering*, 56(4):1063–1069, 2009.

[171] P. Maji. Fuzzy-Rough Supervised Attribute Clustering Algorithm and Classification of Microarray Data. *IEEE Transactions on System, Man and Cybernetics, Part B: Cybernetics*, 41(1):222–233, 2011.

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

[173] P. Maji and A. Mandal. Rough Hypercuboid Based Supervised Regularized Canonical Correlation for Multimodal Data Analysis. *Fundamenta Informaticae*, 148(1-2):133–155, 2016.

[174] 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.

[175] P. Maji and S. K. Pal. Fuzzy-Rough Sets for Information Measures and Selection of Relevant Genes from Microarray Data. *IEEE Transactions on System, Man, and Cybernetics, Part B: Cybernetics*, 40(3):741–752, 2010.

[176] P. Maji and S. K. Pal. *Rough-Fuzzy Pattern Recognition: Applications in Bioinformatics and Medical Imaging.* Wiley-IEEE Computer Society Press, Hoboken, New Jersey, 2012.

[177] P. Maji and S. Paul. Rough Set Based Maximum Relevance-Maximum Significance Criterion and Gene Selection from Microarray Data. *International Journal of Approximate Reasoning*, 52(3):408–426, 2011.

[178] P. Maji and S. Paul. Robust Rough-Fuzzy C-Means Algorithm: Design and Applications in Coding and Non-coding RNA Expression Data Clustering. *Fundamenta Informaticae*, 124(1-2):153–174, 2013.

[179] P. Maji and S. Paul. Rough-Fuzzy Clustering for Grouping Functionally Similar Genes from Microarray Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10(2):286–299, 2013.

[180] P. Maji and S. Paul. *Scalable Pattern Recognition Algorithms: Applications in Computational Biology and Bioinformatics.* Springer-Verlag, London, 2014. ISBN: 978-3-319-05629-6.

[181] A. Mandal and P. Maji. A New Method to Address Singularity Problem in Multimodal Data Analysis. In *Proceedings of 7th International Conference on Pattern Recognition and Machine Intelligence*, pages 43–51, 2017.

[182] A. Mandal and P. Maji. Regularization and Shrinkage in Rough Set Based Canonical Correlation Analysis. In *Proceedings of International Joint Conference on Rough Sets*, pages 432–446, 2017.

[183] 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.

[184] A. Mandal and P. Maji. Adaptive Generalized Multi-View Canonical Correlation Analysis for Incrementally Update Multiblock Data. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–14, 2022.

[185] A. Mandal and P. Maji. Multi-View Regularized Discriminant Canonical Correlation Analysis: Sequential Extraction of Relevant Features from Multiblock Data. *IEEE Transactions on Cybernetics*, pages 1–13, 2022.

[186] 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.

[187] A. Markos and A. I. D'Enza. Incremental Generalized Canonical Correlation Analysis. In A. F. X. Wilhelm and H. A. Kestler, editors, *Analysis of Large and Complex Data*, pages 185–194. Springer International Publishing, 2016.

[188] S. Mehrkanoon and J. A. K. Suykens. Regularized Semipaired Kernel CCA for Domain Adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 29(7):3199–3213, 2018.

[189] R. Memisevic, L. Sigal, and D. J. Fleet. Shared Kernel Information Embedding for Discriminative Inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):778–790, 2012.

[190] C. A. G. Van. Mieghem, N. Bruining, J. A. Schaar, E. Mcfadden, N. Mollet, F. Cademartiri, F. Mastik, J. M. R. Ligthart, G. A. R. Granillo, M. Valgimigli, G. Sianos, W. J. van der. Giessen, B. Backx, M. M. Morel, G. Es, J. D. Sawyer, J. Kaplow, A. Zalewski, A. F. W. van der Steen, P. J. de Feyter, and P. W. Serruys. Rationale and Methods of the Integrated Biomarker and Imaging Study (IBIS): Combining Invasive and Non-invasive Imaging with Biomarkers to Detect Subclinical Atherosclerosis and Assess Coronary Lesion Biology. *The International Journal of Cardiovascular Imaging*, 21(4):425–441, 2005.

[191] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, volume 2, pages 3111–3119, 2013.

[192] K. S. Miller. On the Inverse of the Sum of Matrices. *Mathematics Magazine*, 54(2):67–72, 1981.

[193] Q. Mo and R. Shen. iClusterPlus: Integrative Clustering of Multiple Genomic Data Sets. *R package version 1.19.0*, 2018.

[194] 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.

[195] 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.

[196] V. N. Murthy, S. Maji, and R. Manmatha. Automatic Image Annotation Using Deep Learning Representations. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 603–606, 2015.

[197] 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.

[198] 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.

[199] I. Muslea, S. Minton, and C. A. Knoblock. Active Learning with Multiple Views. *Journal of Artificial Intelligence Research*, 27(1):203–233, 2006.

[200] A. Nazarpour and P. Adibi. Two-Stage Multiple Kernel Learning for Supervised Dimensionality Reduction. *Pattern Recognition*, 48(5):1854–1862, 2015.

[201] TCGA Research Network. Integrated Genomic Analyses of Ovarian Carcinoma. *Nature*, 474(7353):609–615, 2011.

[202] TCGA Research Network. Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *The New England Journal of Medicine*, 372(26):2481–2498, 2015.

[203] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal Deep Learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, page 689–696, 2011.

[204] M. A. Nicolaou, V. Pavlovic, and M. Pantic. Dynamic Probabilistic CCA for Analysis of Affective Behavior and Fusion of Continuous Annotations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1299–1311, 2014.

[205] A. A. Nielsen. Multiset Canonical Correlations Analysis and Multispectral, Truly Multitemporal Remote Sensing Data. *IEEE Transactions on Image Processing*, 11(3):293–305, 2002.

[206] 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.

[207] S. V. Parikh and J. A. De Lemos. Biomarkers in Cardiovascular Disease: Integrating Pathophysiology into Clinical Practice. *The American Journal of the Medical Sciences*, 332(4):186–197, 2006.

[208] E. Parkhomenko, D. Tritchler, and J. Beyene. Genome-wide Sparse Canonical Correlation of Gene Expression With Genotypes. *BMC Proceedings*, 1(1), 2007.

[209] E. Parkhomenko, D. Tritchler, and J. Beyene. Sparse Canonical Correlation Analysis With Application to Genomic Data Integration. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–34, 2009.

[210] S. Paul and P. Maji. μHEM for Identification of Differentially Expressed miRNAs Using Hypercuboid Equivalence Partition Matrix. *BMC Bioinformatics*, 14(1):266, 2013.

[211] S. Paul and P. Maji. Rough Sets for Insilico Identification of Differentially Expressed miRNAs. *International Journal of Nanomedicine*, 8:63–74, 2013.

[212] S. Paul and P. Maji. City Block Distance and Rough-Fuzzy Clustering for Identification of Co-Expressed microRNAs. *Molecular BioSystems*, 10(6):1509–1523, 2014.

[213] S. Paul and P. Maji. Gene Expression and Protein–Protein Interaction Data for Identification of Colon Cancer Related Genes using f-information Measures. *Natural Computing*, 15(3):449–463, 2016.

[214] Z. Pawlak. *Rough Sets: Theoretical Aspects of Reasoning about Data.* Kluwer Academic Publishers, Dordrecht, 1991.

[215] Y. Peng, D. Zhang, and J. Zhang. A New Canonical Correlation Analysis Algorithm With Local Discrimination. *Neural Processing Letters*, 31(1):1–15, 2010.

[216] B. A. Pour, S. M. Shams, and S. Strother. A Hybrid LDA+gCCA Model for fMRI Data Classification and Visualization. *IEEE Transactions on Medical Imaging*, 34(5):1031–1041, 2015.

[217] S. R. Prasad, P. A. Humphrey, J. R. Catena, V. R. Narra, J. R. Srigley, A. D. Cortez, N. C. Dalrymple, and K. N. Chintapalli. Common and Uncommon Histologic Subtypes of Renal Cell Carcinoma: Imaging Spectrum with Pathologic Correlation. *Radiographics*, 26(6):1795–1806, 2006.

[218] 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.

[219] N. Quadrianto and C. H. Lampert. Learning Multi-view Neighborhood Preserving Projections. In *Proceedings of the International Conference on Machine Learning*, pages 425–432, 2011.

[220] N. Rappoport and R. Shamir. Multi-Omic and Multi-View Mlustering Algorithms: Review and Cancer Benchmark. *Nucleic Acids Research*, 46(20):10546–10562, 2018.

[221] 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.

[222] L. J. Revell and A. S. Harrison. PCCA: A Program for Phylogenetic Canonical Correlation Analysis. *Bioinformatics*, 24(7):1018–1020, 2008.

[223] S. A. Rifkin and J. Kim. Geometry of Gene Expression Dynamics. *Bioinformatics*, 18(9):1176–1183, 2002.

[224] S. Rogers, A. Klami, J. Sinkkonen, M. Girolami, and S. Kaski. Infinite Factorization of Multiple Non-Parametric Views. *Machine Learning*, 79(1):201–226, 2009.

[225] 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.

[226] R. Rosipal and L. J. Trejo. Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space. *Journal of Machine Learning Research*, 2:97–123, 2002.

[227] S. T. Roweis and L. K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326, 2000.

[228] J. Rupnik and J. Shawe-Taylor. Multi-View Canonical Correlation Analysis. 2010.

[229] J. Rupnik, P. Skraba, J. Shawe-Taylor, and S. Guettes. A Comparison of Relaxations of Multiset Cannonical Correlation Analysis and Applications, 2013.

[230] M. Salzmann, C. H. Ek, R. Urtasun, and T. Darrell. Factorized Orthogonal Latent Spaces. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9, pages 701–708, 2010.

[231] 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.

[232] Y. Sasaki. The Truth of the F-measure. https://www.toyota-ti.ac.jp/Lab/Denshi/COIN/people/yutaka.sasaki/F-measure-YS-26Oct07.pdf, October 2007.

[233] R. S. Savage, Z. Ghahramani, J. E. Griffin, B. J. D. L. Cruz, and D. L. Wild. Discovering Transcriptional Modules by Bayesian Data Integration. *Bioinformatics*, 26(12):i158–i167, 2010.

[234] L. F. Schoenfeldt and R. Cudeck. MBFACT: Multiple Battery Factor Analysis by Maximum Likelihood. *Applied Psychological Measurement*, 4(3):417–418, 1980.

[235] A.-K. Seghouane and A. Iqbal. The Adaptive Block Sparse PCA and Its Application to Multi-subject FMRI Data Analysis Using Sparse mCCA. *Signal Processing*, 153:311–320, 2018.

[236] H. S. Seung and D. D. Lee. Cognition. The Manifold Ways of Perception. *Science*, 290(5500):2268–2269, 2000.

[237] F. Shang, L. C. Jiao, and F. Wang. Graph Dual Regularization Non-negative Matrix Factorization for Co-clustering. *Pattern Recognition*, 45(6):2237–2250, 2012.

[238] J. Shao, L. Wang, Z. Zhao, F. su, and A. Cai. Deep Canonical Correlation Analysis with Progressive and Hypergraph Learning for Cross-Modal Retrieval. *Neurocomputing*, 214:618–628, 2016.

[239] X. Shen, Q. Sun, and Y. Yuan. A Unified Multiset Canonical Correlation Analysis Framework Based on Graph Embedding for Multiple Feature Extraction. *Neurocomputing*, 148:397–408, 2015.

[240] 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.

[241] Y. Shin and C. Park. Analysis of Correlation Based Dimension Reduction Methods. *International Journal of Applied Mathematics and Computer Science*, 21(3):549–558, 2011.

[242] A. Shon, K. Grochow, A. Hertzmann, and R. Rao. Learning Shared Latent Structure for Image Synthesis and Robotic Imitation. In *Advances in Neural Information Processing Systems*, volume 18, 2005.

[243] 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.

[244] K. Somandepalli, N. Kumar, R. Travadi, and S. Narayanan. Multimodal Representation Learning Using Deep Multiset Canonical Correlation. *arXiv preprint arXiv:1904.01775*, 2019.

[245] S. Sonnenburg, G. Ratsch, C. Schafe, and B. Scholkopf. Large Scale Multiple Kernel Learning. *Journal of Machine Learning Research*, 7(57):1531–1565, 2006.

[246] 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.

[247] N. Srivastava and R. Salakhutdinov. Multimodal Learning with Deep Boltzmann Machines. *Journal of Machine Learninig Research*, 15(1):2949–2980, 2014.

[248] G. Strang. *Introduction to Linear Algebra.* Wellesley, Cambridge Press, 2016.

[249] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller. Multi-view Convolutional Neural Networks for 3D Shape Recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 945–953, 2015.

[250] M. Sugiyama. Dimensionality Reduction of Multimodal Labeled Data by Local Fisher Discriminant Analysis. *Journal of Machine Learning Research*, 8(37):1027–1061, 2007.

[251] S. Sun, L. Mao, Z. Dong, and L. Wu. *Multiview Machine Learning.* Springer Singapore, Singapore, 2019. ISBN: 978-981-13-3029-2.

[252] S. Sun, X. Xie, and M. Yang. Multiview Uncorrelated Discriminant Analysis. *IEEE Transactions on Cybernetics*, 46(12):3272–3284, 2016.

[253] S. Sun, X. Xie, and M. Yang. Multiview Uncorrelated Discriminant Analysis. *IEEE Transactions on Cybernetics*, 46(12):3272–3284, 2016.

[254] 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.

[255] T. Sun, S. Chen, J. Yang, X. Hu, and P. Shi. Discriminative Canonical Correlation Analysis with Missing Samples. In *Proceedings of the WRI World Congress on Computer Science and Information Engineering*, volume 6, pages 95–99, 2009.

[256] T. Sun, S. Chen, J. Yang, and P. Shi. A Novel Method of Combined Feature Extraction for Recognition. In *Proceedings of the Eighth IEEE International Conference on Data Mining*, pages 1043–1048, 2008.

[257] X. Suo, V. Minden, B. Nelson, R. Tibshirani, and M. Saunders. Sparse Canonical Correlation Analysis. *arXiv*, 2017.

[258] I. Sutskever, J. Martens, and G. E. Hinton. Generating Text with Recurrent Neural Networks. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1017–1024, 2011.

[259] A. F. Syafiandini, I. Wasito, S. Yazid, A. Fitriawan, and M. Amien. Multimodal Deep Boltzmann Machines for Feature Selection on Gene Expression Data. In *Proceedings of the International Conference on Advanced Computer Science and Information Systems*, pages 407–412, 2016.

[260] L. Tang, Z. Yang, and K. Jia. Canonical Correlation Analysis Regularization: An Effective Deep Multiview Learning Baseline for RGB-D Object Recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 11(1):107–118, 2019.

[261] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, 2000.

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

[263] 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.

[264] M. Tenenhaus, V. E. Vinzi, Y.-M. Chatelin, and C. Lauro. PLS Path Modeling. *Computational Statistics and Data Analysis*, 48(1):159–205, 2005.

[265] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, Inc., USA, 4th edition, 2008. ISBN: 9781597492720.

[266] Q. Tian, C. Ma, M. Cao, S. Chen, and H. Yin. A Convex Discriminant Semantic Correlation Analysis for Cross-View Recognition. *IEEE Transactions on Cybernetics*, 52(2):849–861, 2022.

[267] 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.

[268] 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.

[269] N. H. Timm. *Applied Multivariate Analysis*. Springer, New York, 2002.

[270] P. Tiwari, S. Viswanath, G. Lee, and A. Madabhushi. Multi-modal Data Fusion Schemes for Integrated Classification of Imaging and Non-imaging Biomedical Data. In *Proceedings of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 165–168, 2011.

[271] L. Tran, X. Yin, and X. Liu. Disentangled Representation Learning GAN for Pose-Invariant Face Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1283–1292, 2017.

[272] W. D. Travis, E. Brambilla, A. P. Burke, A. Marx, and A. G. Nicholson. Introduction to The 2015 World Health Organization Classification of Tumors of the Lung, Pleura, Thymus, and Heart. *Journal of Thoracic Oncology*, 10(9):1240–1242, 2015.

[273] L. R. Tucker. An Inter-Battery Method of Factor Analysis. *Psychometrika*, 23:111–136, 1958.

[274] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, 1995.

[275] 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.

[276] M. V. D. Velden and T. H. A. Bijmolt. Generalized Canonical Correlation Analysis of Matrices with Missing Rows: A Simulation Study. *Psychometrika*, 71(2):323–331, 2006.

[277] J. Via, I. Santamaria, and J. Perez. A Learning Algorithm for Adaptive Canonical Correlation Analysis of Several Data Sets. *Neural Networks*, 20(1):139–152, 2007.

[278] H. D. Vinod. Canonical Ridge and Econometrics of Joint Production. *Journal of Econometrics*, 4(2):147–166, 1976.

[279] S. Virtanen, A. Klami, and S. Kaski. Bayesian CCA via Group Sparsity. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 457–464, 2011.

[280] 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.

[281] S. Virtanen, A. Klami, S. Khan, and S. Kaski. Bayesian Group Factor Analysis. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22, pages 1269–1277, 2012.

[282] S. Viswanath and A. Madabhushi. Consensus Embedding: Theory, Algorithms and Application to Segmentation and Classification of Biomedical Data. *BMC Bioinformatics*, 13(26), 2012.

[283] C. Wang. Variational Bayesian Approach to Canonical Correlation Analysis. *IEEE Transactions on Neural Networks*, 18(3):905–910, 2007.

[284] 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.

[285] 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.

[286] T. Wang, D. Zhao, and Y. Feng. Two-Stage Multiple Kernel Learning with Multiclass Kernel Polarization. *Knowledge-Based Systems*, 48:10–16, 2013.

[287] W. Wang, R. Arora, K. Livescu, and J. A. Bilmes. On Deep Multi-View Representation Learning. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, volume 37, pages 1083–1092, 2015.

[288] W. Wang, R. Arora, K. Livescu, and J. A. Bilmes. Unsupervised Learning of Acoustic Features Via Deep Canonical Correlation Analysis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4590–4594, 2015.

[289] W. Wang, R. Arora, K. Livescu, and N. Srebro. Stochastic Optimization for Deep CCA via Nonlinear Orthogonal Iterations. In *Proceedings of the 53rd Annual Allerton Conference on Communication, Control, and Computing*, pages 688–695, 2015.

[290] 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.

[291] W. Wang and Z.H. Zhou. Analyzing co-training style algorithms. In *Machine Learning: ECML 2007*, pages 454–465, 2007.

[292] 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.

[293] P. A. White. The Computation of Eigenvalues and Eigenvectors of a Matrix. *Journal of the Society for Industrial and Applied Mathematics*, 6(4):393–437, 1958.

[294] D. M. Witten, R. Tibshirani, and T. Hastie. A Penalized Matrix Decomposition, With Applications to Sparse Principal Components and Canonical Correlation Analysis. *Biostatistics*, 10(3):515–534, 2009.

[295] D. M. Witten and R. J. Tibshirani. Extensions of Sparse Canonical Correlation Analysis With Applications to Genomic Data. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–27, 2009.

[296] H. Wold. Estimation of Principal Components and Related Models by Iterative Least Squares. *Journal of Multivariate Analysis*, pages 391–420, 1966.

[297] 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.

[298] 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.

[299] C. Xu, D. Tao, and C. Xu. Multi-View Learning With Incomplete Views. *IEEE Transactions on Image Processing*, 24(12):5812–5825, 2015.

[300] M. Xu, Z. Zhu, X. Zhang, Y. Zhao, and X. Li. Canonical Correlation Analysis With $L_{2,1}$-Norm for Multiview Data Representation. *IEEE Transactions on Cybernetics*, 50(11):4772–4782, 2020.

[301] 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.

[302] 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.

[303] Y. Yamanishi, J. P. Vert, and M. Kanehisa. Protein Network Inference from Multiple Genomic Data: A Supervised Approach. *Bioinformatics*, 20:i363–i370, 2004.

[304] F. Yan and K. Mikolajczyk. Deep correlation for matching images and text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3441–3450, 2015.

[305] J. Yang and X. Zhang. Feature-Level Fusion of Fingerprint and Finger-Vein for Personal Identification. *Pattern Recognition Letters*, 33(5):623–628, 2012.

[306] X. Yang and W. Liu. Multiple Scale Canonical Correlation Analysis Networks for Two-View Object Recognition. In *Neural Information Processing*, pages 325–334, 2017.

[307] X. Yang, W. Liu, D. Tao, and J. Cheng. Canonical Correlation Analysis Networks for Two-View Image Recognition. *Information Sciences*, 385-386:338–352, 2017.

[308] 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.

[309] Z. Yang, X. Zhuang, K. Sreenivasan, V. Mishra, T. Curran, R. Byrd, R. Nandy, and D. Cordes. 3D Spatially-adaptive Canonical Correlation Analysis: Local and Global Methods. *NeuroImage*, 169:240–255, 2018.

[310] F. Yger, M. Berar, G. Gasso, and A. Rakotomamonjy. Adaptive Canonical Correlation Analysis Based on Matrix Manifolds. In *Proceedings of the 29th International Conference on Machine Learning, Edinburgh, Scotland, UK*, 2012.

[311] S. Yu, B. Krishnapuram, R. Rosales, and R. B. Rao. Bayesian Co-training. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

[312] S. Yu, B. Krishnapuram, R. Rosales, and R. B. Rao. Bayesian Co-training. *The Journal of Machine Learning Research*, pages 2649–2680, 2011.

[313] D. Zhai, H. Chang, S. Shan, X. Chen, and W. Gao. Multiview Metric Learning with Global Consistency and Local Smoothness. *ACM Transactions on Intelligent Systems and Technology*, 3(3), 2012.

[314] L. Zhang. Riemannian Newton Method for the Multivariate Eigenvalue Problem. *SIAM Journal on Matrix Analysis and Applications*, 31(5):2972–2996, 2010.

[315] 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.

[316] Z. Zhang, M. Zhao, and T. W. S. Chow. Binary- and Multi-class Group Sparse Canonical Correlation Analysis for Feature Extraction and Classification. *IEEE Transactions on Knowledge and Data Engineering*, 25(10):2192–2205, 2013.

[317] 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.

[318] 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.

[319] H. Zhao, D. Sun, and Z. Luo. Incremental Canonical Correlation Analysis. *Applied Sciences*, 10(21), 2020.

[320] J. Zhao, X. Xie, X. Xu, and S. Sun. Multi-View Learning Overview: Recent Progress and New Challenges. *Information Fusion*, 38:43–54, 2017.

[321] H. Zheng, H. Wang, and D. H. Glass. Integration of Genomic Data for Inferring Protein Complexes from Global Protein-Protein Interaction Networks. *IEEE Transactions on Systems, Man, and Cybernetics, Part B, Cybernetics*, 38(1):5–16, 2008.

[322] W. Zheng. Multichannel EEG-Based Emotion Recognition via Group Sparse Canonical Correlation Analysis. *IEEE Transactions on Cognitive and Developmental Systems*, 9(3):281–290, 2017.

[323] 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.

[324] Y. Zhou, N. Hu, and C. J. Spanos. Veto-Consensus Multiple Kernel Learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2407–2414, 2016.

[325] Y. Zhou, H. Lu, and Y. M. Cheung. Bilinear Probabilistic Canonical Correlation Analysis via Hybrid Concatenations. In *Proceedings of 31st AAAI Conference on Artificial Intelligence*, pages 2949–2955, 2017.

[326] X. Zhu, Z. Huang, H. Tao Shen, J. Cheng, and C. Xu. Dimensionality Reduction by Mixed Kernel Canonical Correlation Analysis. *Pattern Recognition*, 45(8):3003–3016, 2012.

[327] X. Zhuang, Z. Yang, T. Curran, R. Byrd, R. Nandy, and D. Cordes. A Family of Locally Constrained CCA Models for Detecting Activation Patterns in fMRI. *NeuroImage*, 149:63–84, 2017.

[328] X. Zhuang, Z. Yang, K. Sreenivasan, V. Mishra, T. Curran, R. Nandy, and D. Cordes. Multivariate Group-level Analysis For Task fMRI Data With Canonical Correlation Analysis. *NeuroImage*, 194:25–41, 2019.

[329] G. Ziegler, R. Dahnke, A. D. Winkler, and C. Gaser. Partial Least Squares Correlation of Multivariate Cognitive Abilities and Local Brain Structure in Children and Adolescents. *NeuroImage*, 82:284–294, 2013.

[330] W. Zuobin, M. Kezhi, and G.-W. Ng. Effective Feature Fusion for Pattern Classification Based on Intra-Class and Extra-Class Discriminative Correlation Analysis. In *Proceedings of the 20th International Conference on Information Fusion*, pages 1–8, 2017.