Multimodal Data Fusion and Quantitative Analysis for Medical Applications

BOWEN XIN

Doctor of Philosophy (Computer Science)

Supervisor: Xiuying Wang
Associate Supervisor: David Feng

A thesis submitted in fulfilment of
the requirements for the degree of
Doctor of Philosophy

School of Computer Science
The University of Sydney
Australia

26 October 2021

Declaration

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Abstract

Medical big data is not only enormous in its size, but also heterogeneous and complex in its data structure, which makes conventional systems or algorithms difficult to process. These heterogeneous medical data include imaging data (e.g., Positron Emission Tomography (PET), Computerized Tomography (CT), Magnetic Resonance Imaging (MRI)), and non-imaging data (e.g., laboratory biomarkers, electronic medical records, and hand-written doctor notes). Multimodal data fusion is an emerging vital field to address this urgent challenge, aiming to process and analyze the complex, diverse and heterogeneous multimodal data. The fusion algorithms bring great potential in medical data analysis, by 1) taking advantage of complementary information from different sources (such as functional-structural complementarity of PET/CT images) and 2) exploiting consensus information that reflects the intrinsic essence (such as the genetic essence underlying medical imaging and clinical symptoms). Thus, multimodal data fusion benefits a wide range of quantitative medical applications, including personalized patient care, more optimal medical operation plan, and preventive public health.

Though there has been extensive research on computational approaches for multimodal fusion, there are three major challenges of multimodal data fusion in quantitative medical applications, which are summarized as feature-level fusion, information-level fusion and knowledge-level fusion:

- **Feature-level fusion.** The first challenge is to mine multimodal biomarkers from high-dimensional small-sample multimodal medical datasets, which hinders the effective discovery of informative multimodal biomarkers. Specifically, efficient dimension reduction algorithms are required to alleviate "curse of dimensionality" problem and address the criteria for discovering interpretable, relevant, non-redundant and generalizable multimodal biomarkers.

- **Information-level fusion.** The second challenge is to exploit and interpret inter-modal and intra-modal information for precise clinical decisions. Although radiomics and multi-branch deep learning have been used for implicit information fusion guided with supervision of the labels, there is a lack of methods to explicitly explore inter-modal relationships in medical applications. Unsupervised multimodal learning is able to mine inter-modal relationship as well as reduce the usage of labor-intensive data and explore potential undiscovered biomarkers; however, mining discriminative information without label supervision is an upcoming challenge. Furthermore, the interpretation of complex non-linear cross-modal associations, especially in deep multimodal learning, is another critical challenge in information-level fusion, which hinders the exploration of multimodal interaction in disease mechanism.
- **Knowledge-level fusion.** The third challenge is quantitative knowledge distillation from multi-focus regions on medical imaging. Although characterizing imaging features from single lesions using either feature engineering or deep learning methods have been investigated in recent years, both methods neglect the importance of inter-region spatial relationships. Thus, a topological profiling tool for multi-focus regions is in high demand, which is yet missing in current feature engineering and deep learning methods. Furthermore, incorporating domain knowledge with distilled knowledge from multi-focus regions is another challenge in knowledge-level fusion.

To address the three challenges in multimodal data fusion, this thesis provides a multi-level fusion framework for multimodal biomarker mining, multimodal deep learning, and knowledge distillation from multi-focus regions. Specifically, our major contributions in this thesis include:

- To address the challenges in **feature-level fusion**, we propose an Integrative Multimodal Biomarker Mining framework to select interpretable, relevant, non-redundant and generalizable multimodal biomarkers from high-dimensional small-sample imaging and non-imaging data for diagnostic and prognostic applications. The feature selection criteria including representativeness, robustness, discriminability, and non-redundancy are exploited by consensus clustering, Wilcoxon filter, sequential

forward selection, and correlation analysis, respectively. SHapley Additive exPlanations (SHAP) method and nomogram are employed to further enhance feature interpretability in machine learning models.

- To address the challenges in **information-level fusion**, we propose an Interpretable Deep Correlational Fusion framework, based on canonical correlation analysis (CCA) for 1) cohesive multimodal fusion of medical imaging and non-imaging data, and 2) interpretation of complex non-linear cross-modal associations. Specifically, two novel loss functions are proposed to optimize the discovery of informative multimodal representations in both supervised and unsupervised deep learning, by jointly learning inter-modal consensus and intra-modal discriminative information. An interpretation module is proposed to decipher the complex non-linear cross-modal association by leveraging interpretation methods in both deep learning and multimodal consensus learning.
- To address the challenges in **knowledge-level fusion**, we proposed a Dynamic Topological Analysis framework, based on persistent homology, for knowledge distillation from inter-connected multi-focus regions in medical imaging and incorporation of domain knowledge. Different from conventional feature engineering and deep learning, our DTA framework is able to explicitly quantify inter-region topological relationships, including global-level geometric structure and community-level clusters. K-simplex Community Graph is proposed to construct the dynamic community graph for representing community-level multi-scale graph structure. The constructed dynamic graph is subsequently tracked with a novel Decomposed Persistence algorithm. Domain knowledge is incorporated into the Adaptive Community Profile, summarizing the tracked multi-scale community topology with additional customizable clinically important factors.

Publications

Journal Papers

(1) **Xin B**, Huang J, Zhang L, Zheng C, Zhou Y, Lu J, Wang X, ’Dynamic Topology Analysis for Spatial Patterns of Multifocal Lesions on MRI’, Medical Image Analysis, 2021. (Accepted)

(2) Huang J, **Xin B**, Wang X, Qi Z, Dong H, Li K, Zhou Y, Jie L, ’Multi-parametric MRI Phenotype with Trustworthy Machine Learning for Differentiating CNS Demyelinating Diseases’, Journal of Translational Medicine, 2021 Dec 19(1):1-2. (Co-first author)

(3) Lv L, **Xin B**, Hao Y, Yang Z, Xu J, Wang L, Wang X, Song S, Guo X, ’Radiomic Analysis for Predicting Prognosis of Colorectal Cancer from Preoperative 18F-FDG PET/CT’, Journal of Translational Medicine, 2021. (Co-first author, under review)

(4) Wang L, Dong T, **Xin B**, Xu C, Guo M, Zhang H, Feng D, Wang X, Yu J ’Integrative Nomogram of CT Imaging, Clinical, and Hematological Features for Survival Prediction of Patients with Locally Advanced Non-small Cell Lung Cancer’, European Radiology, 2019 Jan 14:1-0.

(5) Li H, Xu C, **Xin B**, Zheng C, Zhao Y, Hao K, Wang Q, Wahl RL, Wang X, Zhou Y, ’18F-FDG PET/CT radiomic analysis with machine learning for identifying bone marrow involvement in the patients with suspected relapsed acute leukemia’, Theranostics. 2019 Sep 16:4730

(6) Li J, Yang Z, **Xin B**, Hao Y, Wang L, Song S, Xu J, Wang X, ’Quantitative Prediction of Microsatellite Instability in Colorectal Cancer with Preoperative PET/CT-Based Radiomics’, Frontiers of Oncology, 2021:2790.

(7) Liu Q, Li J, **Xin B**, Sun Y, Fulham M, Wang X, Song S, '18F-FDG PET/CT Radiomics for Preoperative Prediction of Lymph Node Metastases and Nodal Staging in Gastric Cancer', Frontiers of Oncology, 2021:3540.

(8) Jiang C, Zhao L, **Xin B**, Ma G, Wang X, Song S, '18F-FDG PET/CT Radiomic Analysis for Classifying Hepatocellular Carcinoma and Cholangiocarcinoma and Predicting Microvascular Invasion in Primary Liver Cancer', Frontiers of Oncology, 2021. (Under review)

Conference Papers

(9) **Xin B**, Zeng S, Wang X, 'Self-supervised Deep Correlational Multi-view Clustering', In International Joint Conference of Neural Network, pp. 1-8, 2021. (Oral Presentation)

(10) **Xin B**, Huang J, Zhou Y, Lu J, Wang X, 'Interpretation on Deep Multimodal Fusion for Diagnostic Classification', In International Joint Conference of Neural Network, pp 1-8, 2021. (Oral Presentation)

(11) **Xin B**, Zhang L, Huang J, Lu J, Wang X, 'Multi-level Topological Analysis Framework for Multifocal Diseases', In IEEE International Conference on Control, Automation, Robotics and Vision, pp 666-671, 2020. (Oral Presentation)

(12) **Xin B**, Xu C, Wang L, Dong T, Zheng C, Wang X, 'Integrative Clustering and Supervised Feature Selection for Clinical Applications', In IEEE International Conference on Control, Automation, Robotics and Vision, pp 1316-1320, 2018. (Oral Presentation)

(13) Li J, **Xin B**, Yang Z, Xu J, Song S, Wang X, 'Harmonization Centered Ensemble for Small and Highly Imbalanced Medical Data Classification', In IEEE International Symposium on Biomedical Imaging (ISBI), pp 1742-1745, 2021.

(14) Xue Z, **Xin B**, Wang D, Wang X, 'Radiomics-Enhanced Multi-task Neural Network for Non-invasive Glioma Subtyping and Segmentation', In RNO-AI@MICCAI, pp 81-90, 2019.

Authorship Attribution Statement

Chapter 3 of this thesis is published as **Publication 2**. I co-designed the study with clinicians, conducted the experiments, analyzed the data and wrote the manuscript.

Chapter 4 of this thesis is published as **Publication 4**. I co-designed the study with clinicians, conducted the experiments, analyzed the data and wrote the manuscript.

Chapter 5 of this thesis is published as **Publication 9 and 10**. I designed the study, conducted the experiments, analyzed the data and wrote the manuscript.

Chapter 6 of this thesis is published as **Publication 1**. I designed the study, developed the methodology, analyzed the data and wrote the manuscript.

Acknowledgements

I am deeply grateful to all people who provided valuable support and assistance for my Ph.D. study and thesis.

First, I would like to express my sincere thanks to my supervisor Assoc. Prof. Xiuying Wang, for her unwavering support, insightful advice, and patient guidance. The immense knowledge and plentiful professional insights have deeply impressed me and encouraged me all the time in my academic research and daily life. Without her constant guidance and feedback, this PhD degree would not have been achievable.

I also want to thank my associate supervisor, Prof. David Feng, IEEE fellow, for his kind support. With his dedication and keen interest in research, he has set a great role model for me.

Big thanks to clinical doctors whom I worked with for their collaborative support. The active collaboration on the stages of research projects benefits me along my Ph.D. journey. I would like to offer special thanks to Dr. Jing Huang and Prof. Jie Lu from Beijing Xuanwu Hospital; Prof. Shaoli Song from Fudan Cancer Hospital; Dr. Linlin Wang, and Dr. Taotao Dong from Shangdong Cancer Hospital for their professional opinions on our collaborated research work.

Furthermore, thanks profusely to all fellow students in our research groups and all my friends for their treasured support, encouragements and company all along the way.

Last but not least, I would like to say a heartfelt thank to my patients, especially my mom and dad, for their endless love and deep belief in me. I would like to thank Yunjie Zhou, who has been by my side during all PhD period, with whom I accomplished this incredible journey together.

Glossary

ACC: Accuracy. 102, 129, 134

ACP: Adaptive Community Profile. 133, 136

ADASYN: Adaptive Synthetic. 101

AMD: Age-related Macular Degeneration. 44

ASD: Autism Spectrum Disorder. 41

AUC: Area Under the Curve. 54, 102

BAC: balanced accuracy. 129, 134

BRF: Balanced Random Forest. 128, 131

CA: cross-modal association. 98, 113

cc: clustering coefficient. 149

CCA: Canonical Correlation Analysis. 8

CCRT: concurrent chemotherapy and radiotherapy. 67, 71

CEA: Carcinoembryonic Antigen. 15

CE-SVM: Cost-effective Support Vector Machine. 128, 131

C-Net: Community-level Network. 117, 121

CNS: Central Nervous System. 45, 46

CPH: Cox Proportional Hazard. 66, 73, 74

CRP: C-reactive protein. 68

CT: Computerized Tomography. 2, 67

Cyfra 211: cytokeratin 19 fragments. 15

DFS: Disease-free Survival. 42

DMFusion: Deep Multimodal Fusion. 8

DNA: deoxyribonucleic acid. 2

DNN: Deep Neural Network. 30

DSS: Disease-specific Survival. 42

DTA: Dynamic Topology Analysis. 114, 116, 117, 153

DTI: Diffusion-tensor Imaging. 18

EDSS: Expanded Disability Status Scale. 47, 101, 130

EEG: electroencephalogram. 2

ESR: erythrocyte sedimentation rate. 68

FDG: fluorodeoxyglucose. 67

FIM: Fisher Information Metrics. 40

fMRI: functional MRI. 2, 18

GLCM: Gray Level Co-occurrence Matrix. 17, 49, 72

GLSZM: Gray Level Size Zone Matrix. 17, 49, 72

G-Net: Global-level Network. 117, 119

GTV: Gross Tumor Volume. 71

HOM: homogeneity. 103, 104

ICS: Integrative Clustering and Supervised. 67, 71

IDH: Isocitrate Dehydrogenase. 41

KNN: K-nearest Neighbors. 33, 139

KPS: Karnofsky performance scores. 15, 69

LA-NSCLC: locally advanced non-small cell lung cancer. 66, 67

LASSO: Least Absolute Shrinkage and Selection. 23–25

LBP: Local Binary Patterns. 102

LMR: Lymphocyte/ Monocyte ratio. 16, 69

LoG: Laplace of Gaussian. 17, 49

lv: lesion volume. 149

MI: Mutual Information. 20, 104

MM-RF: Multi-parametric Multivaraite Random Forest. 45

MRI: Magnetic Resonance Imaging. 1, 2, 45, 46

MS: Multiple Sclerosis. 45, 46, 130

MVC: Multi-view Clustering. 26

NC: Normal Control. 41

NLR: Neutrophil/ Lymphocyte ratio. 16, 69

NMI: normalized mutual information. 103

NMO: Neuromyelitis Optica. 45, 46, 130

NSCLC: non-small cell lung cancer. 67

NSE: Neuron-Specific Enolase. 15

OCT: Optical Coherence Tomography. 44

OS: Overall Survival. 42

PCA: Principle Component Analysis. 4

PCM: Pearson Correlation Matrix. 76

PD: Persistence Diagram. 37, 39

PET: Positron Emission Tomography. 2

PFF: Parallel Feature-level Fusion. 91

PFS: Progression-free Survival. 42

PLR: Platelet/ Lymphocyte ratio. 16, 69

PR_AUC: area under the precision recall curve. 129

RF: Random Forest. 52

RFE: Recursive Feature Elimination. 23–25

RF-SFS: Random Forest-based sequential forward selection. 50

ROC_AUC: Area Under the Receiver Operator Characteristic Curve. 22, 52, 129

ROI: Region of Interest. 2, 25, 48

RSD: Relative Standard Deviation. 52

RSF: Random Survival Forest. 66, 73

SAH: Soft Assignment Hardening. 95

SBS: Sequential Backward Selection. 22

SDC: Self-supervised Deep Correlation. 8

SDC-MVC: Self-supervised Deep Correlational Multi-view Clustering. 92

SEN: Sensitivity. 102, 129, 134

SFF: Serial Feature-level Fusion. 91

SFS: Sequential Forward Selection. 22, 128

SGD: Stochastic Gradient Descent. 103

SHAP: SHapley Additive exPlanations. 8, 45

SOTA: State-of-the-art. 58, 106

SPE: Specificity. 102, 129, 134

SVM: Support Vector Machine. 23

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Motivation

Due to the digital revolution and the advancement of medical technology, increasing amount of medical data are being generated at an unprecedented speed. The digital healthcare data have reached 150 exabytes (10^{18}) in 2011 and will exceed yottabytes (10^{24}) in near future [1]. A study from Ponemon Institute estimated that 30 percent of all electronic data generated in 2012 was from the healthcare industry alone [2]. These figures show the promising potential of medical big data. Medical big data is not only enormous in its size, but also heterogeneous and complex in its data form, which is difficult for conventional systems or algorithms to process. These heterogeneous data include various of structured, and unstructured data (e.g., biomedical imaging, laboratory biomarkers, electronic medical records, and hand-written doctor notes [3]). Thus, information fusion in multimodal data are in urgent need to process such complex, diverse and heterogeneous multimodal data, and has become a vital research topic.

Multimodal data fusion is an emerging field from data mining, which is aimed to integrate data from different distributions, sources, and types for more informed decisions. In a narrow sense, multimodal fusion means the integration of signals and imaging from different devices (multi-modality) or different imaging modes (multi-parametric). For example, medical imaging data and non-imaging biomarker status are expected to be used in multimodality fusion [4], while T1 and T2 Magnetic Resonance Imaging (MRI) data can be used for multi-parametric fusion [5]. In a broader sense, multimodal fusion also refers to the integration of different perspectives of information from the same data (multi-view) or different Region of Interests (ROIs) from the same data (multi-focus) [6]. For example, shape features and texture features of tumor on Positron Emission Tomography (PET) images can be regarded as multi-view inputs providing complementary perspectives [7], while multifocal lesions on the same imaging could provide comprehensive multi-focus information for more precise diagnosis and treatment planning [8].

Multimodal data fusion brings enormous benefits to medical data analysis. On the one hand, multimodal fusion takes advantage of complementary information from different sources, such as structural-functional complementarity (e.g., fusion of PET and Computerized Tomography (CT) imaging [7], [9]), physiological-physical complementarity (e.g., fusion of T1/T2 MRI [5]), resolution complementarity (e.g., fusion of PET/CT imaging [9]), and spatial-temporal complementarity (e.g., fusion of electroencephalogram (EEG) and functional MRI (fMRI) [10]). On the other hand, multimodal fusion exploits essential and structural consensus information from different sources. For example, underlying genetic associations, pathological alternations could be captured from clinical symptoms, medical imaging, and laboratory biomarkers.

For quantiative medical applications, multimodal data fusion contributes to the precise medicine, including personalized care, better planned clinical operations, and improve preventional public health. Firstly, multimodal fusion provides comprehensive information to create predictive models for personalized care (e.g., genomic deoxyribonucleic acid (DNA) sequence for cancer care [11]), which improves the best-practice treatments for patient, enables early detection and diagnosis before symptom signs of patients. Secondly, multimodal data fusion could be used for improving clinical operations by mining multimodal data for better ways of diagnosing and treating patients [12]. Thirdly, multimodal health data can be used for timely diagnosis of the challenging diseases and early prevention of the outbreaks of infectious diseases, thus creating benefits for all the human being [1].

To summarize, the motivation for multimodal data fusion includes 1) urgent demand in the big data era for fusing complex heterogenous data, 2) enormous benefits compared with mono-modal analysis, 3) contributions to medical applications including personalized diagnosis, prognosis, treatment recommendation and disease prevention.

1.2 Challenges

Although there has been extensive research on computational approaches for multimodal data fusion, there are three major challenges for data fusion in quantitative medical applications. The first challenge is high-dimension small-sample multimodal datasets, which may contain a large number of redundant, irrelavant and noisy information, hindering the effective discovery of informative multimodal features. The second challenge is to effectively exploit inter-modal relationships and intra-modal information for precise clinical decisions and interpretation of complex inter-modal relationships. The third challenge is to integrate information from multi-focus regions from the same source of data.

1.2.1 Biomarker Discovery from High-dimensional Small-sample Multimodal Data

The integration of imaging data, which is high dimensional itself, with other modalities (imaging or non-imaging data) further exacerbates the dimensionality problem, thus hindering the discovery of effective multimodal biomarkers. The major issue with the high dimensional data is the "curse of dimensionality" [13], which refers to the phenomenon that the feature space becomes sparse with the increase of the dimension. The highly sparse feature space requires to increase the number of samples exponentially for building a precise model [14]; however, large sample sizes are often not available in medical studies due to the confidentiality issue, law or politic issue, or inconsistency of acquisition protocols [15]. Thus, dimension reduction is required to address the high-dimension small-sample multimodality issue in medical studies.

More specifically, there are four major criteria to reduce dimension and mine multimodal biomarkers from high-dimensional small-sample data in medical studies.

- The first criterion is interpretability, which means that the discovered biomarkers are required to be biological meaningful and interpretable. As a counter example, projection methods such as Principle Component Analysis (PCA) [16] and embedding methods such as graph embedding [17] are not ideal for the biomarker discovery due to the lack of interpretability [18].
- The second criterion is relevancy. In other words, the discovered biomarkers are supposed to be relevant with the study outcome, such as diagnostic labels or prognostic results. This criterion requires the algorithm to remove the noisy and irrelevant features, while retaining the discriminative information.
- The third criterion is non-redundancy. If the final feature set contains highly correlated features, it not only unnecessarily increases the dimensionality, but also affects the predictive capability and interpretability.
- The fourth criterion is generalizability. As imaging acquisition protocol tends to be different in different institute [19], generalizablibility is essential for mined imaging-centric biomarkers to be reproducible across imaging devices and institutions.

To summarize, the first challenge (in feature-level fusion) is to mine interpretable, relevant, non-redundant and generalizable multimodal biomarkers from high-dimensional small-sample multimodal data.

1.2.2 Exploitation and Interpretation on Inter-modal and Intra-modal Information in Multimodal Fusion

Due to the complexity of relationship among different modalities, it is important to exploit inter-modal and intra-modal information during the fusion process, which is difficult to be achieved by conventional dimension reduction in feature-level fusion. The challenges of information-level fusion arise from three perspective:

- In a supervised setting, two major fusion frameworks, radiomics and multi-branch deep learning integrate information in an implicit way, dependent on label information. For example, radiomics mines informative imaging features from regions of interest in medical imaging, before concatenated with non-imaging features for predictive modeling using label information [8]. Similarly, multi-branch deep learning feeds multimodal inputs into the neural network framework with different channels, supervising the feature learning process using label information. However, the implicit fusion has difficult revealing and modeling complex inter-modal consensus and intra-modal discriminative information underlying multimodal data.
- Unsupervised multimodal learning is valuable for medical studies because the acquisition of label information is labor-intensive, cost-ineffective, and sometimes not viable [20]. Unsupervised deep multi-view clustering has been proposed to mine essential consensus inter-modal information; however, a major limitation of consensus multi-view clustering algorithms is that multi-view representation learning and clustering are often decoupled, thus the representations can hardly benefit from feedback from the clustering process.
- Interpreting the complex nonlinear cross-modal association, especially in deep-network-based fusion models, remains an unsolved challenge, which is essential for uncovering the disease mechanism. Early research has investigated methods to interpret cross-modal associations in linear multimodal fusion models such as linear CCA [21]. Such interpretation could be achieved through coefficients in

linear embedding functions [22], canonical loading and cross-loading [23], graphical biplots [24], and probabilistic perspectives [25]. However, linear CCA has only limited capacity to model complex nonlinear interactions. In contrast, deep fusion models (such as deep CCA) are equipped with the strong approximation capability, but are generally more difficult to interpret. The difficulty arises from a great number of nonlinear transformation and operations in deep networks, such as nonlinear activation and kernel convolution. Although recent studies investigated the contribution of input features in deep networks towards classification decision using perturbation-based [26], [27] or propagation-based models [28]–[30], it is still an unsolved challenge of interpreting the nonlinear cross-modal association in deep multimodal fusion.

To summarize, the second challenge (in information-level fusion) is to comprehensively exploit inter-modal and intra-modal information in the supervised or unsupervised setting, and to interpret the non-linear complex cross-modal association of multimodal data.

1.2.3 Knowledge Distillation from Multi-ROI Incorporated with Domain Knowledge

Quantitative knowledge distillation from multi-ROI on a medical imaging is still an unsolved challenge, although characterizing imaging features from single lesions using feature engineering or deep learning has been widely investigated in more recent years. Current feature engineering methods integrate multi-ROI features through averaging; however, the contribution of individual lesions is weakened. Deep learning provides a data-driven approach to learn potential MRI predictors for multi-ROI lesions in MS [31], [32]; however, the extracted deep features are generally difficult to interpret, hardly biologically meaningful [18] and usually requires a large amount of data to train the network. More importantly, both feature engineering and deep learning methods neglect the importance of the inter-lesion spatial relationship of multi-ROI lesions. Thus, a topological profiling tool for multi-ROI lesions is in high demand for systematic analysis of lesion spatial patterns and yet missing in current feature engineering and deep learning methods. Furthermore, there is a lack of method incorporating domain knowledge into the knowledge distillation process from multi-ROI.

To summarize, the third challenge (in knowledge-level fusion) is to quantitatively distill knowledge from inter-ROI relationships with domain knowledge incorporated in the knowledge distillation process.

1.3 Contributions

To address the current challenges in multimodal data fusion, we propose a Multi-level Multimodal Data Fusion framework for feature-level fusion, information-level fusion and knowledge-level fusion.

1.3.1 Feature-level Fusion: Integrative Multimodal Biomarker Discovery Framework for High-dimensional Small-sample Multimodal Data in Diagnostic and Prognostic Prediction

The thesis includes the development of Integrative Multimodal Biomarker Discovery framework, equipped with machine learning techniques, for diagnostic and prognostic predictions. The framework addresses the challenge of mining interpretable, relevant, non-redundant and generalizble multimodal biomarkers from high-dimensional small-sample multimodal data. The framework has the following contributions:

- The framework mines diagnostic multimodal biomarkers with high discriminability, robustness across imaging vendors, from multi-parametric MRI imaging and clinical factors for differentiating two neurological diseases.

- The framework mines prognostic multimodal biomarkers, which are highly relevant, representative and non-redundant from the high-dimensional multimodal feature pool. The mined multimodal biomarkers were composed of CT radiomic features, clinical-pathological and hematological factors for survival prediction of lung cancer.
- The interpretability of multimodal biomarkers was further enhanced with SHapley Additive exPlanations (SHAP) method and nomogram, which has value to permit non-invasive, objective, and dynamic evaluation of lung cancer and can provide a practical reference for individualized patient management.

1.3.2 Information-level Fusion: Interpretable Deep Correlational Fusion Framework for Inter-modal and Intra-modal Information Analysis

The thesis includes the development an Interpretable Deep Correlational Fusion framework based on Canonical Correlation Analysis (CCA) for 1) cohesive fusion of imaging and non-imaging medical data, and 2) assisting with understanding of complex non-linear cross-modal association. The framework has the following contributions:

- A novel Deep Multimodal Fusion (DMFusion) loss is proposed to optimize discovery of informative multimodal representations in a supervised setting, by jointly exploiting inter-modal consensus and discriminative intra-modal information.
- A novel Self-supervised Deep Correlation (SDC) loss is proposed to optimize the unsupervised multi-view clustering, leveraging both multimodal consensus and clustering-oriented discriminative information. This loss couples multi-view learning and self-supervised deep clustering in an end-to-end deep learning network.
- A cross-modal interpretation module is proposed to quantify the importance of input features towards the correlated association.

1.3.3 Knowledge-level Fusion: Dynamic Topology Analysis Framework for Graph-based Knowledge Distillation and Domain Knowledge Incorporation

The thesis includes the development of a Dynamic Topological Analysis framework, based on persistent homology (a higher-order graph model), for quantitatively analyzing the inter-lesion graph knowledge from MRI images, including global geometric structures and local lesion clusters. The framework has the following contributions:

- To distill multi-ROI knowledge based on graph methods, we propose a new K-simplex Filtration to construct a high-level abstraction of global topology for dynamic community identification, based on the connectivity of k-simplex structures in the global Rips complex.
- To quantify the dynamic community structure, we propose a novel Decomposed Community Persistence algorithm to track the dynamic evolution of communities at fine-grained scales.
- To incorporate domain knowledge into graph-based knowledge, we summarize the evolutionary communities incorporated with lesion attributes to integrate community heterogeneity into dynamic community quantification.

1.4 Thesis Organization

The rest of the thesis is organized as Figure 1.1. Literature reviews on algorithms for multimodal data fusion are presented and discussed in Chapter 2 as well as its applications. Our proposed models are presented and discussed in Chapter 3-6. Specifically, feature-level fusion framework for mining multimodal biomarkers for diagnostic and prognostic tasks is presented in Chapter 3-4 respectively. Information-level fusion framework to integrate inter-modal and intra-modal information is presented in Chapter 5. Knowledge-level fusion to integrate graph-based multi-ROI knowledge with domain knowledge is introduced in

Chapter 6. Finally, Chapter 7 exhibits our conclusions and future prospects of multimodal data fusion.

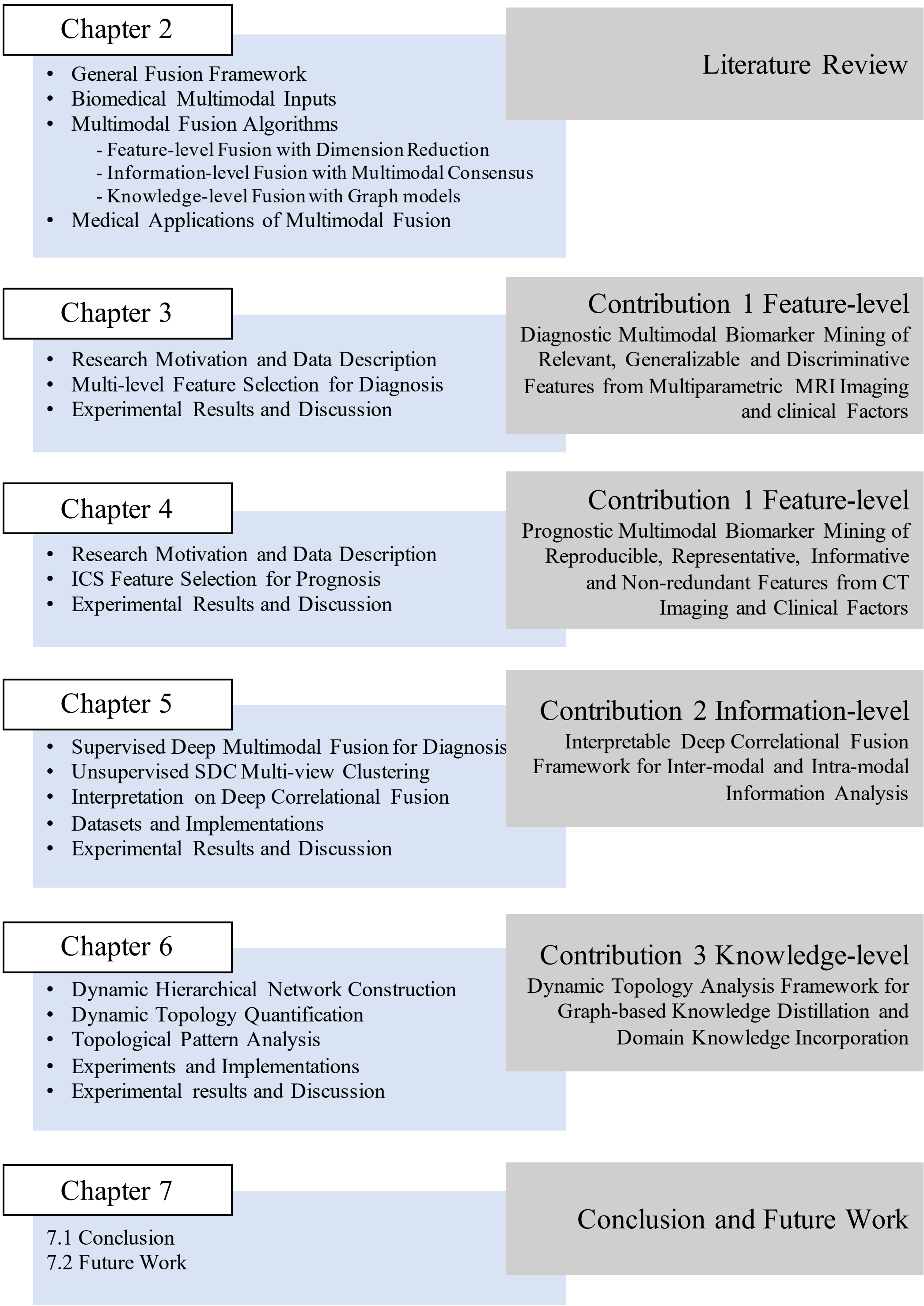

Chapter 2
• General Fusion Framework
• Biomedical Multimodal Inputs
• Multimodal Fusion Algorithms
- Feature-level Fusion with Dimension Reduction
- Information-level Fusion with Multimodal Consensus
- Knowledge-level Fusion with Graph models
• Medical Applications of Multimodal Fusion
Literature Review
Chapter 3
• Research Motivation and Data Description
• Multi-level Feature Selection for Diagnosis
• Experimental Results and Discussion
Contribution 1 Feature-level
Diagnostic Multimodal Biomarker Mining of Relevant, Generalizable and Discriminative Features from Multiparametric MRI Imaging and clinical Factors
Chapter 4
• Research Motivation and Data Description
• ICS Feature Selection for Prognosis
• Experimental Results and Discussion
Contribution 1 Feature-level
Prognostic Multimodal Biomarker Mining of Reproducible, Representative, Informative and Non-redundant Features from CT Imaging and Clinical Factors
Chapter 5
• Supervised Deep Multimodal Fusion for Diagnosis
• Unsupervised SDC Multi-view Clustering
• Interpretation on Deep Correlational Fusion
• Datasets and Implementations
• Experimental Results and Discussion
Contribution 2 Information-level
Interpretable Deep Correlational Fusion Framework for Inter-modal and Intra-modal Information Analysis
Chapter 6
• Dynamic Hierarchical Network Construction
• Dynamic Topology Quantification
• Topological Pattern Analysis
• Experiments and Implementations
• Experimental results and Discussion
Contribution 3 Knowledge-level
Dynamic Topology Analysis Framework for Graph-based Knowledge Distillation and Domain Knowledge Incorporation
Chapter 7
7.1 Conclusion
7.2 Future Work
Conclusion and Future Work

FIGURE 1.1. The outline of the thesis.

CHAPTER 2

Literature Review

2.1 Generalized Fusion Framework

In this chapter, we summarize literature review on multimodal fusion related to quantitative medical applications as a Generalized Fusion Framework, which is illustrated in Figure 2.1. Literature review starts with a) Multimodal Medical Inputs; followed by three major streams of b) Multimodal Fusion Algorithms; and wrapped up with current c) Medical Fusion Applications.

Multimodal Medical Inputs can be categorized into imaging data and non-imaging data, as shown in Figure 2.1a. In a more generalized form, the concept of multimodal data can be extended to multi-view features and multi-focus regions because they share the similar aim of integrating different perspectives of information. For Multimodal Fusion Algorithms, literature is summarized as three major streams of algorithms including Feature-level Fusion, Information-level Fusion, and Knowledge-level Fusion. Each level of fusion differs in 1) the depth of representation to fuse, 2) fusion mechanisms, 3) and interpretation methods, as shown in Figure 2.1b. More specifically, feature-level fusion focuses on integrating high-dimensional features extracted from imaging or non-imaging data using carefully designed dimension reduction methods. Feature-level fusion can be enhanced with feature interpretation modules by revealing the feature clinical meaning and its importance. By comparison, information-level fusion focuses on exploiting cross-modal consensus and intra-modal complementary information using co-regularized fusion methods. The information interpretation majorly focuses on revealing the relationship between multi-modalities. For knowledge-level fusion, the major interest is to distill knowledge from inter-connected multi-focus representations using graph-based methods. The knowledge interpretation majorly focuses on the visualization and analysis of graph-based vertex relationships. Lastly, Medical Fusion Applications are summarized into three categories, including diagnostic classification, prognostic regression and unsupervised clustering, as shown in Figure 2.1c.

This chapter is organized as follows: Multimodal Medical Inputs are summarized in Section 2.2. Three-level Multimodal Fusion Algorithms, including feature-level, information-level and knowledge-level, are introduced in Section 2.3-2.5 respectively. Finally, we present the Medical Fusion Applications in Section 2.6.

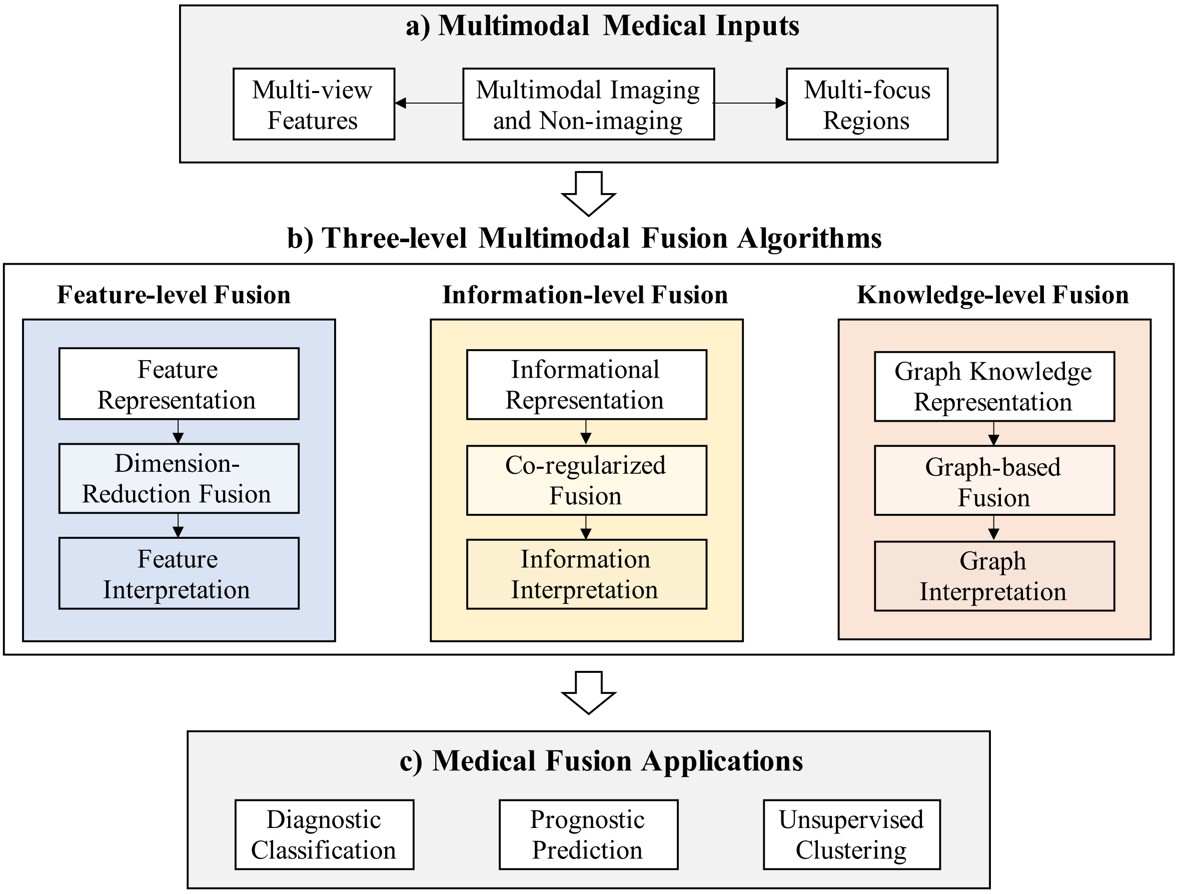

a) Multimodal Medical Inputs
Multi-view Features
Multimodal Imaging and Non-imaging
Multi-focus Regions
b) Three-level Multimodal Fusion Algorithms
Feature-level Fusion
Feature Representation
Dimension-Reduction Fusion
Feature Interpretation
Information-level Fusion
Informational Representation
Co-regularized Fusion
Information Interpretation
Knowledge-level Fusion
Graph Knowledge Representation
Graph-based Fusion
Graph Interpretation
c) Medical Fusion Applications
Diagnostic Classification
Prognostic Prediction
Unsupervised Clustering

FIGURE 2.1. Generalized multimodal fusion framework.

2.2 Biomedical Multimodal Inputs

Biomedical multimodality inputs, including imaging and non-imaging modalities, are widely used to examine body anatomy, functionality, and suspicious symptoms for diagnostic or prognostic purposes. Medical imaging can be generally divided into anatomical imaging and functional imaging, to provide structural and metabolic information, respectively. Three commonly used medical imaging techniques (including CT, PET and MRI) are described in Section 2.2.1 , with a comparison summarized in Table 2.1. Medical non-imaging that are commonly used in the clinical practice, include clinicopathological factors, hematological factors, and other multi-omic factors (Section 2.2.2). The definition of multimodality can be further extended to multi-view features (such as intensity, texture, shape features of tumors from imaging and clinical factors from non-imaging) and multi-focus regions (such as multi-focal lesions, multiple brain regions and multiple cells on imaging). They are described in Section 2.2.3 and 2.2.4, respectively.

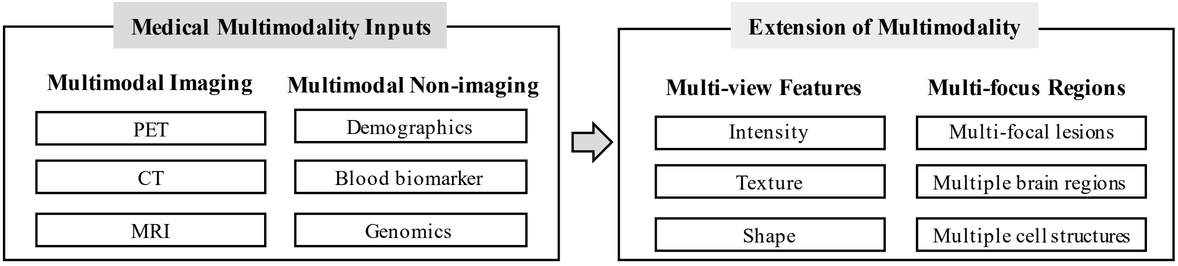
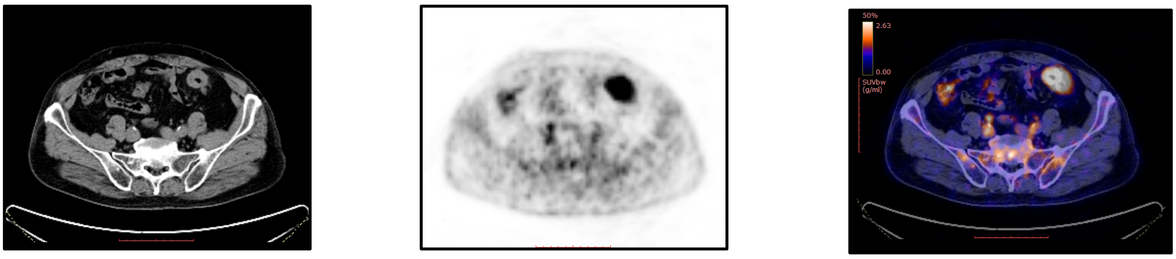


FIGURE 2.2. Medical multimodality inputs.

2.2.1 Medical Imaging

Computerized Tomography (CT). CT deploys a motorized x-ray source to cover multiple angles and creates cross-sectional images using reconstruction algorithms. The mechanism underlying CT images is based on X-ray attenuation of tissues, however, it provides more structural information compared to conventional fixed X-ray source. Other advantages of CT images include comparatively high resolution, excellent contrast on bones, thus they are recommended for anatomical imaging of injuries or diseases (such as lung tumours [33], and



CT PET PET/CT

FIGURE 2.3. Medical images: CT, PET and PET/CT (colorectal cancer).

different types of heart diseases [34]). However, the major limitations of CT are the radiation exposure and low contrast on soft tissues.

Positron Emission Tomography (PET). Different from CT using radioactive substance to check the tissue functions, PET detects the biochemical changes from positron annihilation in body tissues. The change of biochemical substances can reveal the metabolic process of the target body regions; thus helping detect the onset of the disease, which is sometimes even hardly visible on anatomical imaging. Compared with CT, the major advantage of PET is that only a small amount of radioactive substance is required to examine targeted regions. And it has a wide range of diagnostic applications for oncology [35], neurology, and cardiology. However, the limitations of PET imaging include relatively low resolution and high cost.

Magnetic Resonance Imaging (MRI). MRI uses strong magnetic field to generate anatomic imaging reflecting the physiological process of the body. The major advantage is that MRI does not involve any radiation material, which would potentially increase the risk of cancer; thus, it becomes the safest choice in medical procedures. Another advantage of MRI is its more apparent contrast on soft tissues and widely used for the diagnosis of body regions (e.g., brain, spinal cord, breast, and blood vessels) [36], [37]. However, the limitations of MRI include expensive equipment, high maintenance cost, and lack of bone contrast.

T1 and T2 MRI. T1 and T2 are two basic types of MRI, which are commonly used in clinical routines. The images of these two sequences of MRI are illustrated in Figure 2.4. These two MRI sequences are generated using different pulse sequence using different timing

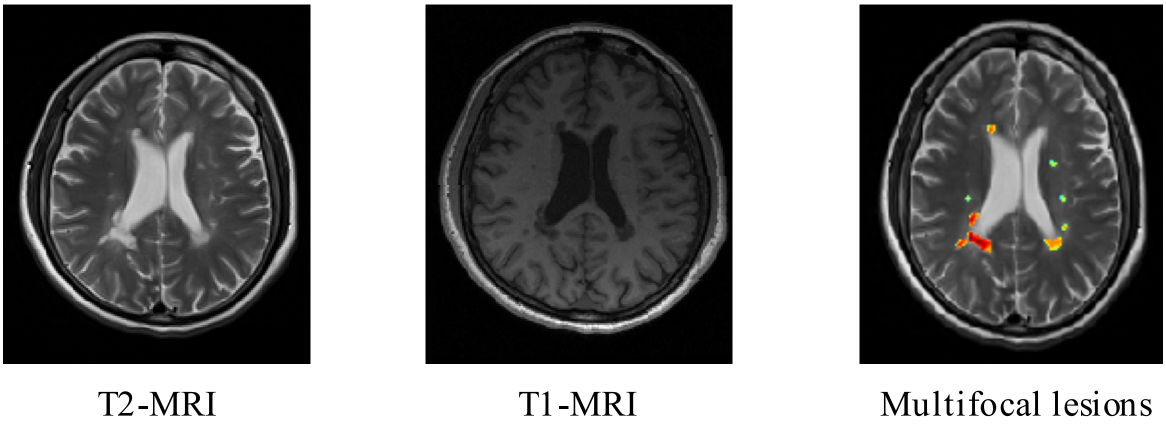


FIGURE 2.4. Medical images: T2-MRI, T1-MRI, and multi-focal lesions on T2 (multiple sclerosis).

to highlight different interested regions. T1 MRI relies on the longitudinal relaxation time of the tissue to generate imaging. The common usage of T1-MRI include fatty tissues, reflecting anatomical information and examine liver lesions. In contrast, T2 MRI relies on the transverse relaxation time to generate imaging. The common applications include detecting white matter lesions in brain, and examining inflammation and edema.

2.2.2 Medical Non-imaging

Clinicopathological factors. Clinicalpathological characteristics are essential for diagnosis and monitoring the disease progression. These factors include age, sex, tumor location, tumor size, node metastasis status, histological type, Karnofsky performance scores (KPS), radiation type and doses, concurrent chemotherapy type, and usage of consolidative chemotherapy, and pre- and post-therapeutical serum tumor biomarkers. These tumor biomarkers include Carcinoembryonic Antigen (CEA), Neuron-Specific Enolase (NSE), and cytokeratin 19 fragments (Cyfra 211). KPS is to quantify patients' ability to tolerate therapy in terms of their physical function and ability to take care of themselves and to perform daily activities.

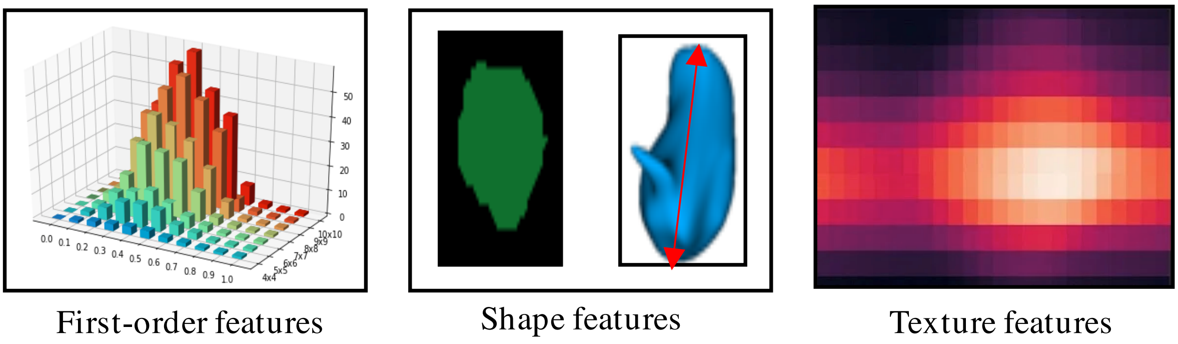
Hematological factors. Hemoatological factors, also known as blood biomarkers, are important indicators for cancer diagnosis and treatment monitoring. The blood biomarkers are

TABLE 2.1. Comparison of medical multimodal imaging.

Modality	**CT**	**PET**	**MRI**
Contrast mechanism	X-ray attenuation of tissues	Photon emmsion after positron annihilation	Emitted RF signal after nuclear spin excitation
Spatial resolution	<=100 μm	1-2 mm	<=100 μm
Acquisition time	10-25 min	10-90 min	5-60 min
Advantages	1) Excellent bone imaging	1) High sensitivity 2) high range of applications	1) Non-ionizing radiation 2) apparent soft tissue contrast
Limitations	1) Radiation dose 2) low soft tissue contrast	1) High cost 2) use of radioactive agents	1) Expensive equipment 2) high maintenance costs 3) lack of bone contrast
Main applications	1) Anatomic imaging (bone)	1) Diagnostic imaging (oncology, neurology, cardiology) 2) pharmacokinetic imaging	1) Anatomical imaging (soft tissue)

collected from routine blood tests. They are highly attractive because they provide clinically relevant noninvaisve predictors, potentially complementary to imaging characteristics. However, the major weakness of blood biomarkers is the inconsistent and non-standard cut-off values for dividing the clear boundary of responders, and non-responders [38]. These blood biomarkers include levels of monocytes, neutrophils, lymphocytes, hemoglobin, and platelet counts. Also, Neutrophil/ Lymphocyte ratio (NLR), Lymphocyte/ Monocyte ratio (LMR), and Platelet/ Lymphocyte ratio (PLR) were calculated for each patient.

Other multi-omic factors. Vast amounts of data from other sources (such as DNA, RNA, and protein) can be mined and analyzed to predict disease risks, treatment response, and prognosis. Such large data are often referred as multi-omic factors (such as genomics, transcriptomics, proteomics) [39]. However, the clinical deployment of multi-omic biomarkers is still hindered by several important factors. Firstly, there are false positive discoveries due to the large amount of biomarker analysis in global 'omics' studies [40]. Secondly, there are technical reproducibility issue due to non-standard multi-omic data acquisition [41].



First-order features

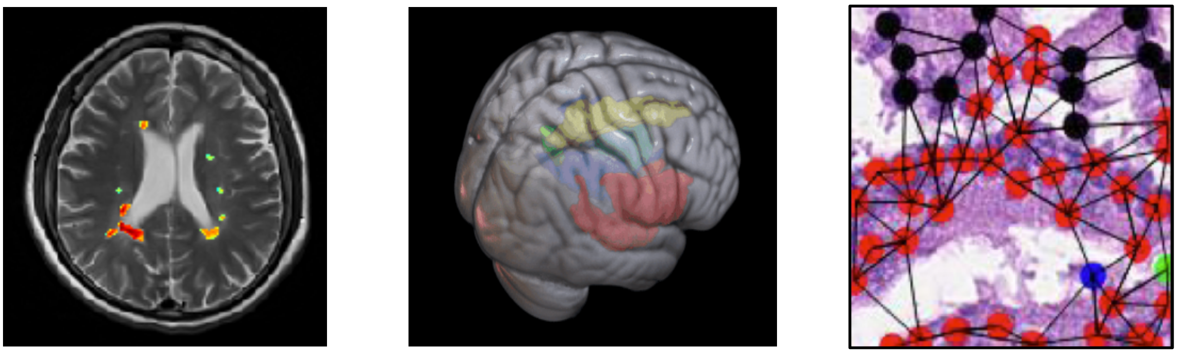
Shape features

Texture features

FIGURE 2.5. Multi-view features: first-order, shape and texture features.

2.2.3 Multi-view Features

In a broader sense, the definition of multimodality can be extended as different perspectives of information from the same source, such as different types of features extracted from the same imaging. Such multiple sets of features are often referred as multi-view features [42], and are widely used in medical applications to characterize tumor heterogeneity and geometric properties. The example of multi-view features include first-order (intensity), shape, texture features and filter-based features, which are usually extracted from a region of interest in medical imaging. These features are illustrated in Figure 2.5. Specifically, first-order features are calculated based on the first-order statistics of the image intensity distribution. Shape features are geometric measurements based on edges, ridge and angle of ROI, including the calculation of the volume, surface area, compactness, and spherical ratio. Texture features are computed with higher-order texture matrices such as Gray Level Co-occurrence Matrix (GLCM) and Gray Level Size Zone Matrix (GLSZM) to quantify lesion heterogeneity. Filter-based features, including Laplace of Gaussian (LoG) and wavelet features were extracted from the filtered images to enhance specific parts of images, such as sharp edges or fine texture. More details on the definition and calculation of intensity, shape and texture features can be found in [43]. Similar to multimodal fusion, effective integration of multi-view features for more precise prediction is a still unsolved challenge [42].



Multifocal lesions

Multiple brain regions

Multiple cells

FIGURE 2.6. Multi-focus regions: multifocal lesions, multiple brain regions and multiple cells.

2.2.4 Multi-focus Regions

Multi-focus fusion is another extension of multimodal fusion, aiming to integrate different region of interest from the same source. Different from conventional single-lesion texture analysis, the challenge of multi-focus fusion is to integrate potentially heterogeneous information in different focused regions, such as different texture characteristics in different lesions. Typical examples of multi-focus fusion include fusion of multi-focal lesions on an MRI imaging [8], fusion of different brain regions using fMRI or Diffusion-tensor Imaging (DTI) [44], [45], fusion of different cells using a pathological imaging [46], [47]. The image examples of multi-focus regions are illustrated in Figure 2.6.

2.3 Feature-level Fusion with Dimension Reduction

In feature-level fusion, high-dimensional data generated in multimodal feature concatenation is a major challenge, which is further exaggerated by the small sample size in medical applications. Feature selection has been a promising solution to high-dimensional small-sample feature fusion problems, because it reduces the effect of the curse of dimensionality and the computational cost, and more importantly it preserves the physical meaning of features and helps understanding of data. Generally, feature selection algorithms can be categorized as filter, wrapper, and embedded methods. Filtering methods act as a preprocessing, removing irrelevant features according to certain ranking criteria (Section 2.3.1). Wrapper methods select features based on their predictive performance given by classifiers (Section 2.3.2). Embedded methods include the feature selection process inside the classifier training process (Section 2.3.3).

TABLE 2.2. Comparison of feature selection techniques.

Feature selection	**Advantage**	**Disadvantage**	**Examples**
Filter	1) Lower computational cost 2) Fast 3) Generalizability	1) No interaction with classifiers 2) suboptimal predictive ability	1) Correaltion-based 2) MI-based 3) Consistency-based
Wrapper	1) Captures feature interaction 2) Interaction with classifier	1) Computational expensive 2) Risk of overfitting	1) SFS 2) Genetic algorithm
Embedded	1) Captures feature interaction; 2) Interaction with classifier 3) Relatively low computational cost	1) classifier-dependent selection	1) Lasso 2) SVM-RFE

2.3.1 Filter Methods

Filtering methods select highly ranked features and filter out less-relevant features according to the feature relevancy criteria, as illustrated in Figure 2.7. It is commonly used in practical applications, especially in the medical domain because of its simplicity, low computational cost, time-efficiency and more likely to avoid overfitting [48], [49]. Different filter methods use different criteria to measure feature relevance, which is the capacity of features to differentiate classes. Examples of relevancy measurement include correlation, mutual information, consistency, and etc.

Correlation-based filter. Correlation-based filter [50] ranks features according to their correlation with the targeted outcome in a heuristic way. The desired feature subsets, acquired

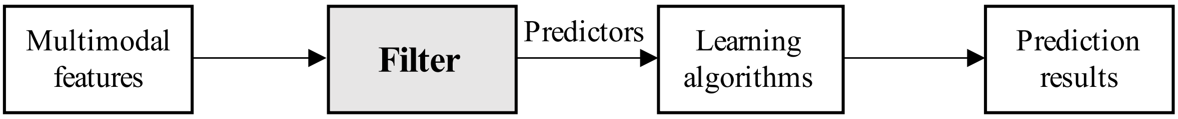


FIGURE 2.7. Feature selection: filter methods.

by correlation-based filter, are features that are highly correlated with the target class while remaining uncorrelated within the subset. Irrelevant features (low correlation with the targets) and redundant features (high correlation within feature sets) should be filtered out. Pearson correlation, also known as linear correlation is mostly used in clinical applications, which are defined as:

$$R(i) = \frac{cov(x_i, Y)}{var(x_i) * var(Y)} \tag{2.1}$$

where x_i is the input features, Y is the output labels, cov denotes the covairance and var denotes the variance. Non-linear correlation can be assessed with Spearman correlation, by measuring the monotonic relationship between the features and outcomes.

Mutual Information (MI)-based filter. MI-based filter [51] relies on information theoretical measurements to assess the relevance of variables. The definition of MI is given below:

$$I(Y, X) = H(Y) - H(Y|X) \tag{2.2}$$

where $H(Y) = -\sum_y p(y)log(p(y))$ represents the entropy in output Y, and $H(Y|X) = -\sum_x \sum_y p(x, y)log(p(y|x))$ is conditional entropy of Y by observing X. The computed MI value $I(Y, X)$ implies the additional information that one variable can offer about the other, thus it proves dependency. If two variables are independent, MI will be zero; otherwise, MI will be greater than zero. The above definition of MI is for discrete variables, and it can be extended to continuous variables if replacing summations with integration.

Consistency-based filter. Consistency-based filter [52] removes irrelevant and redundant features by finding the minimum number of features to form the feature subset, which has the same level of consistency in the class values. However, this method does not take the dependence among features into account.

RELIEF filter. RELIEF filter [53] is a notably sensitive to feature interaction, which ranks the irrelevance to the target using feature value differences between the nearest neighborhood instance pairs. Specifically, if the feature value difference of nearest instance pairs is observed within the same class, the feature rank decreases; otherwise, the feature rank increases. It has inspired a family of RELIEF-based algorithms, such as ReliefF [54] and adapted to a wide range of applications. However, the drawback of the RELIEF filter is that arbitrary threshold is required for feature selection.

To summarize, filter methods have advantages such as computationally light and less likely to overfit because they do not rely on learning algorithms. There are several drawbacks: 1) the predictive performance of selected subset is not optimized [55]; 2) feature interaction is largely discarded [56]; 3) there is a lack of idea method to select the ideal subsequent learning algorithm and the optimal dimension of features [50].

2.3.2 Wrapper Methods

Unlike filter methods using relevance criteria for feature selection, wrapper methods (Figure 2.8) use predictive performance of features given by learning algorithms to select features. Wrapper methods can be further divided into univariate wrapper and multivariate wrapper. Univaraite wrapper ranks the predictive capability of individual features based on the classifiers' results. However, such methods ignore feature interactions. To search for the optimal feature combination, since evaluating all subset combination of features (2^N) is a NP-hard problem, wrapper methods often employ search algorithms to reduce complexity, which can be broadly classified into sequential selection wrapper and heuristic search wrapper. The sequential selection wrapper starts with an empty or full feature set to search for an optimal subset by adding or removing features. The heuristic search wrapper evaluates subsets generated in the search space to optimize the objective function.

Univariate wrapper. Univariate wrapper methods have been widely used in clinical studies to select predictive prognostic and diagnostic biomarkers based on the predictive performance

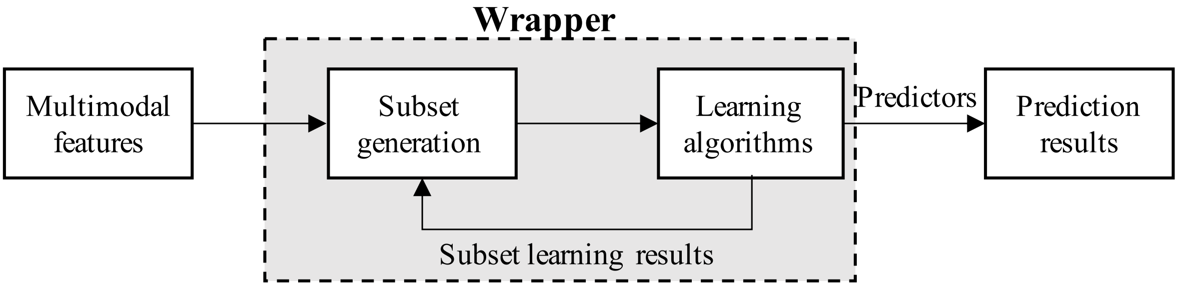


FIGURE 2.8. Feature selection: wrapper methods.

of corresponding regression or classification models. For example, Univariate Cox [57] is a standard feature selection technique that is widely used in survival analysis. It ranks features by C-index computed with Cox proportion hazard model, a classic supervised method designed for time-to-event regression analysis. Another example of univariate wrapper uses Random Forest [58] to select informative features using the classification results, such as Area Under the Receiver Operator Characteristic Curve (ROC_AUC). Though univariate methods are usually efficient and fast compared with multivariate methods, the interaction among features are not adequately considered for the optimal performance.

Sequential selection wrapper. The sequential selection wrapper [59], [60] is named after its iterative nature of the algorithm. The most prominent example, Sequential Forward Selection (SFS), starts the feature selection process with an empty set and then iteratively adding individual features into the feature set to achieve highest predictive performance in each step. The process is repeated until the number of features meets the requirement. The other similar variant, Sequential Backward Selection (SBS), begins the feature selection process from the complete feature sets and then iteratively removes the feature giving the lowest decrease of performance at each step. However, SBS is often computational expensive for high-dimensional feature set to acquire a low-dimensional subset. The other improved version of SFS is name sequential floating forward selection, which introduces additional backtracking steps. An additional step is added to SFS to exclude a feature from the selected subset at each iterate. If the reduced feature set yields better performance, this feature is permanently dropped for the next round of SFS. The drawback of the SFS and SFFS is that redundant features (highly correlated) may be included in the final feature set [61].

Heuristic search wrapper. Compared with the deterministic style used by sequential selection wrapper, Genetic algorithms [62], [63] employ a randomized heuristic approach to identify the optimal feature subset using a designed objective function. The parameters in genetic algorithms can be randomly modified, mimicking the genetic mutation and evolution, to search for the most predictive feature subset. However, it is computationally expensive, particularly for high-dimensional multimodal problems [64]. In many problems, genetic algorithms tend to converge towards local optimal or even arbitrary points rather than the global optimal [65].

2.3.3 Embedded Methods

The embedded methods, as illustrated in Figure 2.9 combine the merits of filter and wrapper methods, embedding the feature selection in the learning algorithm as part of the training process. The main aim of embedded methods is to reduce the computation time evaluating feature combinations in the wrapper methods. The common strategy of embedded methods is to leveraging weights in the classifier as criteria to rank and filter out features. Typical examples include Least Absolute Shrinkage and Selection (LASSO) and Support Vector Machine (SVM)-Recursive Feature Elimination (RFE).

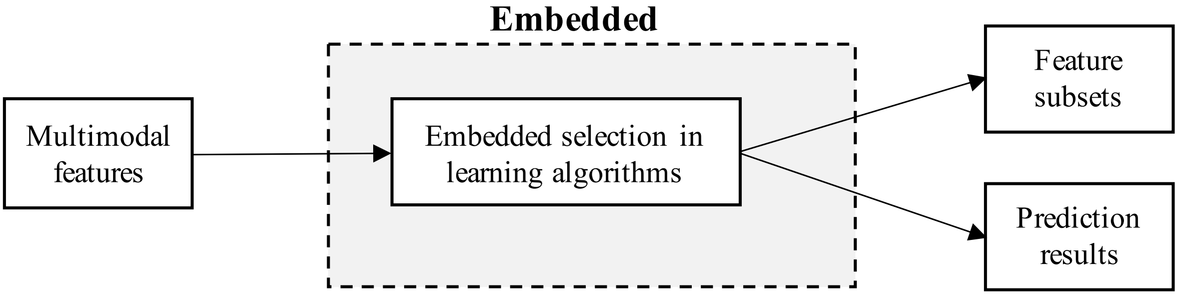


FIGURE 2.9. Feature selection: embedded methods.

Lasso. Lasso is an example of embedded feature selection methods based on a linear model [66]. It penalizes the sum of absolute values of the parameters in the linear model, thus the sum has to be less than a fixed upper bound. To achieve this goal, a regularization process is applied on the parameters of the linear model to shrink some parameters to zero.

These features associated with zero weight will be filtered out, while the non-zero features will be used to minimize the prediction. In this way, the filter process is embedded in the linear model to select discriminative features and reduce the complexity of the model. The cost function for LASSO can be formulated as:

$$\sum_{i=1}^{M}(y_i - \hat{y}_i) = \sum_{i=1}^{M}(y_i - \sum_{j=0}^{p} w_j * x_{ij})^2 + \lambda \sum_{j=0}^{p} |w_j| \tag{2.3}$$

where x is the input, y is the output, w is the weight and $\sum_{j=0}^{p} |w_j| < t$ for some $t > 0$.

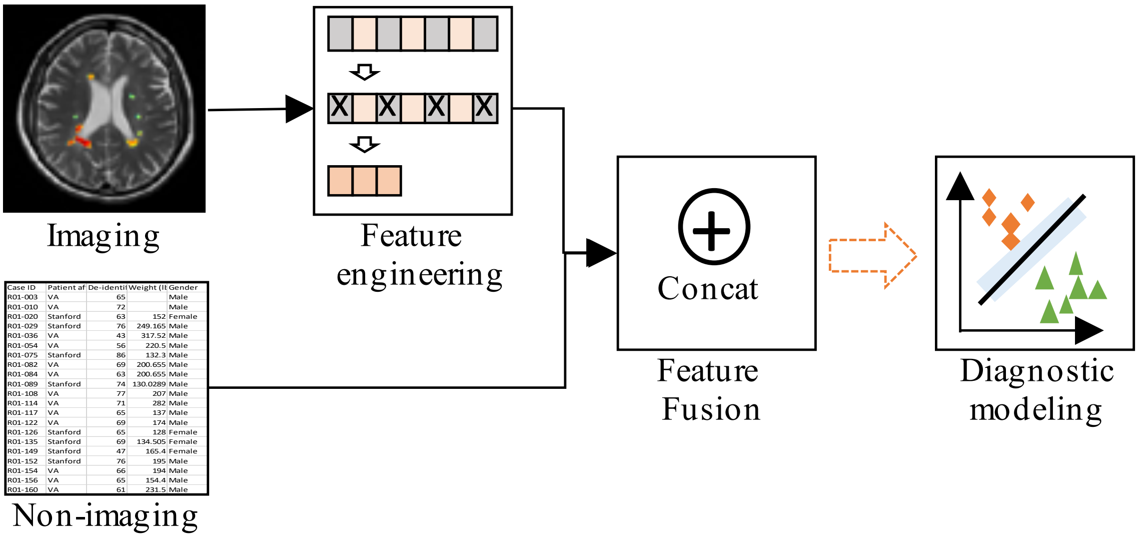
SVE-RFE. RFE is an example equipped with non-linear classifier [67]. In RFE, the full set of features is initially fit with a learning algorithm, and then the least important features (computed with coefficients in the learning algorithm) will be removed until the desired number of features is selected. Compared with sequential selection wrapper, RFE is more computational efficient as less number of feature combination is required to compute.

2.4 Information-level Fusion with Multimodal Consensus

In information-level multimodal fusion, the methods can be categorized into two groups depending on the availability of the label information - Supervised Multimodal Learning (Section 2.4.1) and Unsupervised Multimodal Clustering (Section 2.4.2). To better understand Unsupervised Multimodal Clustering, we briefly introduce a major stream of Multimodal Clustering algorithm, Correlational Consensus Learning in Section 2.4.3. Furthermore, the related work for interpreting deep multimodal associations is summarized in Section 2.4.4.

2.4.1 Supervised Multimodal Learning

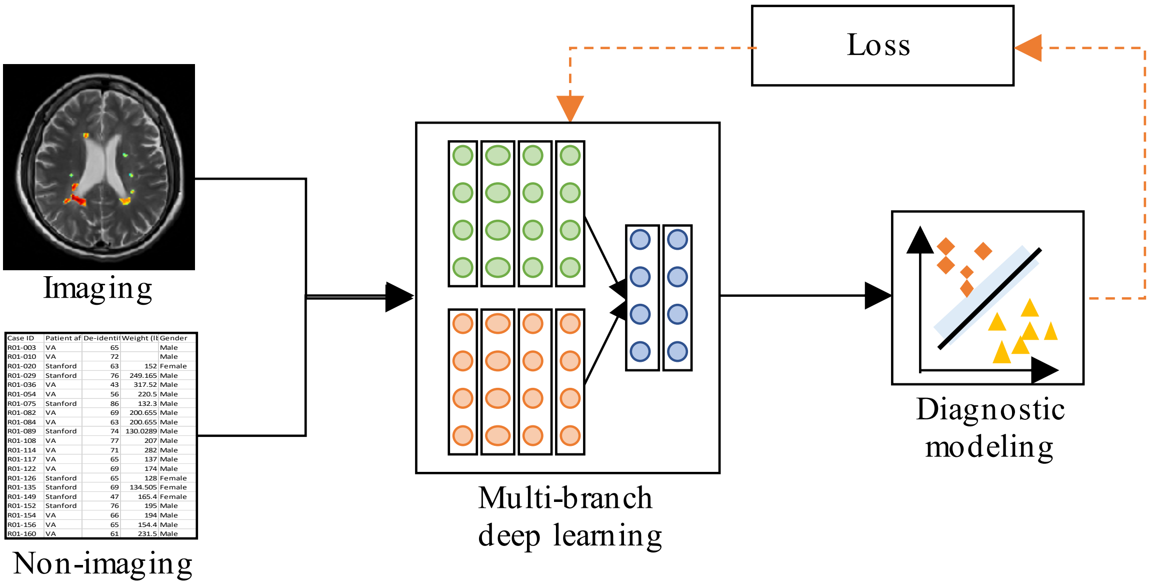
To address the challenge of fusing multimodalies in a supervised setting, radiomics and multi-branch deep learning are two mainstream frameworks in current medical applications.



Case ID	Patient af	De-identi	Weight (lb	Gender
R01-003	VA	65		Male
R01-010	VA	72		Male
R01-020	Stanford	63	152	Female
R01-029	Stanford	76	249.165	Male
R01-036	VA	43	317.52	Male
R01-054	VA	56	220.5	Male
R01-075	Stanford	86	132.3	Male
R01-082	VA	69	200.655	Male
R01-084	VA	63	200.655	Male
R01-089	Stanford	74	130.0289	Male
R01-108	VA	77	207	Male
R01-114	VA	71	282	Male
R01-117	VA	65	137	Male
R01-122	VA	69	174	Male
R01-126	Stanford	65	128	Female
R01-135	Stanford	69	134.505	Female
R01-149	Stanford	47	165.4	Female
R01-152	Stanford	76	195	Male
R01-154	VA	66	194	Male
R01-156	VA	65	154.4	Male
R01-160	VA	61	231.5	Male

FIGURE 2.10. Supervised multimodal fusion: radiomics framework.

Radiomics framework. Instead of directly using raw images, radiomics adopts ROI-based handcraft feature extraction and feature selection to preserve informative features, thus narrowing the dimension gap with clinical factors [68], [69]. In radiomics, simple concatenation with feature selection techniques such as LASSO and RFE [70] is often used to fuse multimodal features for diagnostic prediction. However, handcrafted-based feature engineering in radiomics has limited capacity to model complex multimodal representations.



Case ID	Patient af	De-identi	Weight (lb	Gender
R01-003	VA	65		Male
R01-010	VA	72		Male
R01-020	Stanford	63	152	Female
R01-029	Stanford	76	249.165	Male
R01-036	VA	43	317.52	Male
R01-054	VA	56	220.5	Male
R01-075	Stanford	86	132.3	Male
R01-082	VA	69	200.655	Male
R01-084	VA	63	200.655	Male
R01-089	Stanford	74	130.0289	Male
R01-108	VA	77	207	Male
R01-114	VA	71	282	Male
R01-117	VA	65	137	Male
R01-122	VA	69	174	Male
R01-126	Stanford	65	128	Female
R01-135	Stanford	69	134.505	Female
R01-149	Stanford	47	165.4	Female
R01-152	Stanford	76	195	Male
R01-154	VA	66	194	Male
R01-156	VA	65	154.4	Male
R01-160	VA	61	231.5	Male

FIGURE 2.11. Supervised multimodal fusion: deep multi-branch fusion network.

Multi-branch deep learning framework. In contrast, multi-branch deep learning adopts a data-driven approach to implicitly model the complex multimodal relationship and learn a lower-dimensional representation from different channels of inputs. These multi-branch deep networks adopt supervised architectures such as Multimodal Convolutional Neural Network [71], Multimodal Deep Polynomial Networks [72]. However, multi-branch deep learning, which uses implicit fusion, still has difficulty exploiting cross-modal information caused by large dimension gap between imaging and non-imaging data [73], [74].

2.4.2 Unsupervised Multimodal Clustering

Instead of solely depending on label information to implicitly mine cross-modal associations, unsupervised multimodal clustering explicitly explores inter-modal common structural information through co-regularization for the clustering task. It uses different views of single-source data to partition samples into different groups in an unsupervised manner, thus it is also known as Multi-view Clustering (MVC). These algorithms can be divided into conventional MVC and deep MVC. In addition, an emerging paradigm of deep mono-view clustering for exploiting intra-modal discriminative information is also summarized.

Conventional MVC. To exploit the essential and intrinsic structure hidden in multi-view data, consensus MVC algorithms co-regularize different views to a shared common space for clustering until the consensus is reached. According to the form of the common space, consensus MVC can be divided to common-matrix-based MVC and correlation-based MVC, as illustrated in Figure 2.12. The former branch of algorithms co-regularizes multi-views to a common matrix, which is then used for data partition with an existing clustering method (e.g., k-means). Examples of consensus matrices include common eigenvector matrix used by multi-view spectral clustering [75], common coefficient matrix by multi-view subspace clustering [76] and common indicator matrix by multi-view non-negative matrix factorization clustering [77]. The other branch, correlation-based MVC, projects multi-views to a consensus space by maximizing the correlation between the projected multi-view features and subsequently employing an existing clustering method. Such projection can be obtained through canonical correlation analysis [21] and its variants [78], [79]. However, most of these conventional MVC algorithms adopt linear or shallow embedding functions to reveal intrinsic structure underlying multi-view data, which have difficulty well simulating the complex nonlinear characteristics of large-scale real-world data.

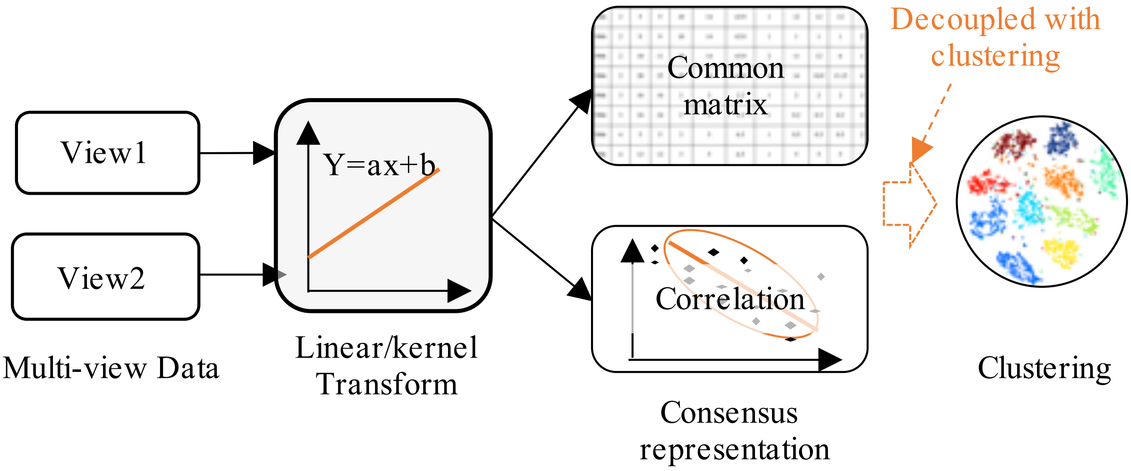


FIGURE 2.12. Unsupervised multimodal fusion: conventional consensus MVC.

Deep MVC. Empowered with strong approximation capacity of neural networks, deep learning has been introduced to model non-linear relationships among multi-views for the clustering task. As illustrated in Figure 2.13, deep MVC is extended from conventional common-matrix-based MVC and correaltion-based MVC. Specifically, conventional common-matrix-based MVC has been extended as deep multi-view spectral clustering [80], deep multi-view subspace clustering [81] by non-linearly projecting multi-view data to the latent common matrix via deep networks constrained with affinity loss or self-representation loss. However, information from different views is entangled in the latent common matrix, thus it is usually difficult to interpret. In contrast, more attention has been recently drawn to deep learning extensions of correlation-based MVC (Deep CCA [82], DCCAE [83], VCCA [84] and DGCCA [85]), as the correlation of the projected multi-view representations can be explicitly measured and interpreted. To be more specific, these correlation-based deep MVC methods non-linearly project multi-view data via deep learning to highly correlated representations enforced by a CCA loss for subsequent clustering. However, a major limitation of the aforementioned consensus MVC algorithms is that multi-view representation learning and clustering are often decoupled, thus the representations can hardly benefit from feedback from the clustering process.

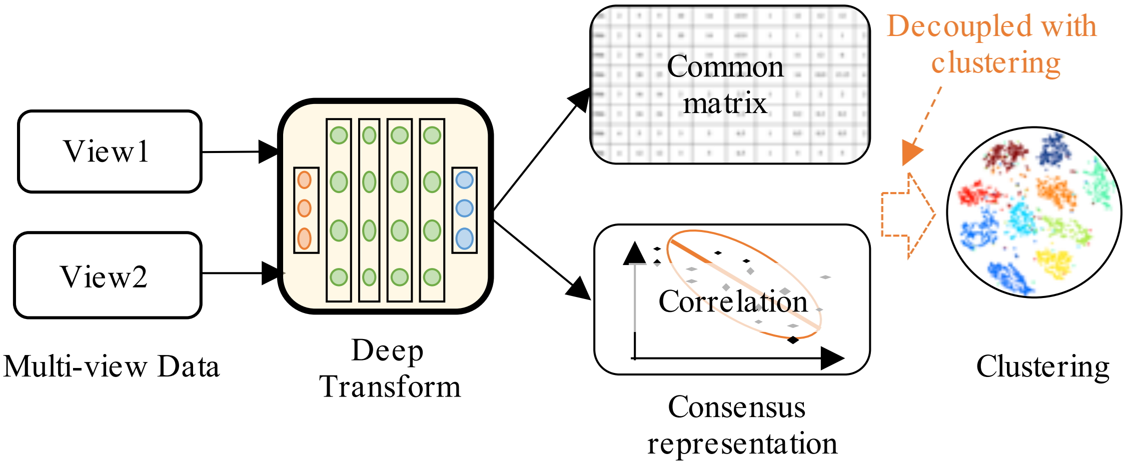


FIGURE 2.13. Unsupervised multimodal fusion: deep consensus MVC.

Deep mono-view clustering. Recent work shows that the joint optimization of feature learning and clustering can yield substantial improvement of performance. Xie *et al.* firstly proposed deep embedding clustering (DEC), in which feature representations and cluster assignments were simultaneously learned by iteratively optimizing the clustering objective [86]. Yang *et al.* proposed to learn a k-means-friendly deep representation via alternatively optimizing representation learning and k-means clustering [87]. Afterwards, more efforts were devoted to improving mono-view deep clustering by adding additional constraints (such as relative entropy loss [88], information maximization loss [89] and image triplet loss [90]) or improving network structures (such as deploying convolutional neural network [91], [92], dual autoencoder [93], variational autoencoder [94]). However, few studies have investigated deep clustering in the multi-view scenario.

2.4.3 Correlational Consensus Learning

To better understand correlation-based MVC (Correlational Consensus Learning), we briefly introduce its two major branches, linear and non-linear CCA algorithms. Given two views of data $X_1 \in \mathbf{R}^{d_1 * N}$ and $X_2 \in \mathbf{R}^{d_2 * N}$, linear CCA aims to find pairs of linear projections of views $(w_1^T X_1, w_2^T X_2)$ where the correlation is maximized:

$$\begin{aligned} &\max_{w_1, w_2} \quad corr(w_1^T X_1, w_2^T X_2) \\ =&\max_{w_1, w_2} \quad \frac{w_1^T \Sigma_{12} w_2}{\sqrt{w_1^T \Sigma_{11} w_1 w_2^T \Sigma_{22} w_2}} \\ =&\begin{cases} \max & w_1^T \Sigma_{12} w_2 \\ \text{s.t.} & w_1^T \Sigma_{11} w_1 = w_2^T \Sigma_{22} w_2 = 1 \end{cases} \end{aligned} \tag{2.4}$$

Σ_{11} and Σ_{22} denote the covariance of X_1 and X_2 respectively, while Σ_{12} denotes the cross-covariance of two views. Assuming pairs of projection vectors (w_1^i, w_2^i) are found, top k pairs of vectors can be assembled into the projection matrix $W_1 \in \mathbf{R}^{d_1 * k}$ and $W_2 \in \mathbf{R}^{d_2 * k}$ respectively. After further constraining the pair of projection uncorrelated with previous pairs, we can obtain the matrix formulation as:

$$\begin{aligned} \max \quad & tr(W_1^T \Sigma_{12} W_2) \\ \text{s.t.} \quad & W_1^T \Sigma_{11} W_1 = W_2^T \Sigma_{22} W_2 = I \\ & w_1^i \Sigma_{11} w_1^j = w_2^i \Sigma_{22} w_2^j, \quad i < j \end{aligned} \tag{2.5}$$

Non-linear extensions of CCA, such as KCCA and DCCA aim to find the non-linear projections of the views with maximized correlation, which can be expressed as:

$$\begin{aligned} &\max_{f_1, f_2} \quad corr(f_1(X_1), f_2(X_2)) \\ =&\max_{f_1, f_2} \quad \frac{cov(f_1(X_1), f_2(X_2))}{\sqrt{var(f_1(X_1)var(f_2(X_2))}} \end{aligned} \tag{2.6}$$

where $f_1(X_1)$ and $f_2(X_2)$ denotes the non-linear mapping for views X_1 and X_2.

Compared with linear CCA, non-linear CCA achieved better performance on larger datasets, as non-linear correlation is more common in real-world datasets. Deep CCA [82] gained momentum as it addressed the scalability issue of KCCA [79] by replacing the kernel functions with neural networks. DCCAE [83] and CorrNet [95] extended DCCA by further constraining DCCA model with autoencoder loss. However, both methods only focused on consensus information, neglecting view-specific information among multi-view data. To address this issue, VCCA [84] was proposed to enhance consensus representations with private variables extracted from each view using additional autoencoder networks. However, multi-view learning and clustering in VCCA were still decoupled, and hence there was no guarantee that the learned multi-view features were discriminative and clustering-relevant.

2.4.4 Deep Multimodal Interpretation

Interpretation on linear canonical correlation. In addition to seeking high predictive performance, there is a surging interest in understanding the complex cross-modal association in diagnostic decisions [96], thus to further uncover hidden disease mechanisms, facilitate understanding of the disease and build trust in statistical models. Early research has investigated methods to interpret cross-modal association in linear multimodal fusion models such as CCA. Such interpretation could be achieved through coefficients in linear embedding functions [22], canonical loadings and cross-loadings [23], graphical biplots [24], and probabilistic perspectives [25]. However, linear CCA has only limited capacity to model complex nonlinear interactions. In contrast, deep fusion models (such as deep CCA) are equipped with the strong approximation capability, but are generally more difficult to interpret. The difficulty arises from a great number of nonlinear transformation and operations in deep networks, such as nonlinear activation and kernel convolution. Although recent studies investigated the contribution of input features in deep networks towards classification decision using perturbation-based [26], [27] or propagation-based models [28]–[30], it is still an unsolved challenge of interpreting the nonlinear cross-modal association in deep multimodal fusion.

Interpretation on deep learning. Interpreting Deep Neural Network (DNN) is a challenging task. Current DNN interpretation methods revealing individual-based importance of input features can be divided into two major categories. The first category of perturbation-based methods systematically perturb the input features and then track the change of the output. Typical examples include occlusion [26] and prediction difference analysis [97]. These approaches are easy to implement, model-agnostic; however, they are often computational expensive [98]. In contrast, propagation-based methods provide computationally trackable approximations using the gradient of network outputs with respect to inputs in a single back-propagation pass. In this way, the sensitivity of the output to small perturbations in input features is conveyed. Typical examples of propagation-based methods include saliency maps [28], deep lift [30] and integrated gradients [29]. Specifically, saliency maps introduce a simple gradient method by applying a first-order linear approximation of the model to detect the sensitivity of the score. However, a major drawback of this simple gradient method is that it neglects the saturation problem [30], thus breaking the sensitivity axiom. Deep lift addresses the sensitivity issue by employing a baseline. In other words, it computes discrete gradients instead of instantaneous gradients at the inputs. However, these two methods violate the Implementation Invariance axiom [29]. Integrated gradients satisfy both axioms of sensitivity and implementation invariance. However, the aforementioned interpretation methods majorly focus on interpreting the contribution of input features in mono-modality towards the prediction results. It is still an unsolved challenge of how to decipher nonlinear cross-modal associations embedded in deep fusion networks.

2.5 Knowledge-level Fusion with Graph Models

2.5.1 Graph Basics and Construction

Network science is able to explicitly model the topological structure and provide insights into the collective behaviour of a complex system, and therefore would be a natural choice when analyzing multifocal lesions. We have witnessed its success in modeling brain network [99]–[101] and pathological cell network [46], [102], [103]. In general, in a network science analysis, firstly a graph model is constructed with vertices and edges depicting the pairwise relationship, and then followed by quantitative feature extraction with graph theory [100], [101] or deep graph embedding [104] for subsequent analysis.

Graph theory is a major stream of network science, which could be used for solving complex problems based on graph structure in the applications of medical imaging. In this section, we will introduce the basic concept, notation, followed by graph construction and quantification.

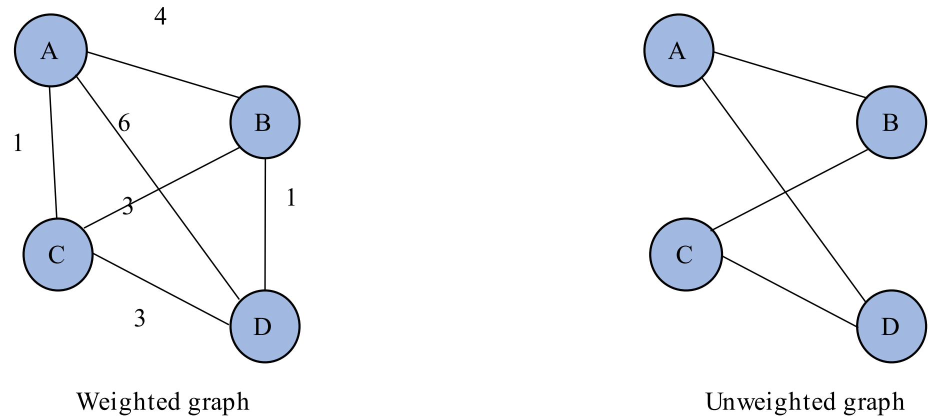


FIGURE 2.14. Graph concept: weighted graphs and unweighted graphs.

Graph concept. Conventionally, a graph is defined as $G = (V, E)$, consisted of a set of vertices $V = \{v_i | i \in [1, N_v]\}$ and their pairwise connection (edges) $E = \{(v_i, v_j) | i, j \in [1, N_v]\}$. Depending whether there is a weight associated with each edge, graphs are divided into weighted graphs and unweighted graphs, as illustrated in Figure 2.14. In a weighted graph, a weight function $W : E \longrightarrow \mathbb{R}$ is associated with each edge, representing the strength of interactions between pairwise vertices. The weights can be computed from Euclidean space, or other metric space. In an unweighted graphs, edges are either existed or not existed with no further weighting. In other words, the unweighted graph can be represented as a adjacency matrix, in which each entry of matrix has value 1 for existed edges and value 0 for non-existed edges. Overall, the weighted graph is a common topological representation in medical applications; however, it is often difficult to visualize, interpret, and analyze, and contains suspicious noisy edges [105]. Thus, proper control of the sparsity level of graph edges is required for informative topology construction, which is a fundamental challenge in network science.

Graph Construction and Edge Sparsity Control. Current graph methods control the sparsity of edges majorly based on three different criteria, as illustrated in Figure 2.15. The

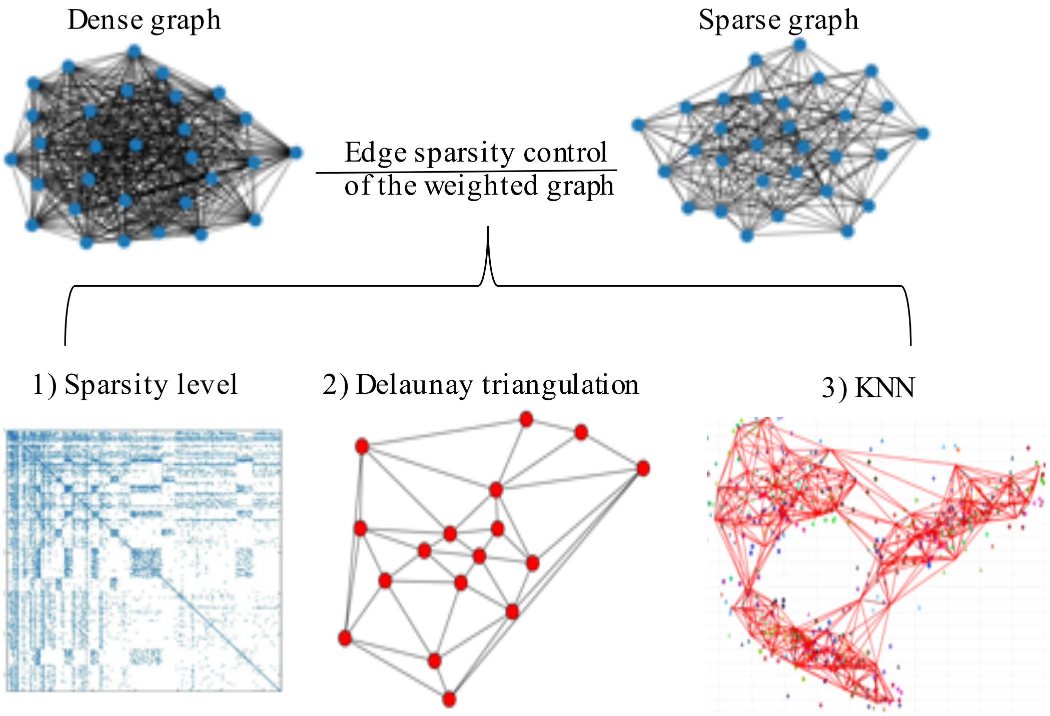

Dense graph
Sparse graph
Edge sparsity control of the weighted graph
1) Sparsity level
2) Delaunay triangulation
3) KNN

FIGURE 2.15. Graph construction and edge sparsity control.

first category of methods [47], [106] thresholded weighted graphs with user-defined sparsity values; however, the choice of the sparsity was rather arbitrary although it influenced the graph topology greatly. Early research [107] also explored the choice of sparsity by statistical tests over possible scales; however, the sparsity was instead dependent on a user-defined p-value. The second category of approaches leveraged computational geometrical structures such as Delaunay triangulation, Voronoi diagram, and minimum spanning tree [46], [102], [103] to control the edge connectivity. However, the constructed geometrical graphs still tended to involve noisy edges from a clinical perspective and required further subjective sparsity thresholding as suggested in a cell graph study [46]. The third category of methods was based on K-nearest Neighbors (KNN) [44], [104], [108], [109] to control sparsity. However, KNN-based methods also relied on a user-defined K value that controls the local relationship of vertices. Until now, there is no widely accepted criterion to define an optimal scale of the edge sparsity to binarise the weighted networks [105]. Instead of trying to determine the optimal scale, a multi-scale solution could be used to bypass this challenge, leveraging topological information at different scales with a mathematical tool named *Persistent Homology*, which will be introduced from Section 2.5.3.

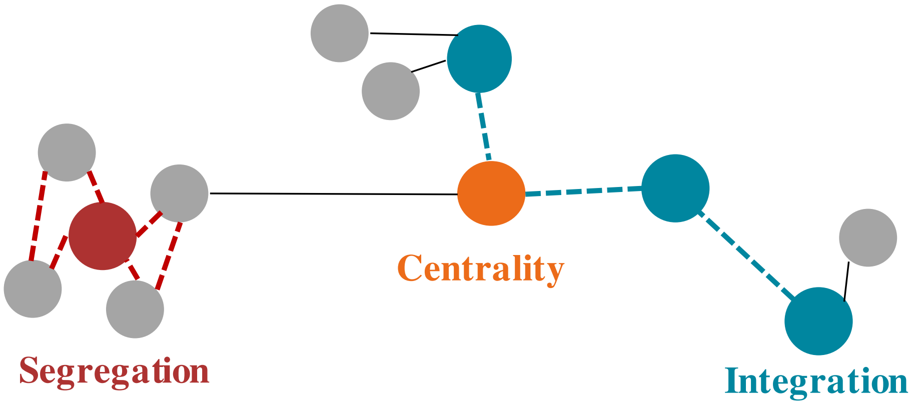


FIGURE 2.16. Graph quantification with graph theoretical measurements, including integration metrics, centrality metrics and segregation metrics.

2.5.2 Graph Quantification.

After graph construction, the next step is to extract quantitative graph features, which is commonly computed using graph metrics provided by graph theory. Broadly, the graph metrics can be divided into three groups, including segregation, integration and centrality measurements, as illustrated in Figure 2.16. Specifically, segregation refers to the graph metrics covering the processing within densely connected groups or clusters. The examples of segregation measurements include clustering coefficients, modularity, hubs. Integration refers to the graph metrics covering the information transmission between vertices. The measurements are usually based on the concept of communication paths and their path length, such as characteristic path length, and global efficiency. The third group, centrality, refers to measurements describing the importance of vertices or edges in the graph. Typical examples include degree centrality, betweenness centrality. More specific definitions of importance graph metrics include average clustering coefficients, average vertex degree, characteristic path length and degree centrality as given below.

Vertex degree is one of the most elementary and important measurements in graph theory, which are common bases for other advanced measurements. The degree of a vertex is defined as the number of edges connecting this vertex with all other adjacent vertices. In an unweighted network, the degree k_i is defined as:

$$k_i = \sum_{j \neq i} a_{ij} \tag{2.7}$$

where $a_{ij} = 1$ denotes the existed connection of vertex i and vertex j; otherwise $a_{ij} = 0$.

In segregation measurements, *local clustering coefficient* is used to measure how densely a vertex is connected [110]. It is defined as the ratio of the number of actually connected edges and possibly connected edges:

$$C_i = \frac{2t_i}{k_i(k_i - 1)} \tag{2.8}$$

where $t(i)$ is the number of the triangles through vertex i and k_i denotes the vertex degree of vertex i. As an extension, global clustering coefficient is defined as the average of local clustering coefficients of all graph vertices.

In integration measurements, *characteristic path length* often represents the efficiency of information transmission and internal structure of the graph. The shorter the path is, the quicker the information can be transmitted in the graph. The characteristic path length is a global measurement by taking average of the shortest path between individual edges. Specifically, we first define shortest path between two vertices i and j as $l_{i \longrightarrow j}$, and its length is denoted as:

$$l_{ij} = \sum_{a_s t \in l_{i \longrightarrow j}} a_{st} \tag{2.9}$$

Then, we can define the characteristic path length L of a network as the average of all shortest paths over all possible pairs of vertices in the graph:

$$L = \frac{1}{N} \sum_i l_i = \frac{1}{N} \sum_i (\frac{1}{N-1} \sum_{i \neq j} l_{ij}) \tag{2.10}$$

where the average shortest path length $l_i = \frac{1}{N-1} \sum_{i \neq j} l_{ij}$.

In centrality measurements, *degree centrality* is most commonly used measurement for centrality. It uses the degree of the vertex to represent its importance in a graph. Formally, the degree centrality $C(i)$ for certain vertex i is the degree of this vertex:

$$C(i) = k_i = \sum_{j \neq i} a_{ij} \tag{2.11}$$

For a more comprehensive review of all graph metrics for segregation, integration and centrality, please refer to the review [45], [111], [112].

2.5.3 Persistent Homology Basics

Persistent homology provides a mathematical tool to distill topological knowledge from higher-order graph structures, which are geometrically invariant across multiple scales. Such invariant features include persistent connected components and persistent holes. The persistent homology analysis usually contains three major steps, simplicial complex construction, persistent diagram generation and persistent homology transformation. In persistent homology, the first step is to construct simplicial complex, a higher-order graph structure, in a multi-scale representation using a technique named filtration. Then, topological invariant features are extracted from the multi-scale graphs and represented as persistent diagram in form of multi-set points. As multi-set points can not be directly processed by machine learning algorithms, quantification methods are proposed to bridge persistent homology analysis with machine learning algorithm using vectorization-based methods or kernel-based methods.

Multi-scale topology construction. In persistent homology, the construction of multi-scale topology includes firstly defining a simplicial complex as the topological structure and then expanding to multi-scale via filtration scheme. More specifically, a simplicial complex is a union of components including not only vertices and edges, but also triangles, tetrahedrons and higher-order polutopes [113], as illustrated in Figure 2.17. As a higher-order extension of conventional graphs, a simplicial complex is able to capture interactions beyond pairwise edges, which accordingly would provide a useful mechanism for analyzing the complicated interactions among multifocal lesions. A simplicial complex is constructed from a set of simplices. A simplex with dimension k (k-simplex) is defined as the convex hull of $k + 1$ independent vertices. For example, a 0-simplex is a point; a 1-simplex is two points connected with an edge; a 2-simplex is a filled triangle with three points, as shown in Figure 2.17. A multi-scale topology then can be constructed from simplicial complexes with filtration technique [114]. Specifically, a nested family of simplicial complexes K^r (filtered simplicial complex) can be induced for a range of scale values $r \in \mathbb{R}$ so that the complex at scale m is embedded in the complex at scale n for $m \leq n$, i.e. $K^m \subseteq K^n$.

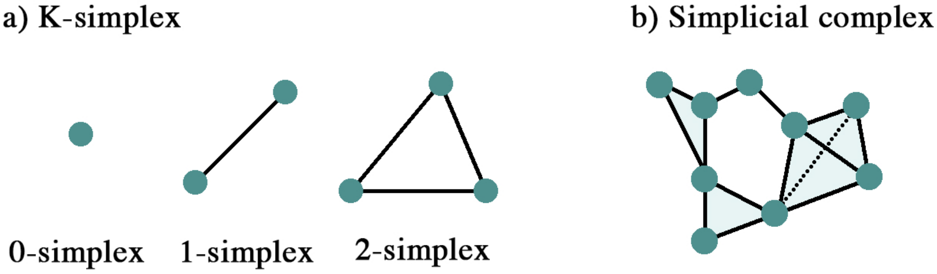


FIGURE 2.17. Higher-order graphs: k-simplex and simplicial complex.

Homology and persistent homology. Persistent homology quantifies the global geometrical structure in the filtered simplicial complex by tracking and recording topological invariants (homology) across different scales [113]. Specifically, homology is a geometrical concept measuring the shape property that is invariant under continuous deformation of a topological object such as a simplicial complex. For instance, a triangle, a square, and a circle are considered homologically equivalent to each other because they all form a single hole and therefore can be easily transformed to each other continuously. For a simplicial complex, homology summarizes the number of connected components as h_0, one-dimensional holes as h_1 and two-dimensional voids as h_2. Persistent homology computes the birth and death time of homological objects in the filtered simplicial complex across different scales r and records these birth-death pairs as a Persistence Diagram (PD) [113]. An example of persistent homology is provided in Figure 2.18.

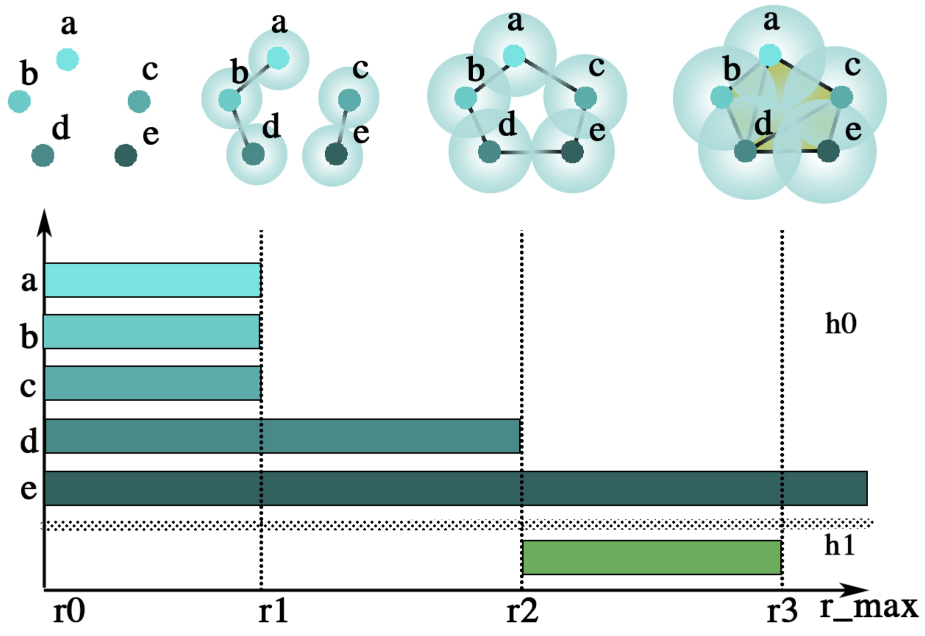


FIGURE 2.18. Global geometric measurements (persistent homology) with Peristence Barcode.

2.5.4 Advanced Topology Quantification for Machine Learning

Since the multi-set form of PD could not be processed by machine learning for classification analysis, advanced quantification methods in persistent homology were proposed to address this issue. Much research focus has been devoted into this area, which is further divided into two major branches of methods. The first branch of methods leveraged implicit similarity measures or kernel representations to quantitatively compare PDs, including bottleneck distance [115], p-Wasserstein distance [115], Persistence Scale Space kernel [116], Persistence Weighted Gaussian kernel [117] and Persistent Fisher kernel [118]. However, these methods were limited only to distance-based learning methods (such as KNN) or kernel-based algorithms (such as SVM). The second branch of methods generated explicit vector representations for PD, including Betti Curves [119], Persistence Landscape [120], and Persistence Image [121]. The typical algorithms from kernel-based methods and vectorization-based methods are introduced as below.

Notation. The aim of the designed distance, kernels or vectors is to measure the similarity of two PDs L^k and L'^k, in which k denotes k^{th} dimension of homology. l_j^k denotes the points in a PD, where j denotes j^{th} point in the PD. Each point l_j^k is represented as $[a_j^k, b_j^k]$.

Bottleneck distance. Bottleneck distance [115] is one of the early proposed metrics to quantify distance between two PDs L^k and L'^k, which is defined as below:

$$d_B(L^k, L'^k) = \inf_{\gamma} \sup_{j} ||l_j^k - \gamma(l'^k_j)||_\infty \tag{2.12}$$

where γ denotes all bijections between two PDs. In other words, the final bottleneck distance can be interpreted as $||l_j^k - l'^k_j||_\infty = \max(|a_j^k - a'^k_j|, |b_j^k - b'^k_j|)$.

p-Wasserstein distance. p-Wasserstein distance [115] between two PDs is defined as

$$d_{W,p} = \inf_{\gamma}(\sum_j ||l_j^k - \gamma(l'^k_j)||^p_\infty)^{1/p} \tag{2.13}$$

where γ still denotes all the bijetions between two PDs and p is a positive number.

Persistence Scale Space kernel. Persistence Scale Space kernel [116] was designed motivated by a heat diffusion problem with Dirichlet boundary condition. It is defined as

$$\kappa_{PSSK}(L_k, L'_k, \sigma) = \frac{1}{8\pi\sigma} \sum_{l \in L_k, l' \in L'_k} e^{-\frac{||l-l'||^2}{8\sigma}} - e^{-\frac{||l-\hat{l}'||^2}{8\sigma}} \tag{2.14}$$

where l' and $\hat{l}'$ are mirrored across diagonal.

Persistence Weighted Gaussian kernel. Persistence Weighted Gaussian kernel [117] was proposed based on kernel mean embedding for reproducing kernel Hilbert space, which is defined as

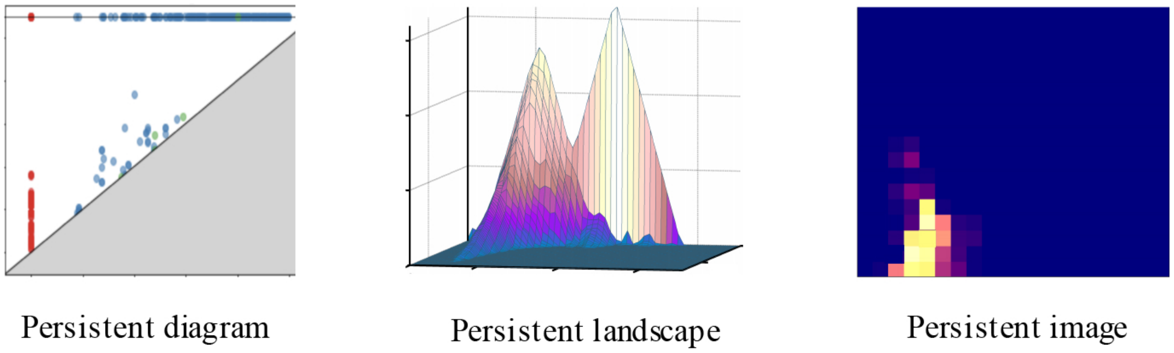
$$\kappa_{PWGK}(L_k, L'_k, \sigma) = \sum_{l \in L_k, l' \in L'_k} w_{arc}(l) w_{arc}(l') e^{-\frac{||l-l'||^2}{2\sigma^2}} \tag{2.15}$$

where $w_{arc} = \arctan(C(b_j^k - a_j^k)^p)$ with positive values C and p.

Persistent Fisher kernel. Persistent Fisher kernel [118] was defined based on Fisher information distance to preserve the geometrical properties of PD, which is defined as:

$$\kappa_{PFK}(L_k, L'_k) = e^{-t_0 d_{FIM}(\rho(x,y,L_k),\rho(x,y,L'_k))} \tag{2.16}$$

where d_{FIM} is Fisher Information Metrics (FIM) between PDs and t_0 is a positive scale number.


Persistent diagram

Persistent landscape

Persistent image

FIGURE 2.19. Advanced topology quantification with vectorization-based methods.

Different from kernel-based methods, vectorization-based methods transform each PD into a vector representation, thus could be fit into a more wide range of machine learning applications.

Betti curve. Betti curve [119] is a one-dimensional piecewise piecewise-constant function computed from rank function of persistent diagram. It is a vector of points sampled evenly in a given range.

Persistence Landscape. Persistent Landscape [120] is generated by repeating the binning process for m times and return m set of sampled values as features.

Persistence Image. Persistence Image [121] computed image-based vector representations that lived in Euclidean space by generating a weighted sum of Gaussians from PD.

2.6 Medical Applications of Multimodal Fusion

Multimodality data fusion is a promising research direction with a wide range of practical applications in the medical domain, especially to support clinical decisions, guide individualized medicine of oncological and neurological diseases. Generally, these applications can be categorized as three groups, including diagnostic classification, prognostic prediction and unsupervised biomarker identification. Prognostic prediction can be further divided into prognostic classification and regression according to the data type of prediction outcome.

2.6.1 Diagnostic Classification

Multimodal fusion can assist with automated diagnostic decisions, such as differential diagnosis, tumor malignancy prediction, tumor staging, tumor subtyping. Specifically, multi-view connectone data have been utilized to support the differential diagnosis of Autism Spectrum Disorder (ASD) and Normal Control (NC) patients [122]. PET/CT images are fused for computer-aided prediction of tumor malignancy of lymphoma [123]. For prediction of tumor staging, multi-parametric T1/T2 MRI have been fused for head-and-neck cancer [124] while PET/CT have been integrated for lung cancer [125]. Lastly, PET/CT images have also been used for tumor subtyping of breast cancer, by predicting the moleduclar characteristics such as ET, PR, Ki67 and HER4 [126].

2.6.2 Prognostic Classification

Prognostic classification refers to applications of multimodal fusion for predicting categorical labels. Such applications include prognostic marker prediction, recurrence prediction, metathesis prediction and therapy response prediction. Specifically, radiomics features and dosiomics features have been utilized for the prediction of acute-phase weight loss [127], an independent prognostic factor in lung cancer. Other examples of prognostic marker prediction include rectal toxicity prediction in prostate cancer [128] and Isocitrate Dehydrogenase (IDH) prediction in brain cancer [42]. PET/CT radiomics features have been utilized for prediction of recurrence in head-and-neck cancer [7]. For metathesis prediction, PET/CT have been used for prediction in soft-tissue sarcomas [129] and PET/MRI have been used for lung cancer [130]. Lastly, multimodal fusion can be used for predicting response of therapy such as immunotherapy [131], induction chemotherapy [132] and neoadjuvant chemotherapy [133].

2.6.3 Prognostic Regression

Prognostic regression in medical applications often specifically refers to survival regression or time-to-event analysis. Different from conventional classification or regression tasks, the interest of survival regression (time-to-event analysis) not only in whether or not an event occur, but also when the event occurs. The popularity of survival regression in medical applications is due to its ability of handling censoring data, a special type of missing data that do not experience the event during the designed follow-up time. Examples of prognostic regression in include prediction of Overall Survival (OS), Progression-free Survival (PFS), Disease-specific Survival (DSS) and Disease-free Survival (DFS). Overall survival is a major application of prognostic analysis, which focuses on the period of time after the diagnosis or treatment of a patient until its death. Alternatively, other types of prognostic analysis include glspfs (the period after treatment until the progression of the disease), DFS (the period after curative treatment until the recurrence of the disease) and DSS (the period after the treatment until the death of the patient due to specific disease). The applications of prognostic regression can be found in various of cancers including brain cancer [134], head-and-neck cancer [135], [136], breast cancer [137] and colorectal cancer [138].

2.6.4 Unsupervised Clustering

Unsupervised biomarker identification is an emerging stream of medical applications powered by multimodality fusion. Compared with the aforementioned diagnostic or prognostic predictions, it does not require large amount of annotated training data, which are sometimes

TABLE 2.3. Medical applications of multimodality fusion.

Applications	**Task**	**Multimodality**	**Disease**
Diagnostic classification	Differential diagnosis	Multi-view connectome	ASD/NC [122]
	Tumor malignancy prediction	PET/CT	Lymphoma [123]
	Tumor staging	T1_T2_MRI	Head and neck [124]
	Tumor staging	PET/CT	Lung cancer [125]
	Tumor subtyping	PET/CT	Breast cancer [126]
Prognostic classification	Prognostic marker prediction (Acute-phase weight loss)	Radiomics and dosiomics	Lung cancer [127]
	Prognostic marker prediction (Rectal toxicity)	Clinical and dosimetric features	Prostate cancer [128]
	Prognostic marker prediction (IDH)	Multi-view features	Brain cancer [42]
	Recurrence prediction	PET/CT	Head and neck [7]
	Metathesis prediction	PET/CT	Soft-tissue sarcomas [129]
	Metathesis prediction	PET/T1, PET/T2	Lung cancer [130]
	Therapy response prediction (immunotherapy)	PET/CT	Lung cancer [131]
	Therapy response prediction (induction chemotherapy)	T1_T2_MRI	Head and neck [132]
	Therapy response prediction (neoadjuvant chemotherapy)	T1, T2, DWI	Breast cancer [133]
Prognostic regression	Overall survial (OS)	T1_MRI and Clinical	Brain cancer [134]
	Overall survial (OS)	CT and clinical	Head and neck [135]
	Overall survial (OS)	MRI and clinical	Breast cancer [137]
	Progression-free Survival (PFS)	Multiparametric MRI	Head and neck [136]
	Disease-specific survival (DSS), disease-free survival (DFS), and overall survival (OS)	PET/CT	Colorectal cancer [138]
Unsupervised biomarker identification	Marker identification (disease detection)	Multi-focus regions	Age-related macular degeneration [20]
	Marker identification (tumor subtype)	Multi-view radiomic features	Breast cancer [139]
	Marker identification (survival)	Multi-view radiomic	Colorectal cancer [140]
	Marker identification (survival)	PET/CT	Lung cancer [141]
	Marker identification (diagnosis)	Hand-craft and deep features	Lung cancer [142]

unavailable, or costly to acquire. In addition, it helps to discover new predictive biomarkers in certain medical domains where no strong predictor has been found. Generally, the discovered biomarker can be used to support disease detection, tumor subtyping, survival prediction, and tumor characterization. Specifically, multi-scale regions on retinal Optical Coherence Tomography (OCT) imaging has been used for unsupervised identification of disease markers [20] of Age-related Macular Degeneration (AMD). OCT is an imaging technique that relies on low-coherence light to capture imaging using optical scattering media. Multi-view radiomics features have been utilized in marker identification of 1) tumor subtypes of breast cancer [139] and 2) survival analysis [140]. Hand-craft features and deep learning features have been used in an unsupervised Tumor characterization task of lung cancer [142].

CHAPTER 3

Feature-level Fusion: Multimodal Biomarker Mining for Trustworthy Differential Diagnosis

Multiple Sclerosis (MS) and Neuromyelitis Optica (NMO) are two common demyelinating diseases in the Central Nervous System (CNS). Misdiagnosis of these two diseases may delay the treatment, resulting in poor prognosis. MRI is routinely used in the differential diagnosis of MS and NMO, however, its specificity is limited because partial lesions in brain white matter of the two diseases share similar lesion appearance, location distribution, and signal characteristics on MRI [143]–[146]. Therefore, it is in high demand for quantitative, repeatable, and objective biomarkers for the differential diagnosis.

This chapter presents a feature-level fusion framework for diagnostic multimodal biomarker mining, from multi-parametric MRI images and clinical non-imaging factors, for the differential diagnosis of MS and NMO. A multi-level feature selection algorithm is proposed to integrate high-throughput radiomic features extracted from T2-MRI imaging, T1-MRI imaging with non-imaging clinical features as a relevant, generalizable and discriminative phenotype, based on univariate Wilcoxon filter and multivariate sequential forward selection. The diagnostic phenotype was used for constructing Multi-parametric Multivaraite Random Forest (MM-RF) model, and interpreted from both case-level and model-level using SHAP methods.

This chapter is organized as follows: Research motivation and dataset description are introduced in Section 3.1 and 3.2, respectively. A multi-level feature selection algorithm to select discriminative and robust features from multi-parametric MRI and clinical modalities is presented in Section 3.3. The experimental results and discussion are analyzed in Section 3.4 and 3.5, respectively.

3.1 Research Motivation

MS and NMO are demyelinating diseases of the CNS, which are the most common causes of neurological disability in young people [147]. In clinical practice, the differential diagnosis of these two diseases is still challenging. It is reported that around 30% of the misdiagnosed MS cases were diagnosed as NMO [143]. There are several factors contributing to the difficulty of differential diagnosis, e.g., they share overlapped features in clinical symptoms such as myelitis, optic neuritis [147], [148], and laboratory examinations (30% of the NMO patients had the same negative results of NMO immunoglobulin G as MS patients [149]). Misdiagnosis can lead to unprecise treatment and sometimes even exacerbation of the disease, as the treatment for MS differs greatly from that of NMO [150].

MRI is routinely used in the differential diagnosis of MS and NMO, however, its specificity is limited because partial lesions in brain white matter of two diseases share similar lesion appearance, location distribution, and signal characteristics on MRI [143]–[146]. In addition to similar neuroimaging characteristics, another common cause of MS misdiagnosis is the subjective visual observation and analysis, such as misinterpretation and misapplication of abnormal MRI findings as suggested by Solomon *et al.* [143]. Therefore, it is in high demand for quantitative, repeatable and objective measurements for the differential diagnosis.

Radiomics is an emerging field with a surge of interest due to its capability to extract quantitative biomedical imaging “markers” for automated objective diagnosis [35], [151], and potentially to foster individualized diagnosis. Empowered with machine learning, radiomics methodology mines the valuable underlying information that could be beyond the perception capacity of human beings and has been successfully applied for differential diagnosis of other CNS diseases [152], [153]. Although radiomic models are able to produce promising diagnostic results with higher accuracy, clinicians often find it difficult to interpret the results from machine learning models. To be clinically applicable, there is an urgent need to address the “lacking interpretability” problem [154].

In this study, we aimed to investigate a quantitative and objective MRI-based radiomics platform, equipped with individualized result interpretation, to provide clinicians with trustworthy assistance for diagnostic differentiation.

3.2 Dataset Description

3.2.1 Patient Characteristics

This study was approved by the institutional review board of Xuanwu Hospital, Capital Medical University, and written informed consent was obtained from all participants. Totally 116 participants were recruited including 78 relapsing-remitting MS and 38 NMO patients. The first cohort included patients from April 2004 to December 2004 who underwent brain scanning on 1.5T MRI (Sonata; Siemens Medical Systems, Erlangen, Germany) with an 8-channel head coil. The second cohort included patients from November 2009 to April 2014 who had brain scanning on 3T MRI (Siemens Magnetom Trio Tim System, Munich, Germany). As a proportion of patient cohorts were recruited before the introduction of the new MS and NMO criteria, our diagnosis of MS and NMO was based on the 2010 McDonald criteria, and the revised NMO diagnostic criteria, respectively [155], [156]. None of these patients who had been treated with medication within three months before the MRI was obtained. The demographic and clinical characteristics including Expanded Disability Status Scale (EDSS) score [157] and Disease Duration of the patients were recorded. Disease duration is defined by the time since the diagnosis.

Seventy-eight relapsing-remitting MS patients (mean age ± SD: 36.5 years ± 10.0), 38 NMO patients (mean age ± SD: 40.9 years ± 11.7) participated in this study. The percentages of males out of all patients were 34.6%, 18.4%, respectively. There were no significant differences in sex and age between MS and NMO patients. NMO group showed a trend towards higher EDSS score than the MS group (p = 0.005). Other demographic characteristics of the participants were provided in Table 3.1.

TABLE 3.1. Patient characteristics of MS and NMO patients for the diagnostic task. P values were calculated using Two-sample t-test for continuous variables (denoted as ^{a}), and Chi-squared test for categorical variables (denoted as ^{b}).

	3T MRI cohort		**1.5T MRI cohort**	
Characteristics	MS (n =38)	NMO (n =30)	MS (n =40)	NMO (n =8)
Age, year, mean ± SD	35.7±9.5	41.5±10.8	37.4±10.6	38.5±15.4
Female / male	25/13	23/07	26/14	8/0
EDSS, mean ± SD	3.1±1.7	3.8±1.7	2.8±1.4	4.1±1.6
Disease duration, ± SD month, mean	62.5±56.4	61.7±56.3	50.8±50.9	84.0±67.0
	Both cohorts			
Characteristics	MS (n =78)	NMO (n =38)	*P*	
Age, year, mean ± SD	36.5±10.0	40.9±11.7	0.053^a	
Female / male	51/27	31/07	0.114^b	
EDSS, mean ± SD	2.9±1.5	3.8±1.6	0.005^a	
Disease duration ± SD month, mean	56.8±54.1	66.4±58.4	0.396^a	

3.2.2 Imaging Processing and Feature Extraction

For the MRI lesion segmentation, manual segmentation is used for both MS and NMO patients since there is no public segmentation algorithm for NMO available. More specifically, segmentation of hyperintense brain lesions volume on T2 sequences was performed by a neuroradiologist with more than 9 years of experience (Jing Huang, Xuanwu Hospital) using MRIcro software[1], and validated by a senior neuroradiologist (Zhigang Qi, Xuanwu Hospital), who had more than 20 years of experience. The Volume of Interests (ROIs) delineated on

[1] https://people.cas.sc.edu/rorden/mricro/

T2 sequence were mapped to T1-MPRAGE sequence through rigid image registration to automatically obtain the corresponding VOIs from T1-MPRAGE sequence.

In terms of feature extraction, from the VOIs of both MRI sequences, we extracted 1118 quantitative radiomic features for each sequence that embraced 18 intensity, 68 texture, 344 LoG features, and 688 wavelet features [43]. LoG and wavelet filters were applied before the texture feature extraction with aims to reduce the impact of noise. After feature extraction, all features were standardized to be comparable in scale. Intensity features were calculated based on the first-order statistics of the image intensity distribution. Texture features were computed with higher-order texture matrices such as GLCM and GLSZM to quantify lesion heterogeneity. Filter-based features, including LoG and wavelet features were extracted from the filtered images to enhance specific parts of images, such as sharp edges or fine texture.

3.3 Multi-Level Feature Fusion for Diagnostic Decisions

In addition to image segmentation and feature selection, the rest of radiomics pipeline includes feature selection (phenotype building), machine learning modeling, and quantitative interpretation of results. An overview is provided in Figure 3.1.

3.3.1 Multi-Level Feature Selection

Our Multi-level Feature Selection algorithm aims to solve two key challenges in feature selection in the clinical context, including: 1) selecting relevant and discriminative features from high-dimensional small-sample multimodal data, and 2) mining robust features across MRI images with different imaging quality (e.g., MRI images with different magnetic field strength), which is often neglected by feature selection algorithms [19]. To address these two challenges, we design a Multi-level Feature Selection algorithm, composed of univariate-level and multivariate-level module, to jointly explore feature relevancy, robustness and discriminability.

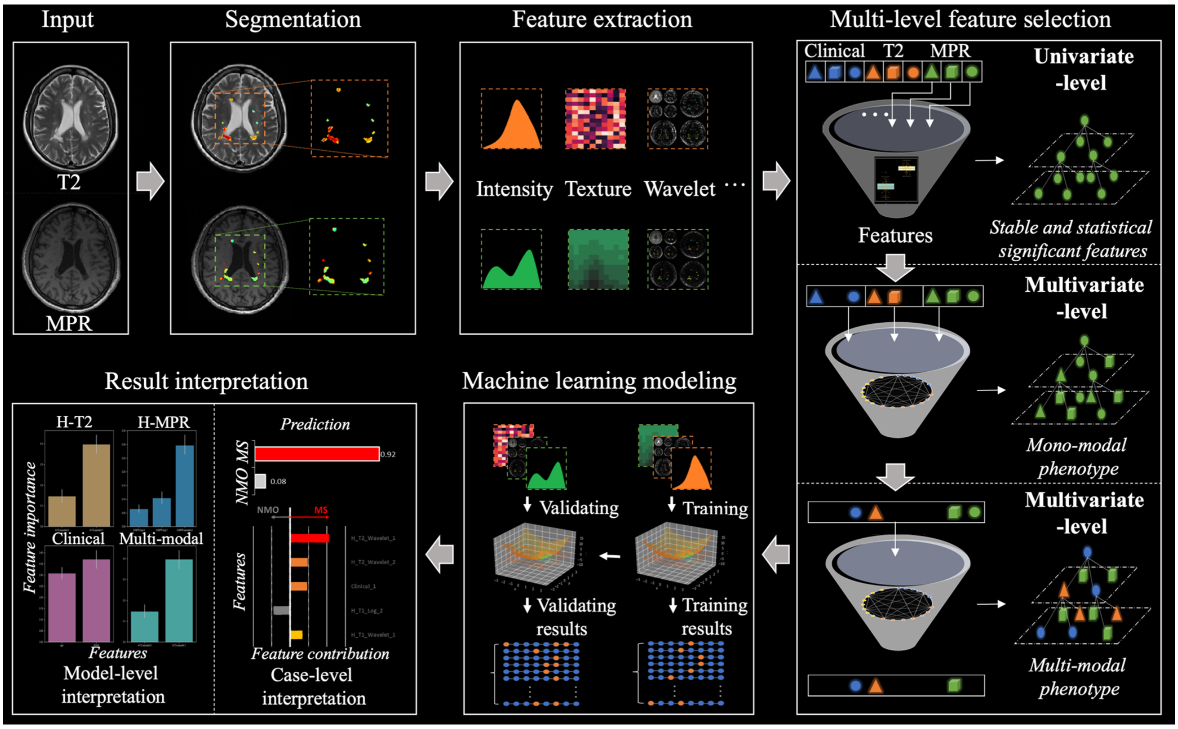


FIGURE 3.1. Flowchart of radiomics pipeline for differential diagnosis of MS and NMO.

In univariate-level module, we design a statistical filter scheme on the basis of Wilcoxon Rank-sum test to simultaneously select (1) robust features by testing the statistical consistency across MRI with different image quality and (2) relevant features through calculating statistical relevancy towards the outcome. Specifically, robust features across different magnet strength of MRI scanners (1.5T and 3T) were first selected with Wilcoxon [158]. Then, discriminative features were selected by assessing whether there was a significant distribution difference between MS patients and NMO patients via Wilcoxon test [159].

In multivariate-level module, we propose a pyramid searching structure to first exploit intra-modal feature relationships and then explore inter-modality relationships. This pyramid searching scheme boosts feature discriminability and mining efficiency. In contrast, the conventional feature selection often uses a flattened search space by concatenating all features for feature selection. In specific, Random Forest-based sequential forward selection (RF-SFS) was firstly employed to select discriminative features and construct preliminary phenotypes from T2, T1-MPRAGE and clinical features separately [160]. Then, a multi-parametric phenotype was constructed by further applying RF-SFS to a fused feature set of three preliminary phenotypes.

Algorithm 1 Multi-level feature selection

```
Input: Training dataset $X \in \mathbb{R}^{p*q}$, max number of features d
Output: Selected feature set $F^m$
Initialize:
Split Train dateset into $X_{ms} \in \mathbb{R}^{m*q}$ and $X_{nmo} \in \mathbb{R}^{n*q}$.
Split $X_{ms}$ into $X_{ms1.5}$ and $X_{ms3}$.
Split $X_{nmo}$ into $X_{nmo1.5}$ and $X_{nmo3}$.
$F^u \leftarrow \emptyset$, $F^m \leftarrow \emptyset$
Univariate selection of robust features:
for feature $f^i \in$ all features $F^q$ do
    $s_{ms}, p_{ms} \leftarrow wilcoxon(X^i_{ms1.5}, X^i_{ms3})$
    $s_{nmo}, p_{nmo} \leftarrow wilcoxon(X^i_{nmo1.5}, X^i_{nmo3})$
    if $p_{ms} > 0.05$ and $p_{nmo} > 0.05$ then
        $F^u = F^u \cup f^i$
    end if
end for
Univariate selection of relevant features:
for feature $f^i \in F^u$ do
    $s, p \leftarrow wilcoxon(X^i_{ms}, X^i_{nmo})$
    if $p < 0.05$ then
        $F^u = F^u \cup f^i$
    end if
end for
Multivariate selection of discriminative features:
while $dim(F^m) <= d$ do
    $f^{j+} \leftarrow \underset{f^j \in F^u}{\arg\max} J(F^m \cup f^j)$
    $F^m \leftarrow F^m \cup f^j$
end while
```

The mathematical details of multi-level feature selection were summarized in Algorithm 1, which is composed of univariate selection of robust features, relevant features and multivariate selection of discriminative features. The feature selection is performed on the training data $X \in \mathbb{R}^{p*q}$ where p and q are the number of samples and features respectively. A parameter d is required as the max number of features to select. The specific selection process can be divided into three separated stages:

- In univariate selection of robust features, we used Wilcoxon rank-sum test to select features that were robust across 1.5T MRI and 3T MRI in both MS cohort ($X_{ms1.5}$ and X_{ms3}) and NMO cohort ($X_{nmo1.5}$ and X_{nmo3}).
- In univariate selection of relevant features, we performed Wilcoxon rank-sum test to choose features with significant statistical differences between MS cohort (X_{ms}) and NMO cohort (X_{nmo}). The cut-off p-value was set to 0.05.
- With selected features F^u from univariate analysis, we further applied sequential forward selection in multivariate analysis to obtain the final feature set F^m.

3.3.2 Multimodal Model Construction and Validation

To compare the diagnostic performance of preliminary phenotypes of T2, T1-MPRAGE and clinical and the multi-parametric phenotype, three preliminary Multivariate Random Forest models and Multi-parametric Multivariate Random Forest model (MM-RF) were constructed based on the corresponding phenotype respectively. To handle data imbalance, balanced bootstrap mechanism and balanced weighting [161] were incorporated into the Random Forest model. Common evaluation metrics for imbalanced datasets were used for assessing the diagnostic performance of models, including ROC_AUC, accuracy, sensitivity, and specificity. The stability of diagnostic performance was assessed with the mean of Relative Standard Deviation (RSD) of ROC_AUC. The lower the ROC_AUC value, the higher the stability.

To validate the feature selection, our Multi-level Feature Selection algorithm was compared with 8 state-of-the-art feature selection algorithms that were commonly used in radiomics studies [162]–[164]. These algorithms included three Filter Selection methods (Wilcoxon filter [165], Anova filter [166], mRMR filter [167]), two Wrapper Selection methods (Random Forest (RF) wrapper [168], and SFS wrapper [59]), and three Embedded Selection methods (Lasso [66], ElasticNet [169], RFE_SVM [67]).

To rigorously validate the diagnostic performance of the imaging phenotypes, both 10-fold cross-validation on the training set and independent validation on the testing set were computed. From a total of 116 patients, 86 patients were randomly selected to form the training set, while the rest 30 patients was used for independent testing. Bootstrapping with 1000 times resampling was used in the independent validation.

All statistical analysis was two-sided, with the significance level of 0.05. Multi-level statistical analysis was performed with “scipy”, “sklearn”, “mlxtend”, “mifs”, “imblearn” modules in Python 3.6. Correlation analysis was performed in R 3.5.1. Bootstrapping with 1000 times resampling was used in the independent validation.

3.3.3 Quantitative Interpretation of Results

Lack of interpretability is a key challenge as the basis for for trustworthy decision making [170]. To provide quantitative interpretation, we utilized SHAP method [171] to analyse the differential decisions of our MM-RF model at both individual-level and model-level. SHAP is an abbreviation of Shapley Additive exPlanations by Lundberg and Lee *et al.* designed to explain individual predictions in machine learning. The individual-level interpretation explains the output of an individual prediction by visualizing the important features in the phenotype and unveiling their importance for discrimination decisions. The model-level interpretation computed the average feature importance across all patients and revealed the relationship between the feature value and its importance.

3.4 Experimental Results

3.4.1 Clinical Visual Analysis

Three neuroradiologists with 5, 7, and 10 years of MRI reading experience were involved in the visual assessment of the brain lesion and differential classification of MS and NMO patients. The assessment was based on T1-MPRAGE and T2 MRI sequences, while the clinical data (age, sex, disease duration and EDSS score) was allowed to refer during the assessment. Each neuroradiologist provided a diagnostic result for each patient based on their own clinical experience. In case of any discrepancy, it shall be jointly reviewed to reach an agreement. The assessments such as Area Under the Curve (AUC), diagnosis accuracy, sensitivity, and specificity were calculated.

The AUC of the visual analysis was 0.683 with 95% Confidence Interval (CI) 0.571-0.789. The visual analysis successfully diagnosed 61 out of 86 patients, with accuracy reaching 0.709 (95% CI: 0.616-0.802). In the misdiagnosed 25 cases, 10 NMO patients were misdiagnosed as MS while 15 MS patients were misdiagnosed as NMO. Its sensitivity and specificity were 0.615 and 0.750 respectively.

3.4.2 Radiomic Feature Selection and Phenotype Construction

Figure 3.2 demonstrates the differentiation ability of top univariately selected features and clinical factors. It shows that T2 and MPR radiomics features generally achieved higher AUC than clinical features at univariate level. Figure 3.3 illustrates the process of multivariate SFS feature selection, which shows multivariate imaging phenotypes had higher discriminability compared with clinical features. The multi-parametric phenotype was established with three T2, four T1-MPRAGE, and one clinical feature. These eight features and their corresponding feature identification number (id) were H-T2-waveletHHL-glcm-Idn (116), H-T2-log2-glcm-Autocorrelation (18), H-T2-waveletLLH-glcm-JE (551), H-MPR-waveletLHL-glszm-GLNU (500), H-MPR-log4-gldm-SDLGLE (225), H- MPR-log3-firstorder-Median (95), H-MPR-log5-glcm-Idmn (287), and EDSS.

In univariate feature selection, 450 T2 and 117 T1-MPRAGE robust features across 1.5T and 3T MR images were firstly selected from 2236 radiomic features. After that, 313 T2 and 86 T1-MPRAGE discriminative features were identified from robust features for differentiation of MS and NMO. Seven T2 features, four T1-MPRAGE features, and one clinical feature were selected from 313 T2, 86 T1-MPRAGE features and four clinical features respectively, to form

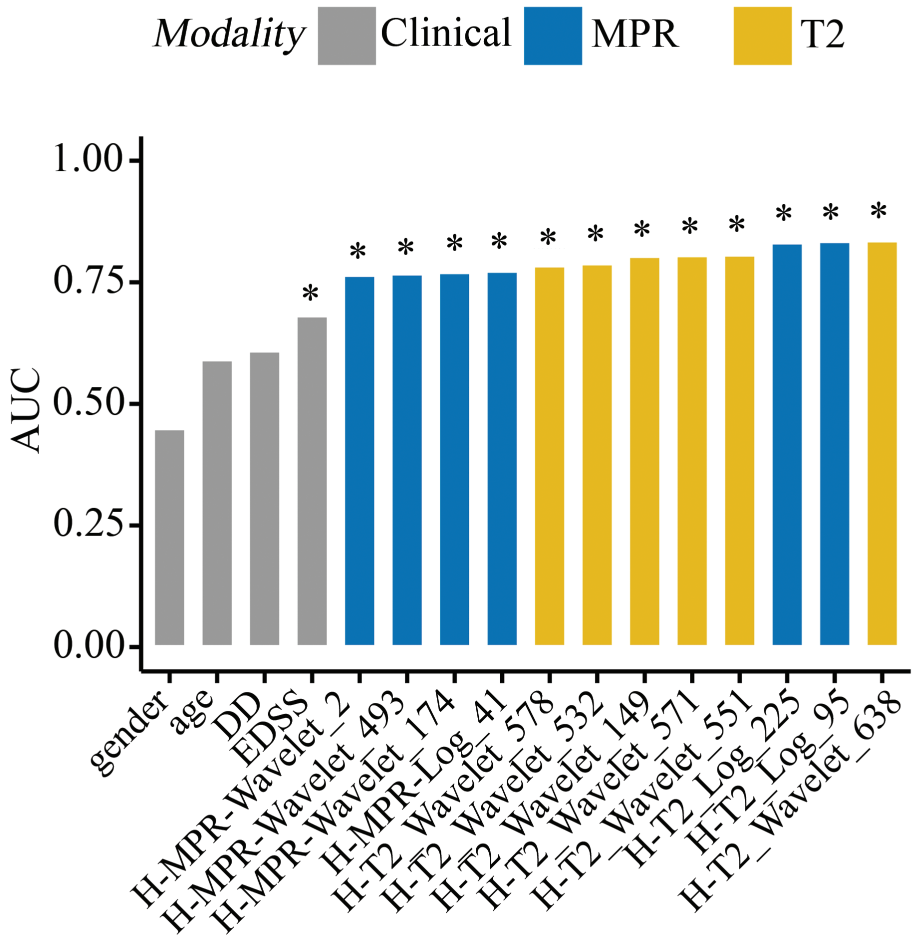


FIGURE 3.2. Results of univariate feature selection and analysis. Top six T2 and T1-MPRAGE features and four clinical factors were selected and ranked according to AUC. The asterisk (*) represents statistical significance ($p<0.05$) with Wilcoxon test.

the corresponding preliminary phenotypes. From 12 fused features from T2, T1-MPRAGE and clinical phenotypes, the multi-parametric phenotype was established with three T2, four T1-MPRAGE and one clinical feature.

3.4.3 Discriminability of Multi-parametric Phenotype

The multi-parametric phenotype was evaluated with both 10-fold cross-validation and independent testing. In cross-validation, the multi-parametric phenotype achieved AUC 0.826

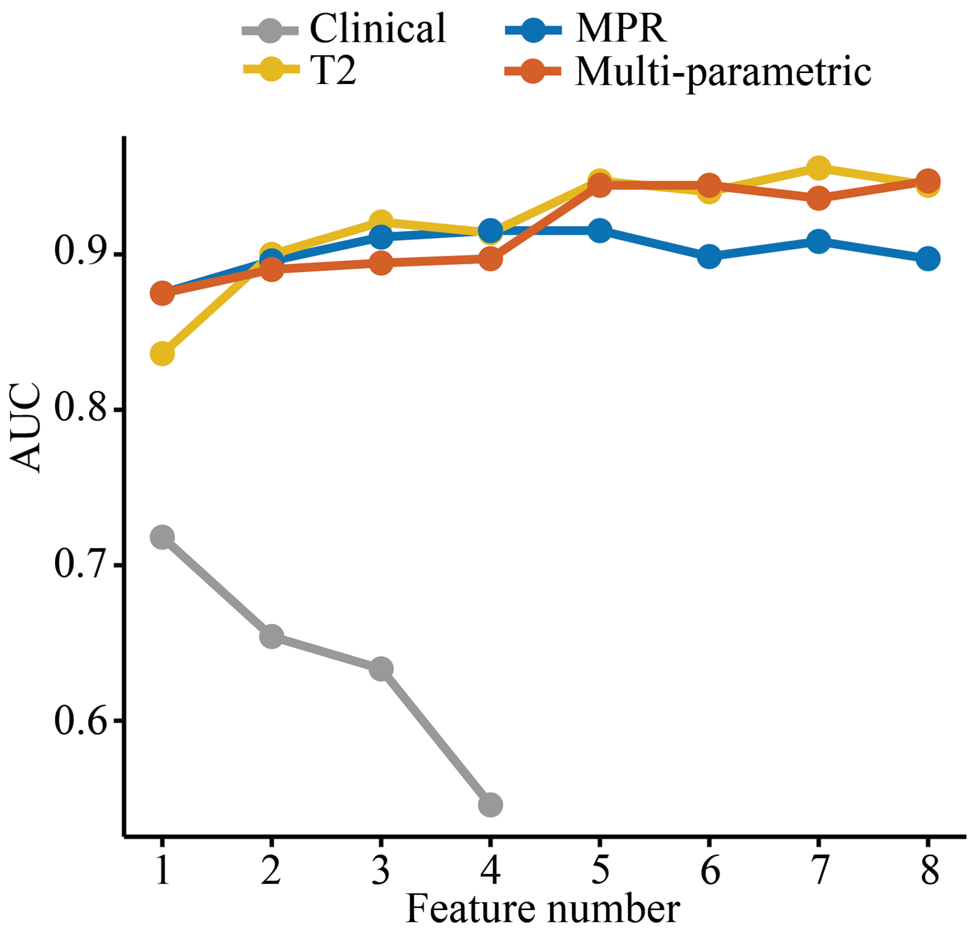


FIGURE 3.3. Results of multivariate feature selection. The algorithm selected a subset of features having the highest AUC. The arrows indicate the stopping points where the highest AUC was achieved.

(95% CI: 0.732-0.912), which was significantly higher than that of visual analysis (p = 0.016). The diagnostic accuracy was 0.849 (95% CI: 0.767-0.919), higher than that of visual analysis (p = 0.008). Its sensitivity and specificity were 0.769 and 0.883 respectively.

In the independent testing, the multi-parametric phenotype based on Random Forest model achieved AUC 0.902 ± 0.027, which was higher than the performance of preliminary T2, T1-MPRAGE and clinical phenotypes (Figure 3.4). The diagnostic AUC of T2, T1-MPRAGE and clinical phenotype was 0.852 ± 0.053, 0.880 ± 0.033 and 0.573 ± 0.055 respectively.

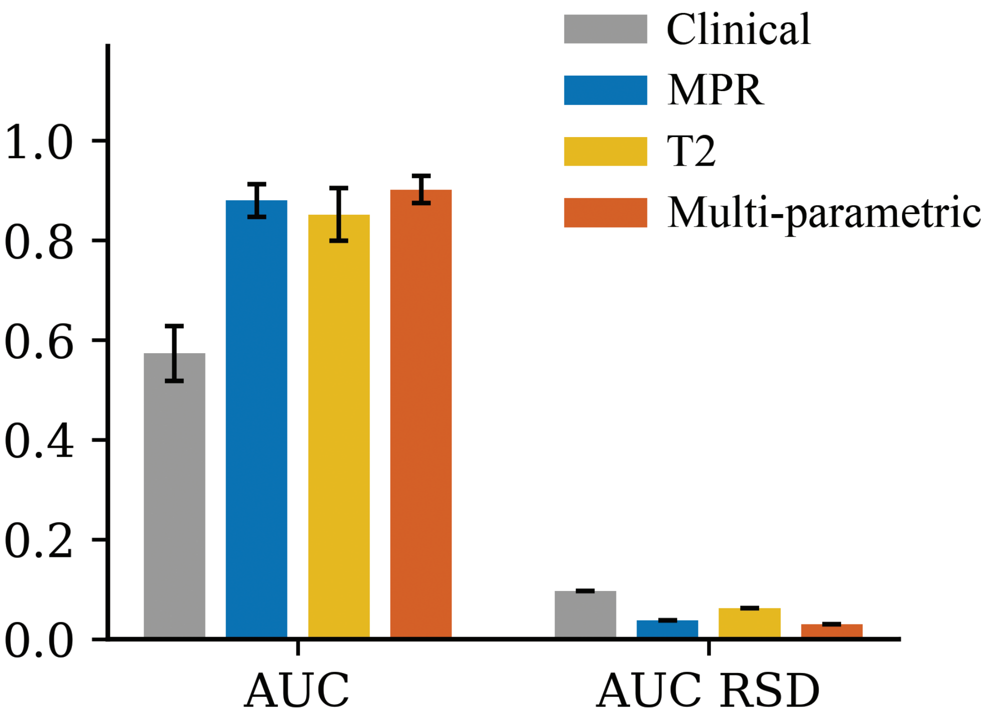


FIGURE 3.4. Comparison of diagnostic performance of T2, T1-MPRAGE, clinical phenotypes and the multi-parametric phenotype in the independent testing.

Figure 3.4 also illustrates that multi-parametric phenotype achieved better stability of diagnostic performance. Other assessments of the multi-parametric phenotype such as diagnostic accuracy, sensitivity, and specificity were 0.871 ± 0.044, 0.873 ± 0.083 and 0.869 ± 0.051, respectively, as reported in Table 3.2. To assess the impact of 3T and 1.5T MRI, a further experiment showed that high diagnostic accuracy were achieved by both 3T (0.856 ± 0.046) and 1.5T (0.976 ± 0.074) cohort ($p < 0.05$).

3.4.4 Evaluation of Multi-level Feature Selection Compared with 8 state-of-the-art Methods

Table 2 shows the performance comparison of our Multi-level Feature Selection methods with 8 state-of-the-art methods. From a combined feature pool of T2, T1-MPR and clinical features,

TABLE 3.2. Diagnostic performance of our method compared with 8 State-of-the-art (SOTA) feature selection algorithms. Abbreviations: AUC = area under the curve; Anova = Analysis of variance; mRMR = Maximum Relevance Minimum Redundancy; RF = Random Forest; SFS = Sequential Forward Selection; Lasso = Least Absolute Shrinkage and Selection Operator; RFE = Recursive Feature Elimination.

Category	**Method**	**Roc_auc**	**Sensitivity**	**Specificity**	**Accuracy**
Filter	Wilcoxon [165]	0.625 ± 0.120	0.604 ± 0.145	0.585 ± 0.137	0.593 ± 0.104
Filter	Anova [166]	0.879 ± 0.037	0.883 ± 0.042	0.726 ± 0.067	0.789 ± 0.036
Filter	mRMR [167]	0.846 ± 0.041	0.823 ± 0.091	0.754 ± 0.076	0.782 ± 0.050
Wrapper	RF [168]	0.846 ± 0.041	0.828 ± 0.091	0.750 ± 0.075	0.781 ± 0.048
Wrapper	SFS [59]	0.858 ± 0.038	0.782 ± 0.073	0.829 ± 0.121	0.810 ± 0.067
Embedded	Lasso [66]	0.873 ± 0.050	0.884 ± 0.046	0.737 ± 0.087	0.796 ± 0.056
Embedded	ElasticNet [169]	0.850 ± 0.064	0.868 ± 0.070	0.719 ± 0.135	0.779 ± 0.086
Embedded	RFE [67]	0.814 ± 0.072	0.846 ± 0.058	0.612 ± 0.141	0.705 ± 0.090
Ours	Clinical	0.573 ± 0.055	0.716 ± 0.190	0.446 ± 0.179	0.554 ± 0.070
Ours	T2 MRI	0.852 ± 0.053	**0.887** ± **0.067**	0.733 ± 0.057	0.795 ± 0.045
Ours	T1-MPR MRI	0.880 ± 0.033	0.798 ± 0.102	0.859 ± 0.073	0.835 ± 0.060
Ours	multimodality	**0.902** ± **0.027**	0.873 ± 0.083	**0.869** ± **0.051**	**0.871** ± **0.044**

these comparison methods selected at most 8 features as in our method, and were evaluated with the same Random Forest model as ours. The experimental results in Table 2 showed that our Multi-Level Feature Selection algorithm outperformed these SOTA methods in comparison. Our multiparametric phenotype achieved highest AUC 0.902 ± 0.027, followed by our MPR phenotype (AUC 0.880 ± 0.033) and Anova Filter (AUC 0.879 ± 0.037).

3.4.5 Case Studies for Individual-level Interpretation

As illustrated in Figure 3.5-3.6, the two selected cases included (a) an MS case and (b) an NMO case, whose lesions were difficult to differentiate due to similar lesion location and signal characteristics. With the extracted phenotype from VOIs in the MR images, our MM-RF classified the cases correctly with 89% confidence for MS case and NMO case with 86% confidence respectively. For the MS case, our case-level interpretation revealed that H-MPR-log3-firstorder-Median, H-MPR-waveletLHL-glszm-GLNU, and EDSS were the three most significant contributors for accurate classification, with 29.86%, 27.61% and 21.07%

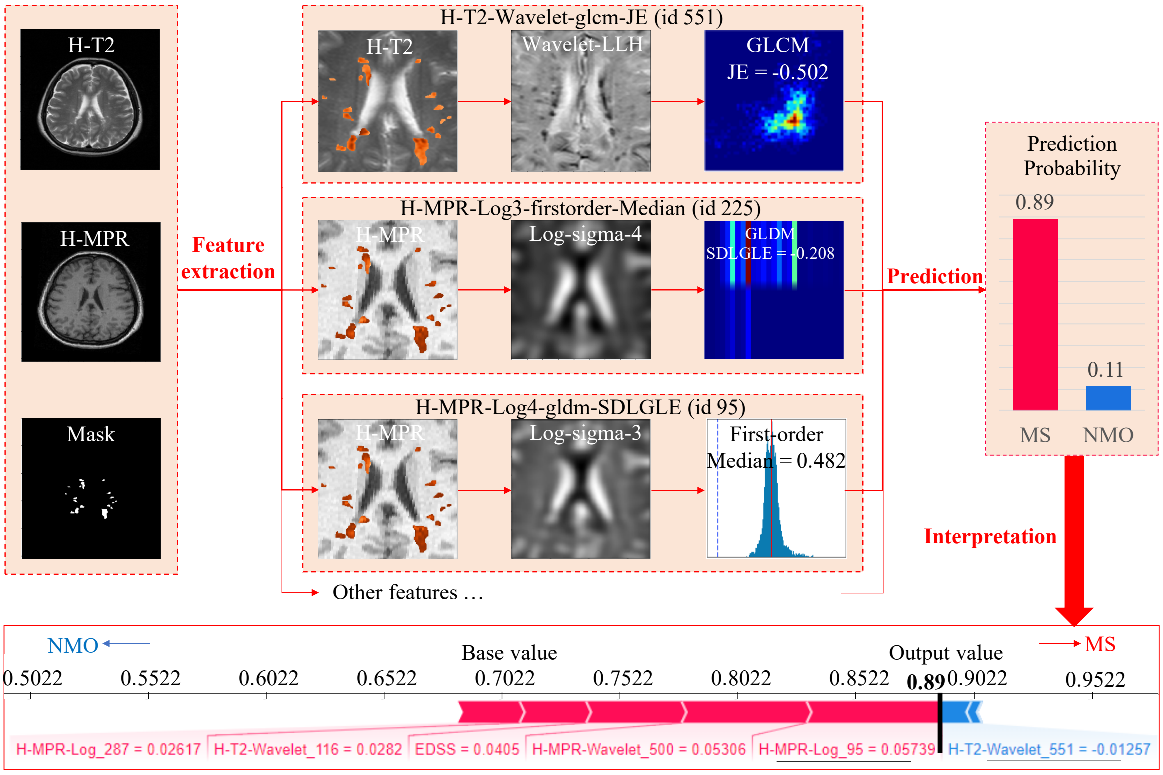


FIGURE 3.5. Results of individual-level interpretation (MS). For each case, visualization of three key radiomic features were provided. The classification results were computed with Random Forest model. Lastly, the classification results were explained by revealing feature contribution.

contribution respectively. As a contrast, for the NMO case, three T1-MPRAGE features (H-MPR-log3-firstorder-Median, H- MPR-log4-gldm-SDLGLE, H-MPR-waveletLHL-glszm-GLNU) were three major contributors, contributing 25.06%, 24.78%, 17.46% towards the correct decision.

3.4.6 Model-level Result Interpretation

Model-level interpretation investigated the feature importance and the relationship between feature value and its importance from the perspective of all patients. Figure 3.7A shows case-level interpretation results of all patients in one graph, which visualizes how feature contributions differ for diverse cases. Of all eight features, H-MPR-log4-gldm-SDLGLE, H-MPR-log3-firstorder-Median, and H-T2-waveletLLH-glcm-JE were the top three important

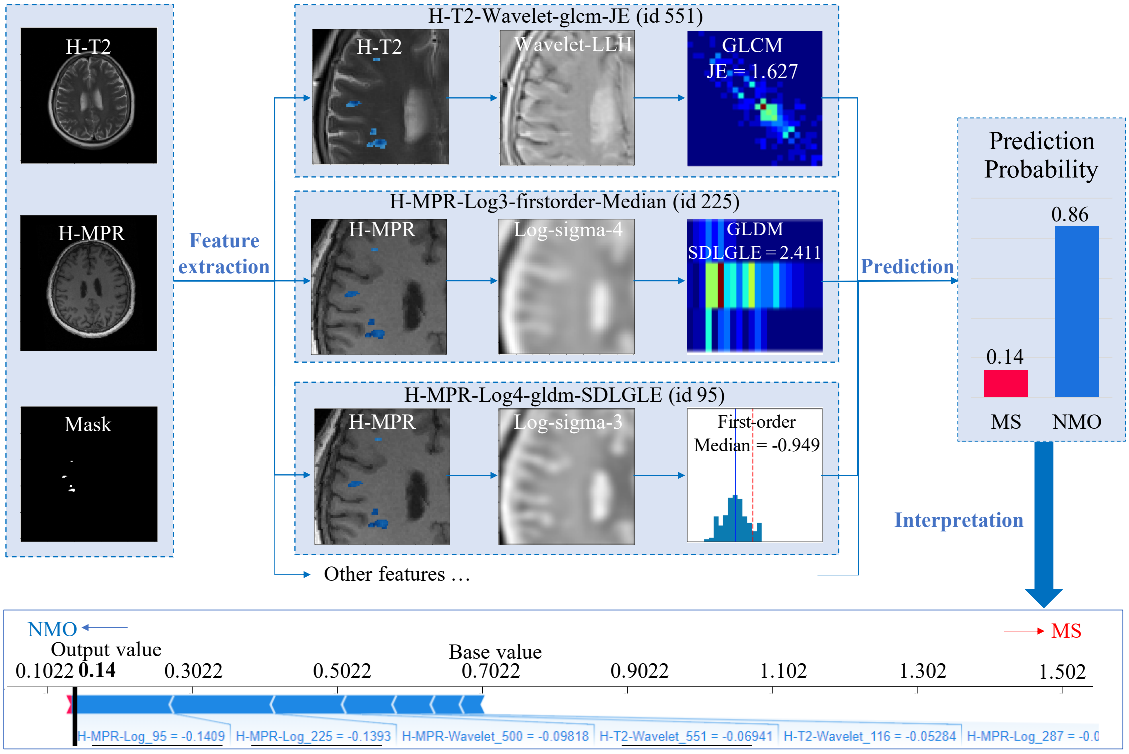


FIGURE 3.6. Results of individual-level interpretation (NMO).

features in the model decision making, as shown in Figure 3.7B and 4C. In terms of relationship between the feature value and its importance, there was a negative linear relationship for H-T2-waveletLLH-glcm-JE (Figure 3.7D) and H-MPR-log4-gldm-SDLGLE (Figure 3.7E), and a positive linear relationship for H-MPR-log3-firstorder-Median (Figure 3.7F).

3.4.7 Association of Selected Radiomic Features with Clinical Variables

As shown in Figure 3.8, sex was found significantly negatively correlated with H-MPR-log4-gldm-SDLGLE (p = 0.008), while positively correlated with H-T2-log2-glcm-Autocorrelation (p = 0.035). EDSS scores was significantly positively correlated with H-T2-waveletLLH-glcm-JE (p = 0.036). Age was significantly negatively correlated with H-MPR-waveletLHL-glszm-GLNU (p = 0.007), and H-T2-waveletHHL-glcm-Idn (p = 0.010), and H-T2-log2-glcm-Autocoorelation (p = 0.035).

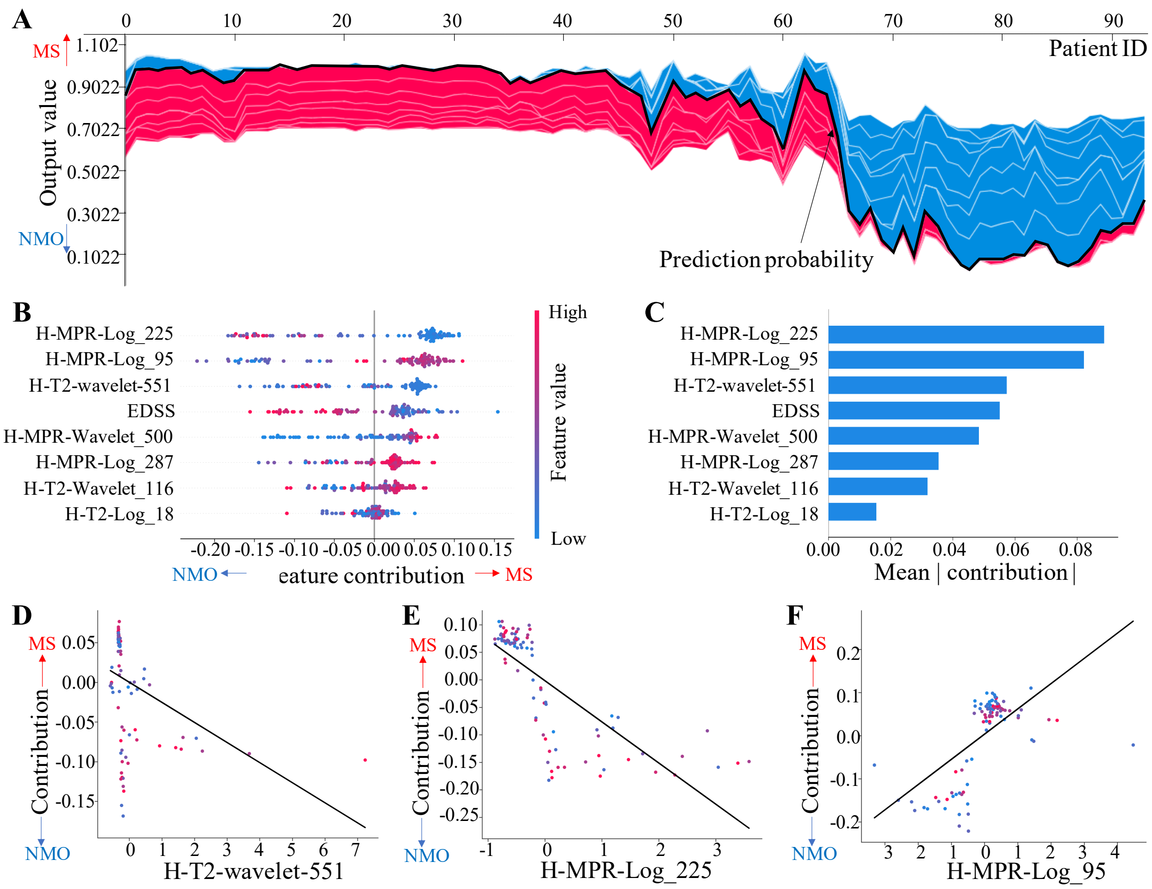


FIGURE 3.7. Results of model-level interpretation. (A) Visualization of feature contribution for all individual cases. Each vertical line corresponds to interpretation for individual diagnosis. Red represents a diagnosis of MS, blue for NMO (B) Summary plot of feature contribution for all individual cases (C) Mean contribution of features in the multi-parametric phenotype (D-F) Relationship between feature value and feature importance. The straight lines were obtained through curve fitting through linear regression.

3.5 Discussion

In this research, we extracted the imaging phenotype from multi-parametric MRI sequences with the machine learning framework for automated differentiating MS from NMO to provide an additional reference for timely differential diagnostic decision making. The major findings of this study include: (1) our multi-parametric phenotype was able to achieve high differential diagnostic performance, generalizability and robustness, mined by our designed Multi-level Feature Selection algorithm; (2) our radiomics platform provided individualized differential

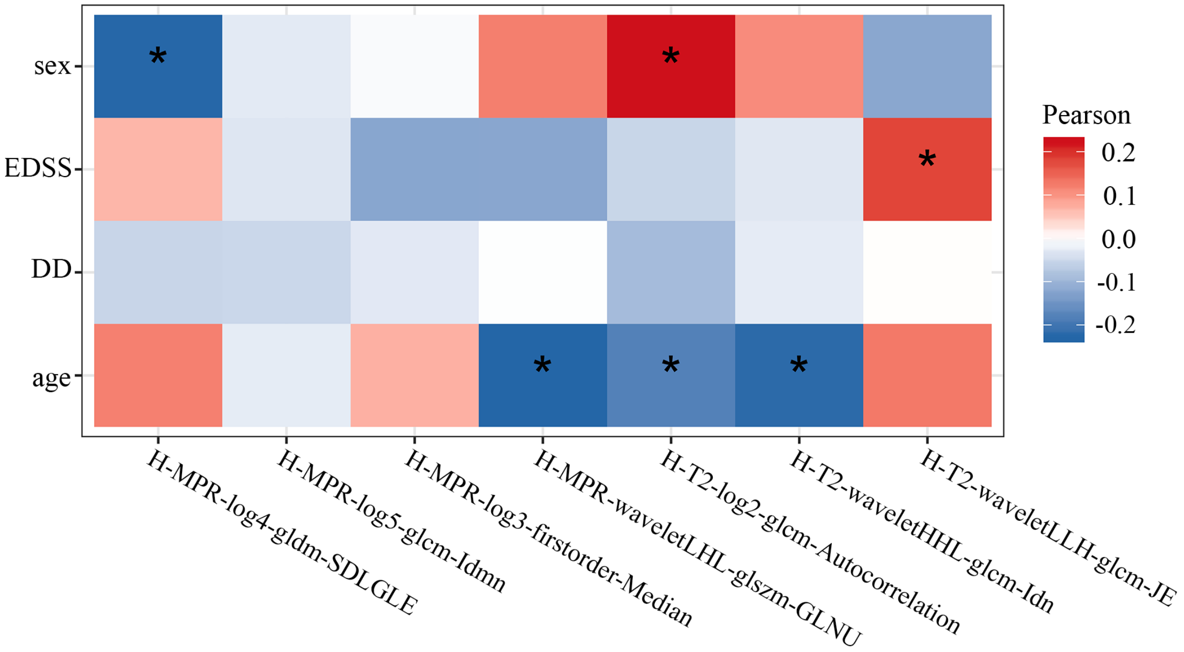


FIGURE 3.8. Results of correlation analysis. Correlation matrix was computed for seven radiomic features in the multi-parametric phenotype and four clinical features (age, sex, DD, EDSS). Red and blue bars show the positive and negtive correlation respectively. The asterisk (*) represents statistical significance.

diagnosis and interpretation which is illustrated with a case study; and (3) the correlation between radiomic and clinical features was revealed to enhance trust in radiomic features.

The first finding of our study is that the multi-parametric phenotype demonstrated high differential diagnostic performance, which statistically outperformed visual analysis in terms of AUC (0.826 vs. 0.683, p = 0.016), and the diagnosis accuracy (0.849 vs. 0.709, p = 0.008) in 10-fold cross-validation. The accuracy of clinical visual analysis in our study complied with the studies [172]–[176] with the reported accuracy ranging from 0.573 to 0.739. In this study, doctors misdiagnosed about 25% of patients with MS as NMO, similar to the previous study [143], which justified the machine learning model can provide valuable assistance for clinical decisions. Remarkably, the multi-parametric phenotype demonstrated the highest discriminative ability (0.902 ± 0.027) in the independent testing, outperforming the discriminative performance of the T2 (AUC 0.852), T1-MPRAGE (AUC 0.880) and clinical phenotypes (AUC 0.573). It indicates that the multi-parametric phenotype successfully fused pathological characteristics by fusion of information about edema, demyelination in T2 images, axonal damage in T1-MPRAGE images [177] and clinical information. This finding is consistent with previous studies in the differential diagnosis of brain tumors where the model embracing MRI-based radiomic features and clinical features can achieve the highest classification accuracy [178]. The present study, for the first time, constructs a multi-parametric phenotype including T2, T1-MPRAGE and clinical information for differentiating MS from NMO.

Our Multi-level Feature Selection algorithm outperforms SOTA feature selection methods because the proposed algorithm comprehensively considers 1) feature robustness across MRI images with different magnetic fields, 2) feature relevancy towards the outcome, and 3) intra-modal and inter-modal feature discriminability. Comparatively, Filter methods [165]–[167] select features using the defined feature relevancy such as mutual information; however, these methods may not take account of the interaction with the learning algorithm, and hence feature discriminability might not be optimized [55]. To address this issue, both Wrapper [59], [168] and Embedded [66], [67], [169] methods involve learning algorithms to assess the predictive performance of feature combinations. However, these methods may be prone to overfitting due to the dependency of learning algorithm [179]. In contrast, our univariate-level selection selects robust and relevant features based on Wilcoxon testing, which addresses feature generalizability issue across different MRI imaging qualities (as analysed in [19]) and facilitates to alleviate the risk of overfitting. Further, our multivariate-level selection boosts feature discriminability by exploiting intra-modal feature interaction and inter-modality interaction using a pyramid search structure. Due to the reduced risk of overfitting and boosted feature discriminability, our algorithm outperforms SOTA methods.

Secondly, individual-level interpretation is provided to articulate machine learning-based decision making for an individual patient and thus to facilitate the trustworthy and individualized differential diagnosis. It is achieved by graphical visualization of important features and unveiling of the quantitative contribution of features in the machine learning models to facilitate understanding on both radiomic features and model decisions, as illustrated in case studies (Figure 3.5-3.6). Specifically, for the case in Figure 3.5, this patient was correctly classified as MS with 89% confidence by our phenotype, in which the top three important features were two T1-MPRAGE features and EDSS. And for the case in Figure 3.6, three T1-MPRAGE radiomic features were major contributors. Interpretability enables doctors to gain insight why those diagnosis is made, thus assisting clinicians to provide precise differential diagnosis [136], [180].

Furthermore, the trust in radiomic features can be enhanced by revealing its connection with the clinical information. Mild correlations were observed between the radiomic features and clinical features (age, sex, and EDSS). Interestingly, we found that one T2 feature was related to EDSS (Figure 3.8). The reason underlying the correlation between EDSS and T2, T1-MPRAGE features might be that EDSS was found correlated with lesion load and brain atrophy [181], [182], while T2 and T1-MPRAGE images could also reflect the information of lesion and brain structure respectively. As a result, we may potentially use radiomic features to objectively and conveniently assist EDSS in evaluating treatment and disability management in the future. Although the above assumptions are preliminary, our study provides a perspective for understanding the clinical significance of radiomic features, in response to the urgent clinical need [183].

We suggest that our multi-parametric phenotype may serve as an objective, quantitative tool to assist the clinical differential diagnosis of MS and NMO. Compared with current diagnostic criteria of these two diseases using MR, our phenotype holds advantages in three aspects: (1) Instead of diagnosis by naked eye based on vague clinical experience, our multi-parametric phenotype provided a quantitative solution by feature extraction of medical images, thus help complement and clarify the current diagnostic criteria. (2) The phenotype reduces inter-observer variety and subjectivity because the whole system is highly automated. (3) The current model was able to achieve high diagnostic accuracy with conventional MR sequence, which is simple and operable in the clinical practice [184].

3.6 Chapter Summary

In this chapter, we present a feature-level fusion framework, based on a Multi-level Feature Selection algorithm, to integrate multiparametric MRI images and clinical non-imaging factors for differential diagnosis of MS and NMO. The designed multi-level feature selection mines the multimodal phenotype, which are relevant to the clinical outcome, robust across 1.5T and 3T MRI by leveraging univariate-level statistical analysis and multivariate-level feature interaction from inter-modalality and intra-modalality. Effective interpretation of mined multimodal phenotype, coupled with machine learning methods, can be used as an adjunct to traditional radiology to support the diagnostic process in the clinical practice.

CHAPTER 4

Feature-level Fusion: Multimodal Biomarker Mining for Prognostic Survival Analysis

Identifying the high risk of recurrence in patients with lung cancer and death would be valuable for guiding the enhanced therapy. Therefore, individualized evaluation of the prognosis for this complex and heterogenous entity is particularly important. Computational feature fusion maximizes the information obtaining from the diagnostic images acquired in routine clinical practice and has proven promising results in the diagnosis, response prediction and survival prognosis for several types of cancer patients [135]–[137]. However, these radiomic models have not taken into account the clinically indispensable clinico-pathological or hematological predictors in locally advanced non-small cell lung cancer (LA-NSCLC) studies, and the prognostic performance of radiomics is yet to further improve.

This chapter presents a feature-level fusion framework for prognostic multimodal biomarker mining, from high-dimensional CT imaging features, clinical features, and hematological features, for survival prediction of LA-NSCLC. The proposed integrative feature selection algorithm is proposed to integrate high-dimensional CT imaging features, clinical features, and hematological features as a prognostic, representative, and non-redundant radiomic signature, which is based on consensus clustering and survival regression. The predictive integrated radiomic signature was subsequently fitted into a final prognostic model using both the Cox Proportional Hazard (CPH) model and the Random Survival Forest (RSF) model. A multimodality nomogram was then established from the fitting model and was cross-validated. Finally, calibration curves were generated with the predicted versus actual survival status.

This chapter is organized as follows: Research motivation and dataset description are introduced in Section 4.1 and 4.2, respectively. An Integrative Clustering and Supervised (ICS) feature selection algorithm is presented in Section 4.3. The experimental results and discussion are analyzed in Section 4.4 and 4.5, respectively.

4.1 Research Motivation

Non-small cell lung cancer (NSCLC) accounts for 85% of lung cancer cases, with approximately one third of those being defined as LA-NSCLC, classified as stage III NSCLC [185], [186]. Although concurrent chemotherapy and radiotherapy (CCRT) is considered the standard treatment, outcomes of LA-NSCLC patients remain poor, with a median survival of 12-23.2 months [187]–[189]. TNM is a cancer staging system, in which T (tumor) indicates the depth of tumor invasion, N (node) indicates whether lymph nodes are affected, M (metastasis) indicates whether the cancer has spread to other parts of the body. The TNM-based one-size-fits-all strategy might not be suitable for all patients. Identification of patients at high risk of recurrence and death would be valuable for guiding the enhanced therapy. Therefore, individualized evaluation of the prognosis for this complex and heterogenous entity is particularly important.

Computational radiomics analysis maximizes the information obtaining from the diagnostic images acquired in routine clinical practice and has proven promising results in the diagnosis, response prediction and survival prognosis for several types of cancer patients [43], [151], [190]. In NSCLC, from CT images, the quantitative measure of cancer volume reduction after chemoradiation provided more clinical information on tumor response than conventional response assessment (Response Evaluation Criteria in Solid Tumors, RECIST) [191]. In addition, 18F-fluorodeoxyglucose (FDG) PET features of lung cancer were found to be significantly correlated with T stages, N status, pathological stages, as well as tumor grades [192]–[195]. Several attempts have been made to improve the performance of predictive models. For instance, a grading system combine neutrophil and SUVpeak in PET images was developed by Schernberg *et al.* , which could effectively stratify patients with better overall survival (hr = 5.8, p = 0.001) [196]. However, these radiomic models have not taken into account the clinically indispensable clinico-pathological or hematological predictors in LA-NSCLC studies, and the prognostic performance of radiomics is yet to further improve.

Emerging evidence demonstrated that hematological inflammatory cells could effectively predict the survival of patients with LA-NSCLC [197]. The mutual interaction between tumor and inflammatory cells promoted the evolution and development of cancers. On one hand, NSCLC could drive the stimulation of inflammatory cells in the tumor microenvironment as well as that in systemic circulation systems. On the other hand, these inflammatory cells could play a pivotal role in the initiation and development of NSCLC [198], [199]. Due to the important roles of systemic inflammatory cells in the biology of NSCLC, our hypothesis is that the incorporation of these inflammatory parameters with current radiomic imaging features could improve the predictive capacity.

4.2 Dataset Description

4.2.1 Characteristics of Patients

This study retrospectively includes 118 cases of LA-NSCLC patients from Shandong Cancer Hospital between January 2014 and January 2016. The institutional review board of Shandong Cancer Hospital has approved this retrospective study of these patients. Inclusion criteria include: patients were aged 18 years or older, were diagnosed as stage III NSCLC confirmed by histopathology and radiographic exams according to the AJCC 8th edition TNM classification and staging system; received CCRT without prior therapy or the operations. The exclusion criteria are patients with 1) autoimmune disease; 2) active lung infections judged by clinicians with the consideration of fever, rales or the abnormal blood test findings of abruptly erythrocyte sedimentation rate (ESR), C-reactive protein (CRP), and neutrophiles; 3) the pneumonitis or abcesses not related to the tumor; or 4) other infections such as gastroenteritis, appendicitis, and cholecystitis.

For all of the 118 LA-NSCLC cases, 48 patients were TNM stage IIIA (40.7%), 58 cases were stage IIIB (49.2%) and 12 patients were stage IIIC (10.2%). The median survival of these patients was 19.8 months (95% confidence interval: 4.0-35.6 months). Other clinico-pathological characteristics were shown in Table 4.1.

4.2.2 Clinicopathological and Hematological Parameters

For each patient, we collected the clinico-pathological characteristics including age at diagnosis, gender, tumor location, tumor size, node metastasis status, histological type, KPS, radiation type and doses, concurrent chemotherapy type, usage of consolidative chemotherapy, pre- and post-therapeutical serum tumor biomarkers including CEA, NSE, and Cyfra 211. KPS is to quantify the patient's ability to tolerate therapy in terms of their physical function and ability to take care of themselves, and to perform daily activities. Hematological inflammatory variables included: levels of monocytes, neutrophils, lymphocytes, hemoglobin, as well as platelet counts. Also, NLR, LMR, PLR were calculated for each patient. For each patient, both the pre- and post-therapeutical hematological variables ("1" and "2" were used as markers, respectively) were obtained.

4.2.3 Follow-up and Prognostic Evaluations

Follow-up data were collected from the most recent medical records of these patients, including the information of physical exams, complete blood count, blood biochemistry, tumor biomarkers, thoracic CT scans, and abdominal ultrasound. In addition, we also acquired the survival information of these patients through telephone enquiries, medical insurance records as well as death certificates. Overall survival in this study was defined as the period from the date of admission to the death date regardless of specific causes of death.

TABLE 4.1. Patient characteristics of lung cancer for the prognostic task.

Characteristic (N=118)	Classification	N	%
Age (years)			
	≤60	62	52.5
	>60	56	47.5
Gender			
	Male	104	88.1
	Female	14	11.9
KPS			
	≥80	112	94.9
	<80	6	5.1
Location			
	Central	82	69.5
	Peripheral	36	30.5
Histology subtype			
	SCC	66	55.9
	Non-SCC	52	44.1
T stage			
	T1	8	6.8
	T2	37	31.3
	T3	21	17.8
	T4	52	44.1
N stage			
	N0	12	10.1
	N1	14	11.9
	N2	57	48.3
	N3	35	29.7
Radiotherapy technique			
	3D-CRT	50	42.4
	IMRT	68	57.6
Radiotherapy doze (Gy)			
	≤60	96	81.4
	>60	22	18.6
Concurrent chemotherapy			
	EP	13	11
	PC	39	33.1
	Others	66	55.9

4.3 Integrative Clutering and Supervised (ICS) Feature Selection for Prognostic Prediction

The overall workflow of this study is illustrated in Figure 4.1. After Feature Acquisition of imaging and non-imaging factors (Figure 4.1a), an ICS Feature Selection algorithm was developed to select the most informative, representative and non-redundant features from high-dimensional multi-view features (Figure 4.1b). In our method, the unsupervised clustering contributed to reducing redundancy by exploring the correlation among features, while supervised learning selects informative and representative features by examining the relation between features and outputs. The method was separately published as a conference paper [200]. After feature selection, the prognostic features were fitted into one predictive model using CPH and RSF models, respectively (Figure 4.1c). Lastly, nomograms for 1-year and 2-year overall survival were generated for these patients (Figure 4.1d).

4.3.1 Image Segmentation and Radiomic Feature Extraction

For each patient in this study, we collected both pre- and post- CCRT contrast-enhanced CT images using a Somatom Definition AS (Siemens Healthineers). The CT parameters were as follows: tube voltage, 120 kVp; tube current, 200 mAs; detector, 64 × 0.625 mm; beam pitch, 1.5. First of all, three-dimensional Gross Tumor Volume (GTV) was interactively segmented and delineated using an in-house segmentation software based on Random Walker algorithm [201], [202]. This delineation procedure was performed twice on all CT images with the interval about 2 months between the first and second evaluations to reduce the operator's biases. Delineation is the action of manually segmenting the regions of interest on medical imaging.

In total, 1045 comprehensive CT image features, including intensity, shape, texture, and wavelets [43], were extracted from 118 LA-NSCLC cases. Intensity features were calculated by the first order statistics through the tumor voxel intensity distributions. Shape features were

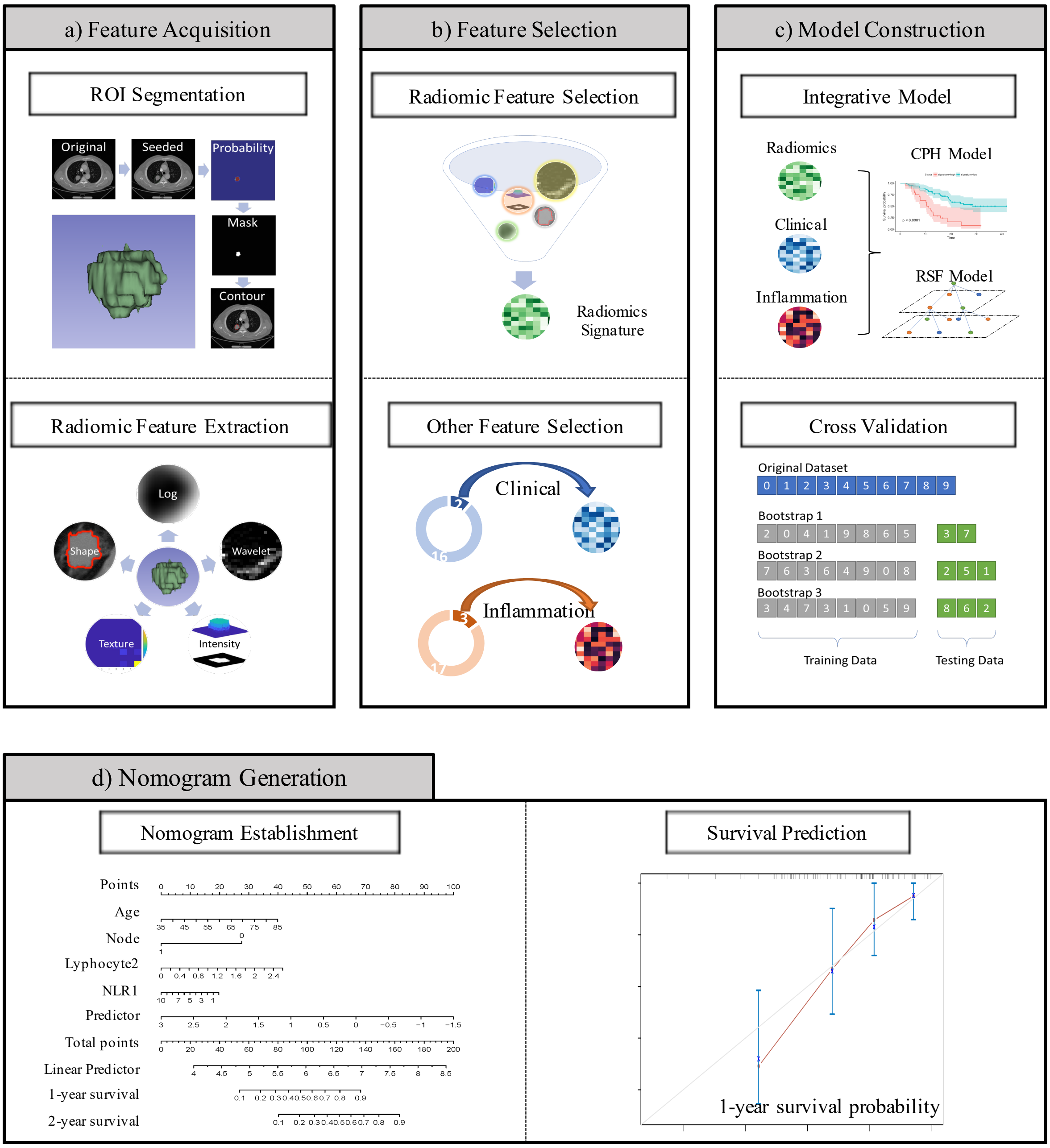


FIGURE 4.1. Workflow of generation of a comprehensive radiomic based nomogram.

extracted to reflect 3D geometric features of the tumor, such as surface area, compactness, and tumor volume. Texture features were described using texture matrix such as GLCM and GLSZM to quantify internal tumor heterogeneity and using Log filters to depict different tumor coarseness with different sigma values [203]. Wavelet features could extract intensity and texture features in the frequency domain using wavelet decomposition on the original images. Specifically, there were 18 first-order intensity features, 13 shape features, 68 texture features, 258 log features and 688 wavelet features extracted. For the implementation of feature extraction, first-order intensity features were extracted with 1000 voxel array shift. LoG features were extracted with sigma set to 1mm, 3mm, and 5mm. Wavelet features were extracted using "coif" through 8 channels including LLL, LLH, LHL, LHH, HHH, HLL, HLH, HHL.

4.3.2 ICS Feature Selection for Prognostic Prediction

The ICS Feature Selection was proposed in our paper [200] and contained four three major components: Reproducible Feature Selection, Prognostic Feature Selection, and Non-redundant Feature Selection. The workflow of ICS Feature Selection was illustrated in Figure 4.2. Firstly, reproducible features were selected from two batches of features extracted from two delineations, in which consistency was assessed with Pearson correlation analysis. Secondly, Prognostic Feature Selection was aimed to select informative and representative features. To this aim, consensus clustering method, combined with CPH, was then used for the selection of prognostic features based on the p-value ranking. Clustering combined with RSF, an ensemble tree method for analyzing right-censored survival data was used to generate trees, and was performed for comparison. Thirdly, redundancy still remaining in the selected feature subsets was further eliminated using pairwise correlation analysis.

Reproducible feature selection. Reproducibility refers to the feature characteristic of retaining high correlation across different batches of delineations, which is important in clinical tasks where inter-observer variety presents. To select reproducible features, Pearson correlation was deployed to compute the association of individual features across two batches of delineations. Pearson correlation is a classic linear correlation measurement. Denote a certain feature computed from the first delineation as X and the same feature from the second delineation as Y, Pearson correlation is formulated as:

$$r = \frac{\sum_i (x_i - \hat{x})(y_i - \hat{y})}{\sqrt{\sum_i (x_i - \hat{x})^2}\sqrt{\sum_i (y_i - \hat{y})^2}} \tag{4.1}$$

where $\hat{x}$ and $\hat{y}$ are the mean values of variables X and Y. The value of r ranges from -1 to 1, where 1 and -1 means that two variables are totally correlated while 0 means two variables are completely independent. After computing Pearson correlation for each feature across delineations, reproducible features were identified using a correlation threshold.

Prognostic feature selection. Prognostic Feature Selection was the most important module in our ICS Feature Selection algorithm. Two criteria were defined for selecting prognostic features - informativeness and representativeness. Informative features refer to the features with the outstanding capability to predict the clinical outcome, i.e. time-to-event survival label for survival prediction. Representative features are defined as the ability to represent its natural grouping and predict the output, as well as with low redundancy.

Informative (Prognostic) feature selection. To identify and rank informative features, supervised feature selection was deployed because of its excellent performance to discover the feature relevance to the output. In our method, we used univariate Cox selection [57] as our supervised feature selection method, a classical wrapper technique designed for time-to-event prediction. The univariate Cox selection was based on CPH model, ranking features by their statistical significance.

The CPH model is one of the most general regression models because it does not assume the underlying survival distribution. The CPH uses the hazard function as a response, which is represented as:

$$h(t, x) = h_0(t)exp(\beta' x) \tag{4.2}$$

where h_0 is an arbitrary baseline hazard and $\beta = (\beta_1\ , \ \ldots\ , \ \beta_p\ )'$ represents an array of unknown regression coefficients. As h_0 is not dependent of covariates, the regression coefficients

β can be obtained by maximizing the partial log-likelihood as:

$$L(\beta) = \prod_{i=1}^{k} \frac{exp(\beta' x_i)}{\sum_{l \in R(t_i)} exp(\beta' x_l)} \tag{4.3}$$

where k denotes distinct ordered survival times and x_i denotes the covariate value with distinct ordered survival times.

In our method, we applied univariate CPH model on individual features to rank the importance of each feature. The importance score was the p-value computed with Wald test, representing the statistical significance of the features. Then, feature informativeness rank can be obtained by sorting the p-values of features ascendingly, where features with low p-values are identified as informative features.

Representative (Prognostic) feature selection. To select representative features, we firstly divided the full feature space into several highly correlated feature clusters and assigned each feature a cluster number using clustering algorithms. The unsupervised clustering is commonly used to reveal hidden structure based on inherent feature information in the datasets without labels. Then, the representative features were selected from each cluster according to the feature informativeness rank obtained by supervised feature selection.

Consensus clustering [204] was adopted in our framework because it could determine the number of clusters within the algorithm and use resampling to improve the stability, robustness, scalability of clustering results. This clustering method determined the final clusters using the consensus across multiple runs of a base clustering algorithm, analogous to ensemble learning in supervised learning. To fit the consensus clustering algorithm, the data were required to be pre-processed, including feature omitting and feature standardization. Then, the number of clusters was computed with the plan that gave the highest median cluster consensus on all clusters. Cluster consensus was defined as the mean of consensus between all pairs of features from the same cluster.

Non-redundant feature selection. Non-redundant feature subsets refer to as small feature subsets as possible with minimum internal correlation within the subset. Although the redundancy had been significantly reduced by selecting representatives from highly correlated feature clusters, there still existed inter-cluster correlation among feature representatives for each cluster. Therefore, we used pairwise correlation analysis to further remove redundancy.

The pairwise correlation analysis assessed the redundancy by constructing a Pearson Correlation Matrix (PCM) for every pair of features in the representative feature subsets. Given selected variables P and Q from the feature subset, the computation of Pearson Correlation was similar as Equation 4.1:

$$r' = \frac{\sum_i (p_i - \hat{p})(q_i - \hat{q})}{\sqrt{\sum_i (p_i - \hat{p})^2}\sqrt{\sum_i (q_i - \hat{q})^2}} \tag{4.4}$$

where $\hat{p}$ and $\hat{q}$ were the mean values of variables X and Y. As the Pearson correlation was a symmetrical measure for the two variables, the PCM was simplified by removing the symmetrical upper triangle in the matrix. Based on the simplified PCM, we further reduced the redundancy using following criteria: Firstly, identify the feature pairs with Pearson correlation higher than a threshold. Secondly, preserve the one with higher feature informativeness (low p-value) in each pair.

4.3.3 Prognostic Model and Nomogram Construction

Prognostic model establishment. Multimodal features and parameters including radiomic, clinico-pathological, as well as inflammatory features were fused into a single predictive model based on multivariate CPH model. Performance of this model was evaluated with the concordance index (C-index). For comparison, in the RSF model, the possible split points for each variable were examined to find the optimal split method.

Cross-validation. Bootstrap based cross-validation was applied to assess and compare the discriminative power of CPH model and RSF model. These prediction models were trained on 10% of total bootstrap samples drawn with replacement from the original data while tested in the observations that were not in the training sets. Then, the C-index was computed for different timepoints (with a constant interval of 1 month) and the mean of those C-indexes were calculated to represent the model discriminative ability [205].

Nomogram construction. Nomogram, a more interpretable, graphical representation of predictive models that can include different types of predictive markers, has become the focus of interest in the cancer research in recent years [206]–[208]. Model with better C-index was chosen for further nomogram construction. Calibration curves of the nomogram were then drawn for 1-year and 2-year overall survival of the patients. The calibration curves illustrated both survival probabilities predicted by the nomogram and the observed probabilities.

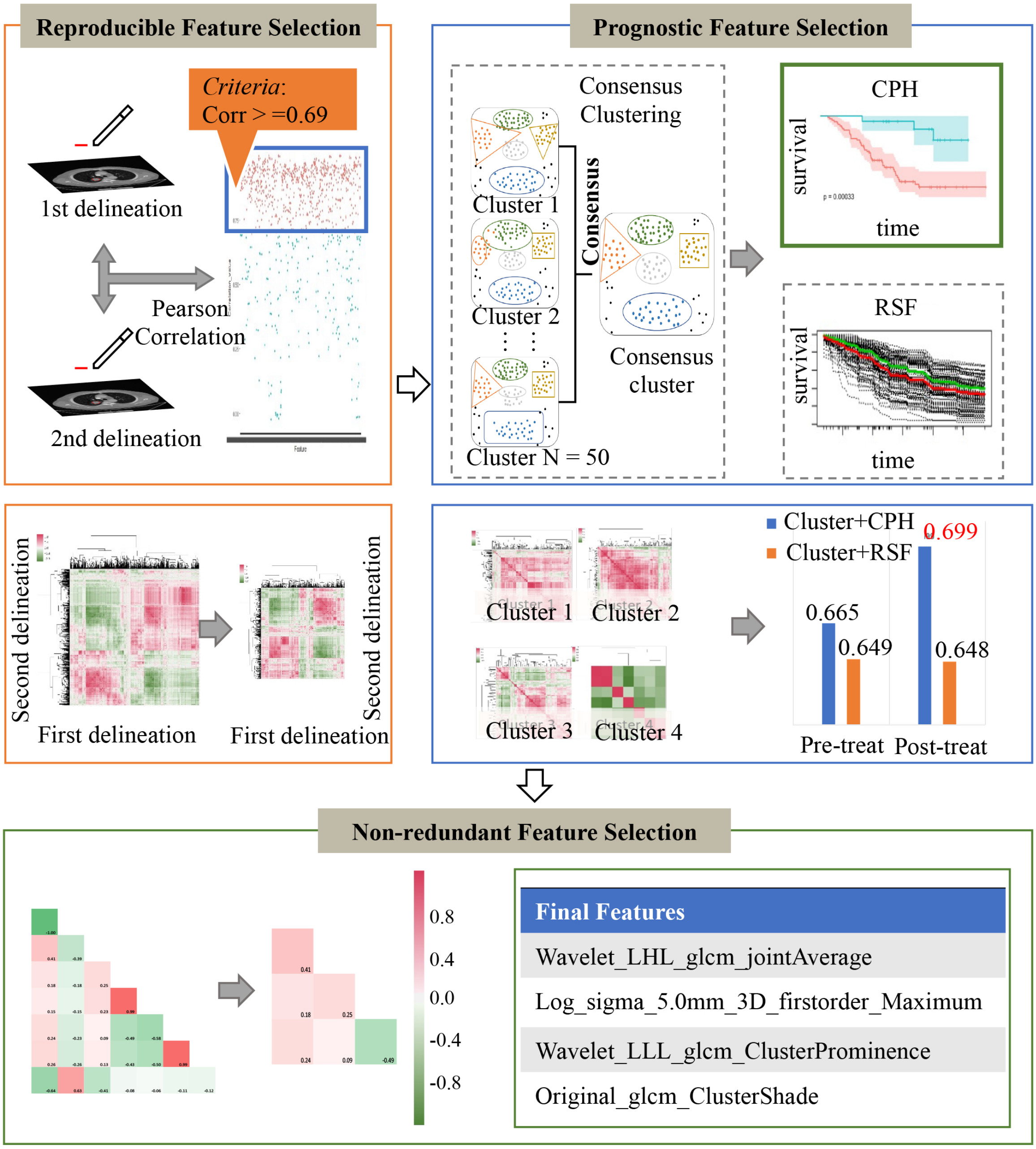
All statistical analyses are two-sided, with the significance level of 0.05. Statistical analyses were performed with “rms”, “Hmisc”, “survival”, “pec”, as well as “randomForestSRC” modules in R programming language and environment (http://www.r-project.or) as well as STATA software (version 14.1, College Satation, TX, USA).

4.4 Experimental Results

4.4.1 Result of Feature Selection

In total, 1,045 radiomic features were extracted from the CT images including 70 sets of pre-CCRT and 97 sets of post-CCRT. We first ranked the stability of the 1045 features using Pearson correlation coefficients calculated between the two delineations. As a result, 829 stable features were selected for the subsequent analyses (Figure 4.2a). Then, two hybrid selection methods, i.e., clustering combined with CPH and clustering combined with RSF, were used and compared in our study. Features selected by clustering combined with CPH were found to be more predictive with a C-index of 0.699 in comparison to 0.648. Radiomic features extracted from post-treatment CT images were found to be superior in prediction than that from pre-treatment CT images (Figure 4.2b). Thus, features from post-treatment

CT images were selected by the method of clustering combined with CPH and were further investigated in the following study (Figure 4.2b).



Final Features
Wavelet_LHL_glcm_jointAverage
Log_sigma_5.0mm_3D_firstorder_Maximum
Wavelet_LLL_glcm_ClusterProminence
Original_glcm_ClusterShade

FIGURE 4.2. Integrative Clustering and Supervised (ICS) Feature Selection. Abbreviation: Corr = Correlation.

With backward elimination algorithm, the least prognostic features were repeatedly removed from the clustered feature subset until the subset was able to achieve the optimal predictive performance. Then, the top eight prognostic features selected from the clusters were analyzed with further correlation analyses to avoid overfitting (Figure 4.2c). After identifying pairs of highly correlated features (Pearson Correlation coefficient > 0.9), the one with higher p-value in each pair was eliminated. Finally, four independent predictive radiomic features including wavelet-LHL_glcm_JointAverage, wavelet-LLL_glcm_ClusterProminence, original_glcm_ClusterShade, and log-sigma-5-0-mm-3D_firstorder_Maximum, were used to generate the radiomic signature which also had good predictive capacity in the Kaplan-Meier analyses of these patients (Figure 4.3).

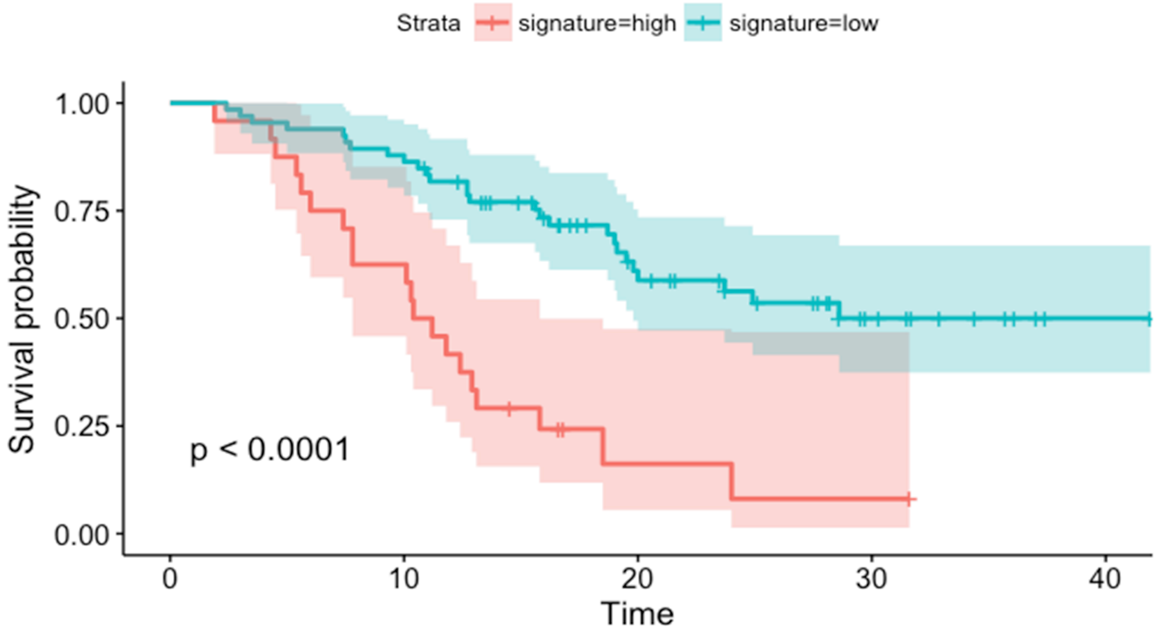


FIGURE 4.3. Predictive capacity of radiomic signature with Kaplan-Meier curve.

Other validation details of the proposed ICS feature selection can be found in our publication [200].

4.4.2 Association of Selected Features with Hematological Inflammatory Variables

Due to the importance of inflammatory factors in the prognosis prediction of patients with NSCLC, we further explored the correlation of selected features with the hematological inflammatory variables. Two of the four selected features were found to be significantly correlated with specific inflammatory factors. In particular, the "wavelet_LHL_glcm_JointAverage" feature was positively correlated with the levels of platelet1 and PLR1 while negatively correlated with levels of LMR2 significantly (p = 0.026, p = 0.045, and p = 0.048, respectively). In addition, the "log_sigma_5_0_mm_3D_firstorder_Maximum" feature was significantly positively correlated with both pre- and post- therapeutic platelet levels (p = 0.013, and p = 0.049, respectively) (Figure 4.4).

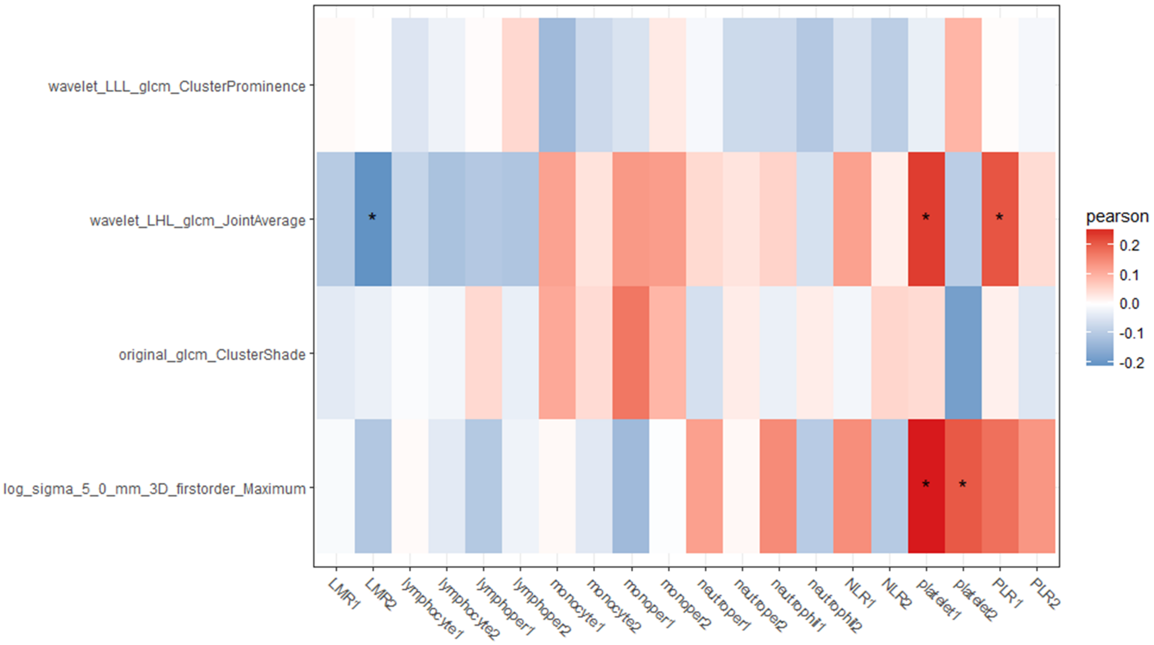


FIGURE 4.4. Correlation analyses of selected radiomic features with hematological inflammatory cells. Selected radiomic features are also analyzed for the correlation with inflammatory cells for included patients. Red bars show the positive correlation while the blue ones denote the negative association. Bars with asterisk denote the correlation has reached significance.

4.4.3 Performance of Multimodality Prediction Model

Patient age and lymph node metastases were found to be independent risk factors in our study using multivariate CPH. In addition, for the inflammatory parameters, lymphocyte2 levels and NLR1 were found to be independent prognostic factors for our patient cases.

Next, multivariate CPH and RSF were used and compared for assessing the performance of the predictive model. C-index of CPH and RSF was not stable until 505 days, and C-index thereafter was selected for our study. The C-index of the CPH model was 0.792 and it retained 0.743 after cross-validation (Figure 4.5-4.6). In comparison, C-index of the RSF model dropped from 0.891 to 0.647 when cross-validation was performed (Figure 4.5-4.6). Again, the CPH model was found to be more stable and was ascertained for the further construction of the nomogram.

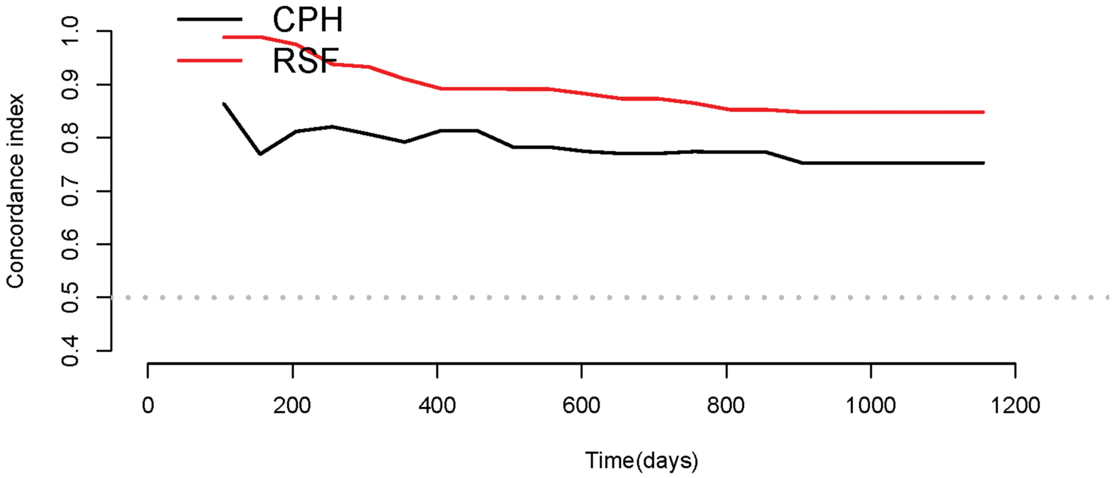


FIGURE 4.5. C-indexes of the model including radiomic, clinico-pathological and inflammatory parameters using CPH and RSF methods.

Importantly, the performance of this integrative model was proven to be superior to radiomic, clinico-pathological or hematological models alone, with C-indexes of 0.699, 0.618, and 0.653, respectively.

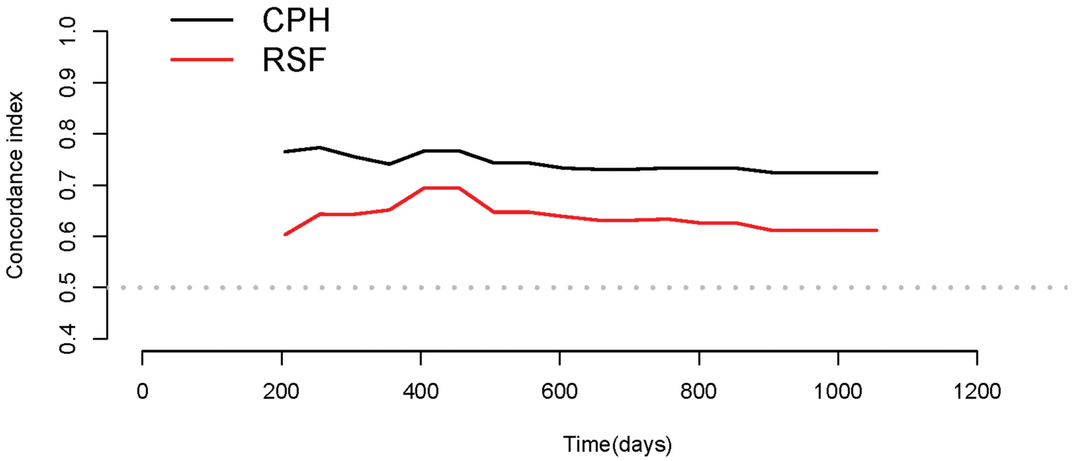


FIGURE 4.6. Cross-validations are performed using CPH and RSF methods.

4.4.4 Performance Interpretation with Nomogram

Nomogram for prediction performance (Figure 4.7) of 1-year and 2-year survival was generated on the basis of the selected radiomic signature, patient age, lymph node metastasis, lymphocyte2 levels, and NLR1.

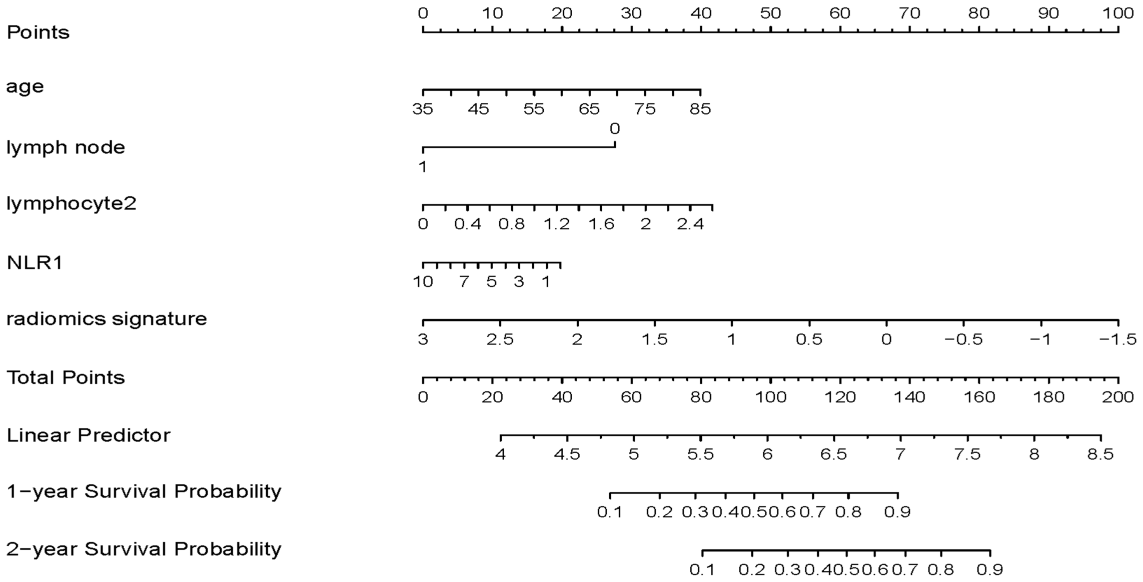


FIGURE 4.7. A comprehensive nomogram for prediction of 1-year and 2-year overall survival for patients with LA-NSCLC.

Furthermore, a calibration curve had been drawn for these patients. The estimated versus observed 1-year and 2-year survival probabilities intersected the 45-degree line, showing that the predicted value approximated the observed value within a 95% confidence interval (Figure 4.8-4.9). This calibration curve shown the agreement between the predicted and actual values.

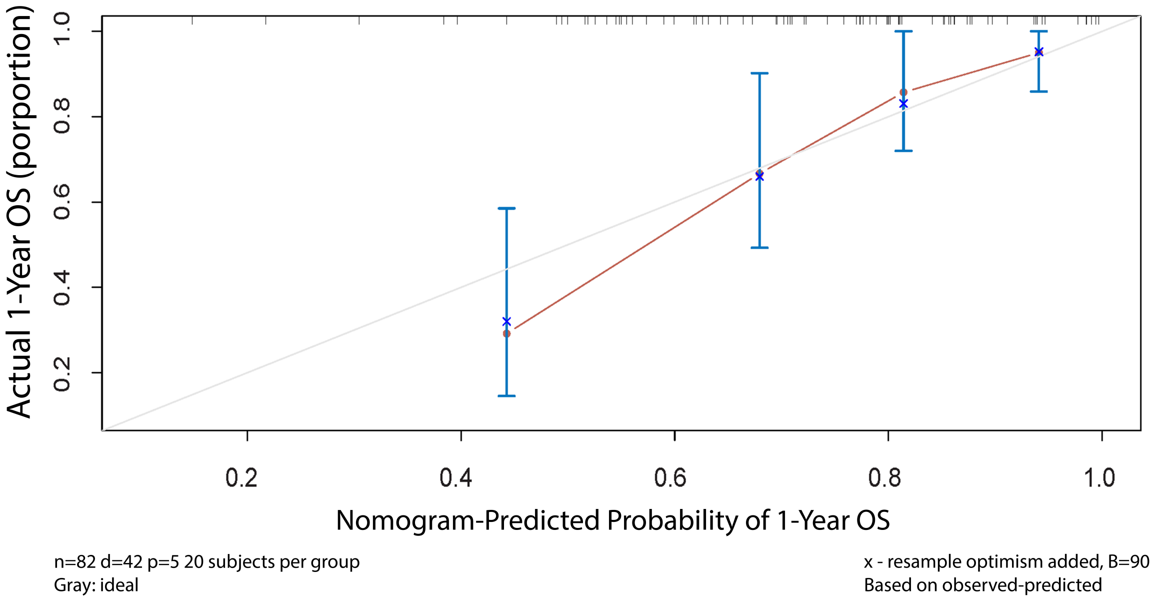


FIGURE 4.8. Calibration curve for estimation of 1-year overall survival predicted by nomogram. Nomogram-estimated overall survival is plotted on the x-axis; actual overall survival is plotted on the y-axis. Dash line represents the ideal agreement.

4.5 Discussion

In this study, we incorporated comprehensive multimodal radiomic, clinico-pathological and hematological factors for the individualized survival prediction of LA-NSCLC patients. To the best of our knowledge, for the first time, a concise nomogram with only five variables can provide a feasible and practical reference to clinical professionals for recommending a more appropriate management for LA-NSCLC patients. We also found that the selected radiomic features were associated with inflammatory variables in these patients, suggesting that the

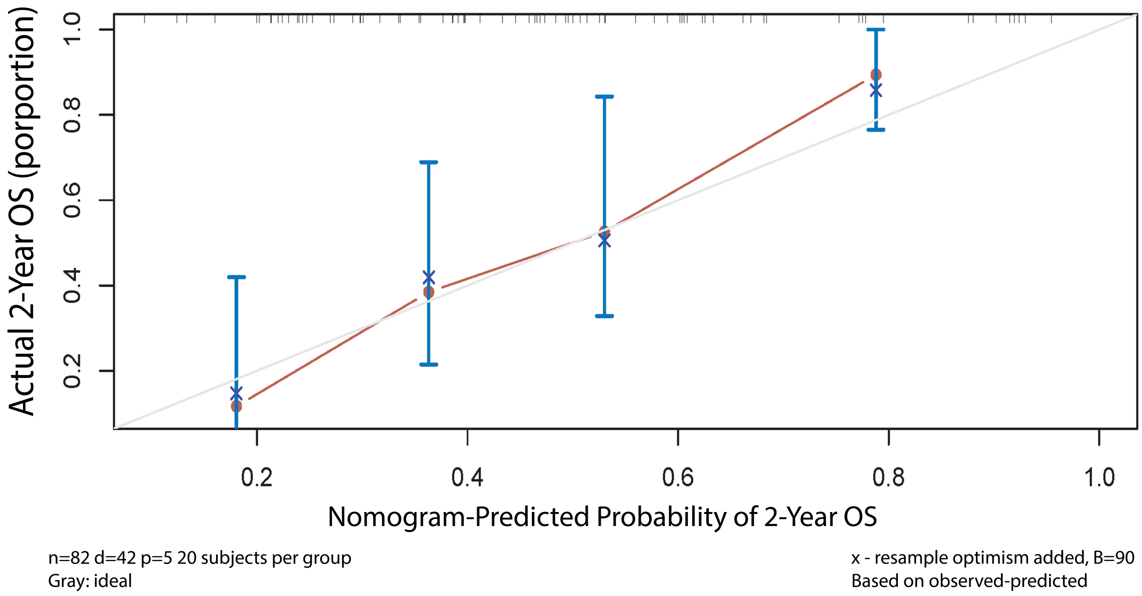


FIGURE 4.9. Calibration curve of 2-year overall survival predicted by nomogram for patients with LA-NSCLC.

inflammatory status may partially account for the poor survival of patients harboring these radiomic features.

The performance of the integrative model was also shown to be superior to the individual model alone in our present study, demonstrating powerful predicting capability using different types of biomarkers. As reported, the C-index of the radiomic model was often between 0.60 and 0.67, which has been improved to 0.72 when combining with clinical and genomic features [205], [209]–[211]. This improvement due to information integration of the distinct sources may reflect that multiple factors of the patient characteristics contribute to a more accurate prediction model. In comparison to using genomic features, our new nomogram incorporating the clinical, hematological and CT imaging data, which are all routinely evaluated in clinical settings, could be more feasible in the clinical practice.

Most studies correlating radiomics with survival outcomes in lung cancer analyzed the baseline features of pre-treatment [212], [213]. However, tumors undergo dynamic changes during treatment, which would be more informative [214]–[216]. Thus, we further analyzed the CT features pre- and post- CCRT dynamically. Rather than keeping the same set of important radiomic features from baseline for analyzing post-treatment CT data [209], we selected the important prognostic features of pre- and post-treatment, respectively. Our analysis on these features found that the performance of the post-treatment features was much higher than that of the baseline features, which demonstrated that post-treatment features were able to better reflect the actual response to CCRT and were more informative and accurate for predicting the patients' prognosis. This finding suggests that CT scan after CCRT is also recommended for LA-NSCLC patients.

Interestingly, our selected radiomic features were found to be associated with inflammatory biomarkers including levels of platelet, LMR and PLR. Platelets play a role in protecting tumor cells from antitumor-immunity, and releasing cytokines for tumor progression [217]; monocytes have been proven as an important factor in favoring tumor invasion and metastasis by producing protease enzymes [218], [219]. In contrast, lymphocytes are a protective factor by inducing cytotoxic cell death and inhibiting tumor cell proliferation and migration [220]. Hence, the elevated platelet or PLR and the decreased LMR are considered to be associated with worse prognosis of patients due to their important roles in the initiation and development of cancers [221]–[223]. Our study found that radiomic feature "wavelet_LHL_glcm_JointAverage" or "log_sigma_5_0_mm_3D_firstorder_Maximum" positively correlated with the levels of platelet1 or PLR1 while negatively correlated with LMR2. These radiomic features may indicate the unfavorable immunological status which at least partly accounts for the prognostic effects of radiomic features in LA-NSCLC patients. Yet, the mechanism underlying the predictive capacities of the radiomic features and their relationship with inflammatory biomarkers still need to be further investigated.

We investigated both the effects of CPH and that of RSF for feature selection and model fitting, and found that CPH was much more stable and reliable than RSF. Although RSF could reach higher C-index in the primary model establishment analysis, the C-index dropped remarkably in the subsequent cross-validation stages, which was consistent with the reported findings in glioblastoma research [203], [224]. We speculate that the selection of algorithms in the machine learning model establishment stage would be influenced by the sample size of the study. Only when the sample size is sufficiently large, could we include a bigger number of parameters in machine learning models while avoid the risk of overfitting.

There are some limitations in our present study. Firstly, due to 32 patients were not confirmed the cause of death, we therefore only analyzed OS for evaluating the patient prognosis. In the future study, it could be better if cancer specific survival is investigated for the prediction of patients with LA-NSCLC. Secondly, this was a retrospective study, and prospective trials in different centers and regions could eliminate the selection bias. In addition, the underlying mechanism for explaining the prognostic role of our nomogram still needs to be further investigated in the future. The analysis of genomic types with different driving genes might be helpful for understanding the biological characteristics of the patients with poor outcomes who harbor the worse integrative features of radiomic, clinico-pathologics and hematology simultaneously.

4.6 Chapter Summary

In conclusion, we have constructed a simple, yet not trivial, nomogram integrating high-dimensional CT imaging features, clinicopathological, and hematological factors, which would have potential as an individualized utility in the clinical practice for LA-NSCLC patients. This nomogram has value to permit non-invasive, comprehensive, and dynamical evaluation of the phenotypes of LA-NSCLC and to predict the survival prognostication for LA-NSCLC patients.

CHAPTER 5

Information-level Fusion: Interpretable Deep Correlational Fusion Framework for Inter-modal and Intra-modal Information Analysis

Fusion of multimodal medical data, in the information-level, is critically important for a more complete understanding of the disease characteristics and therefore essential to accurate computer-aided diagnosis. Although deep multimodal learning has showed strong modeling capacity compared with conventional multimodal modeling, there are still three major challenges, including exploiting inter-modal and intra-modal information in supervised and unsupervised settings and understanding of complex non-linear cross-modal association.

To address these three challenges in information-level fusion, we propose an Interpretable Deep Correlational Fusion framework in this chapter, to optimize the discovery of multimodal biomarkers in both supervised and unsupervised settings and boost the interpretability of multimodal deep learning.

- For the supervised setting, a novel DMFusion loss is proposed to optimize the discovery of discriminative multimodal representations in low-dimensional latent fusion space. It is achieved by jointly exploiting inter-modal correlational associations via CCA loss and intra-modal structural and discriminative information via reconstruction loss and cross-entropy loss.
- For the unsupervised setting, a new unified SDC loss function is proposed to incorporate consensus information into discriminative representations, in which, the

former is learnt by maximizing the canonical correlation among multi-view representations projected by neural networks, and the later is achieved through using confident clustering assignments as supervision.

- For interpreting the complex nonlinear cross-modal association in deep fusion network, we propose a cross-modal association (CA) score to quantify the importance of input features towards the correlated association, by harnessing integrated gradients in deep networks and canonical loading in CCA projection.

This chapter is organized as follows: Section 5.1 presents a supervised deep multimodal fusion model with a novel loss for diagnostic predictions. Section 5.2 introduces a unsupervised deep multi-view clustering model, named self-supervised deep correlational multi-view clustering for computer vision and audio recognition tasks. Section 5.3 presents a new interpretation module for understanding complex non-linear cross-modal associations in deep fusion networks. The experimental implementations and results are sumamrized in Section 5.4 and Section 5.5, respectively.

5.1 Supervised Deep Multimodal Fusion Network for Diagnostic Decisions

Figure 5.1 illustrates the overview of the proposed Deep Multimodal Fusion network and Interpretation module on cross-modal association. Firstly, we introduce the novel DMFusion loss to jointly exploit inter-modal relation-driven association and intra-modal data-driven and target-driven discriminative information, which are then integrated via DMFusion Layer for diagnostic decisions. Secondly, we propose a cross-modal association (CA) score to interpret the importance of input features towards correlational consensus.

The architecture in our DMFusion network consists of three modules: Multimodal Encoders, Multimodal Decoders and DMFusion Layer. Multimodal Encoders $f_x(X;\theta_x)$ and $f_y(Y;\theta_y)$ project inputs $X \in \mathbb{R}^{n*d_x}$, $Y \in \mathbb{R}^{n*d_y}$ to their corresponding representations. Multimodal

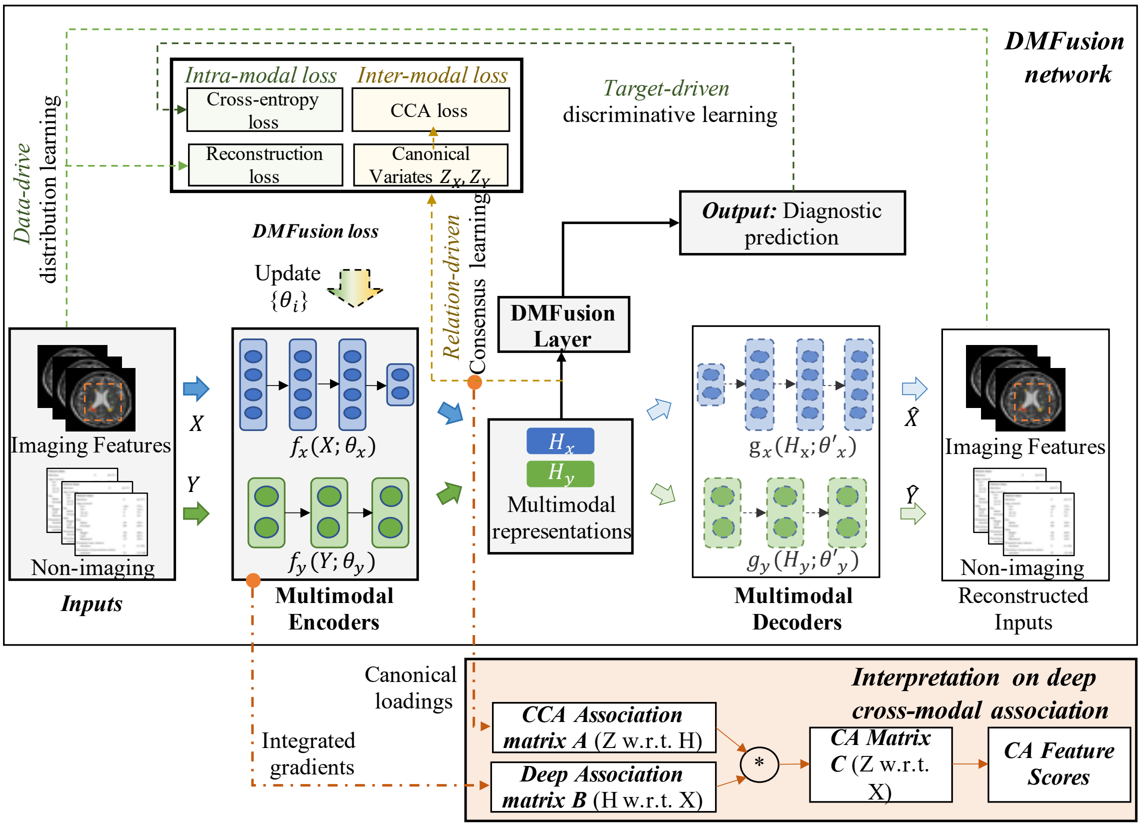


FIGURE 5.1. The flowchart of the proposed interpretable DMFusion network. The architecture of DMFusion consists of Multimodal Encoders, Decoders and DMFusion Layer. A novel DMFusion loss optimizes the multimodal representations with inter-modal common information and intra-modal discriminative information, which are fused in DMFusion Layer for diagnostic prediction. Interpretation of deep cross-modal association is achieved by harnessing canonical loadings from CCA projection and integrated gradients from deep networks.

Decoders $g_x(H_x;\theta'_x)$ and $g_y(H_y;\theta'_y)$ project the encoded representations $H_x \in \mathbb{R}^{n*d_o}$ and $H_y \in \mathbb{R}^{n*d_o}$ to reconstructed input features $\hat{X}$ and $\hat{Y}$ to learn data-driven structural information. Both encoders and decoders are implemented with Multi-layer Preception (MLP) where θ represents network parameters such as weights and bias. The dimension of the inputs is denoted as d_x, d_y, while d_o denotes the dimension of multimodal representations and canonical variates. The number of samples is n.

5.1.1 Multimodal Representations via DMFusion Loss

Our novel DMFusion loss function jointly harnesses relation-driven consensus learning, target-driven discriminative learning and data-driven distribution learning to exploit inter-modal common information and intra-modal discriminative information. DMFusion loss is formulated as:

$$L_{DMF} = \lambda_1 L_{cca} + \lambda_2 L_{ce} + \lambda_3 L_{re} \tag{5.1}$$

where λ_1, λ_2 and λ_3 are trade-off parameters for inter-modal CCA loss and intra-modal cross-entropy and reconstruction loss respectively.

Firstly, we concentrate inter-modal common information in multimodal representations via a deep CCA loss function L_{cca}. To be more specific, L_{cca} aims to maximize the pairwise correlation of canonical variates Z_x, Z_y [83]. The canonical variates are projected from multimodal representations H_x, H_y using the linear projection matrix U and V, respectively. The formulation of L_{cca} is summarized below:

$$\begin{aligned}
&\max_{f_x, f_y} && corr(f_x(X;\theta_x), f_y(Y;\theta_y)) \\
=&\max_{\theta_x,\theta_y,U,V} && \frac{1}{N} tr(U^T f_x(X;\theta_x) f_y^T(Y;\theta_y) V) \\
&\text{s.t.} && U^T \hat{\Sigma_{xx}} U = I \\
& && V^T \hat{\Sigma_{yy}} V = I \\
& && u_i^T f_x(X) f_y^T(Y) v_i = 0, \quad i < j
\end{aligned} \tag{5.2}$$

For accurate covariance estimation, we define $\hat{\Sigma_{xx}} = \frac{1}{N} f_x(X) f_x^T(X) + r_x I$ and similarly for $\hat{\Sigma_{yy}}$ where regulation parameters $r_x > 0$, $r_y > 0$ are introduced. By negating Formula 5.2, inter-modal loss L_{cca} can be obtained:

$$\begin{aligned}
L_{cca} &= -\frac{1}{N} tr(U^T f_x(X;\theta_x) f_y^T(Y;\theta_y) V) \\
&\text{s.t. same constraints as Equation 5.2}
\end{aligned} \tag{5.3}$$

As the common representation among multimodalities is not necessarily discriminative, we seek to capture intra-modal discriminative information via both target-driven and data-driven approaches. Specifically, target-driven intra-modal loss is formularized as a cross-entropy loss on the encoded representation H_x, H_y to boost the relevancy between the representations and the true label:

$$L_{ce} = \frac{1}{N}\sum_{i=1}^{N} l_i log(\sigma(f_x(x_i))) + l_i log(\sigma(f_y(y_i))) \tag{5.4}$$

where l_i is the diagnostic ground-truth of i^{th} sample and σ is a probability function implemented as softmax in our method.

To capture the data-driven intra-modal structural information, we encode the hidden information underlying data distribution into multimodal representations by enforcing the similarity between reconstructed multimodal features $\hat{X}, \hat{Y}$ and input features X, Y via a reconstruction loss.

$$L_{re} = \frac{1}{N}\sum_{i=1}^{N} (||x_i - g_x(f_x(x_i))||^2 + ||y_i - g_y(f_y(y_i))||^2) \tag{5.5}$$

where $\hat{x_i} = g_x(f_x(x_i)) \in \hat{X}$ and $\hat{y_i} = g_y(f_y(y_i)) \in \hat{Y}$.

With the formulated DMFusion loss, optimization of the DMFusion network is achieved by minimizing the loss with RMSprop. The loss is backpropagated to the representation encoders f_x and f_y and iteratively tunes the network parameters θ_x and θ_y in order to enrich the inter-modal common information and intra-modal discriminative information in modality-specific representations H_x and H_y. The fusion scheme for H_x and H_y will be illustrated in the next section.

5.1.2 Fused Representation via DMFusion Layer

Effective fusion of inter-modal information and modality-specific information is essential for boosting classification performance. Yuan *et al.* [225] have investigated two fusion schemes to fuse the outputs of linear CCA through Serial Feature-level Fusion (SFF) and Parallel

Feature-level Fusion (PFF). SFF achieved outstanding performance on face recognition. In our work, we extend SFF as a DMFusion Layer to fuse the correlational outputs from non-linear CCA network.

For conventional linear CCA, SFF proposed by Yuan *et al.* [225] are formularized as

$$SFF = \begin{pmatrix} w_x^T X \\ w_y^T Y \end{pmatrix} = \begin{pmatrix} w_x & 0 \\ 0 & w_y \end{pmatrix}^T \begin{pmatrix} X \\ Y \end{pmatrix} \tag{5.6}$$

For non-linear CCA networks, we extend SFF to a DMFusion Layer to cohesively integrate modality-specific knowledge and inter-modal common information into a comprehensive representation P, which is used for diagnostic prediction.

$$P = \begin{pmatrix} H_x \\ H_y \end{pmatrix} = \begin{pmatrix} f_x(X;\theta_x) \\ f_y(Y;\theta_y) \end{pmatrix} \tag{5.7}$$

where $h_x^i \in H_x$ and $h_y^i \in H_y$ are encoded modality-specific representations. The fused representation P is then leveraged for diagnostic classification.

5.2 Self-supervised Deep Correlational Multi-view Clustering

In this section, we propose Self-supervised Deep Correlational Multi-view Clustering (SDC-MVC) network. We firstly introduce a novel SDC loss function. Then, we investigate a feature-level fusion scheme for multi-view representations followed by the specifics of optimization. Figure 5.2 shows the architecture of our SDC-MVC network, which is composed of three modules, including multi-view representation module, DSF fusion layer, and self-supervised clustering layer.

Notations. For simplicity, the method is demonstrated with two views X_1 and X_2, which is extensible to multi-views. In multi-view representation module, the representation of each view is learnt through a non-linear multilayer perceptron (MLP) neural network, denoted as

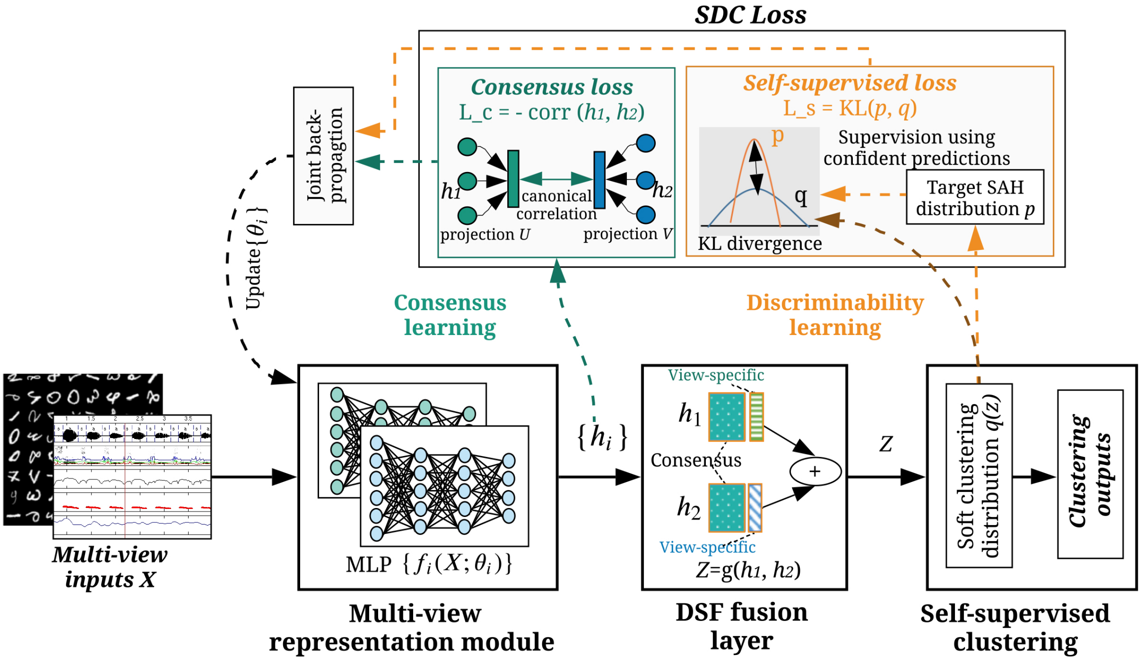


FIGURE 5.2. The overview of the proposed SDC-MVC. Its architecture includes multi-view representation module, DSF fusion layer and self-supervised clustering. A novel SDC loss is proposed to jointly optimize consensus and discriminative representations for multi-view clustering by 1) maximizing canonical correlation of the projected multi-view representations through *consensus loss*; 2) iteratively refining representations and clusters using an auxiliary target distribution p through *self-supervised loss*. Deep Serial Feature-level (DSF) Fusion layer integrates consensus and view-specific discriminative information in multi-view representations for clustering.

$f_1(X_1;\theta_1) : X_1 \rightarrow H_1$ for $X_1 \in \mathbf{R}^{d_1 * N}$ and $f_2(X_2;\theta_2) : X_2 \rightarrow H_2$ for $X_2 \in \mathbf{R}^{d_2 * N}$. θ represents all learnable parameters including weights and bias. The dimension of representations $H_1, H_2 \in \mathbb{R}^{L*N}$ is denoted as L and the number of samples is denoted as N.

5.2.1 SDC Loss Function

The major contribution is the novel SDC loss, which enables joint optimization of multi-view correlational representation and self-supervised deep clustering in a fully unsupervised manner. It can be formularized as:

$$L = \lambda L_c + L_s \tag{5.8}$$

where $\lambda > 0$ is a coefficient to balance the multi-view learning and deep clustering. L_c denotes the correlational loss to constrain multi-view correlation, while L_s denotes the self-supervised loss to constrain learning from clustering results. This objective function can also be seen as adding additional regularization to self-supervised clustering. It is important to optimize correlational representation loss alongside self-supervised loss; otherwise, correlational consensus would tend to collapse during the optimization of self-supervised loss, even though the representation is pre-trained.

The correlational loss L_c is implemented with deep CCA. Following the non-linear CCA formula in Equation 2.6, maximization of CCA can be expressed as:

$$\begin{aligned} &\max_{f_1,f_2} && corr(f_1(X_1;\theta_1), f_2(X_2;\theta_2)) \\ = &\max_{\theta_1,\theta_2,U,V} && \frac{1}{N} tr(U^T f_1(X_1;\theta_1) f_2^T(X_2;\theta_2) V) \\ &\text{s.t.} && U^T \hat{\Sigma_{11}} U = I \\ & && V^T \hat{\Sigma_{22}} V = I \\ & && u_i^T f_1(X_1) f_2^T(X_2) v_i = 0, \quad i < j \end{aligned} \tag{5.9}$$

where $U \in \mathbf{R}^{L*N}$ and $V \in \mathbf{R}^{L*N}$ are projection matrixes for the output of $f_1(X_1;\theta_1)$ and $f_2(X_2;\theta_2)$ respectively. For accurate covariance estimation, we define $\hat{\Sigma_{11}} = \frac{1}{N} f_1(X_1) f_1^T(X_1) + r_1 I$ and similarly for $\hat{\Sigma_{22}}$ where regulation parameters $r_1 > 0$, $r_2 > 0$ are introduced. By negating formula 5.9, L_c can be obtained:

$$\begin{aligned} &L_c = -\frac{1}{N} tr(U^T f_1(X_1;\theta_1) f_2^T(X_2;\theta_2) V) \\ &\text{s.t. same constraints as Equation } 5.9 \end{aligned} \tag{5.10}$$

In the absence of label information, self-supervised loss L_s enforces the representation $f_1(X_1;\theta_1)$ and $f_2(X_2;\theta_2)$ to learn the discriminative information from its high confident clustering predictions. In this section, we formulate L_s for the scenario that only a single view from correlated views is used for clustering, which is commonly used in consensus-based clustering [83], [226]. To further exploit discriminative view-specific features for more comprehensive representations, we propose the DSF fusion scheme with a new soft assignment function in the next section. L_s is illustrated with $h_1 \in f_1(X_1)$ and it is similar for h_2. Self-supervised loss L_s takes two steps including soft assignment and KL divergence minimization. Firstly, soft clustering assignment is calculated by measuring the distance between an embedded point h_1^i and centroids μ_1^j with Students' t-distribution [227]:

$$q_{ij} = \frac{(1 + |h_1^i - \mu_1^j|^2)^{-1}}{\sum_{j'}(1 + |h_1^i - \mu_1^{j'}|^2)^{-1}} \tag{5.11}$$

where centroid $\mu_1^j \in \mathbf{R}^{L*k}$ can be obtained via k-means clustering on the pretrained network. The number of centroid k is pre-defined.

Secondly, KL divergence of the centroid-based probability and an auxiliary target distribution is minimized to refine the representation by learning high confidence predictions. Instead of using a naive delta distribution, we implement Soft Assignment Hardening (SAH) distribution [86] to improve class separation and imbalanced data prediction. The q_{ij} is raised to the second power to push the network to learn from confident prediction and soft assignment is normalized to prevent large clusters from distorting the hidden distribution.

$$p_{ij} = \frac{q_{ij}^2 / \sum_i q_{ij}}{\sum_{j'} q_{ij'}^2 / \sum_i q_{ij'}} \tag{5.12}$$

Lastly, self-supervised loss L_s can be computed as KL divergence between the soft assignment q_{ij} and target distribution p_{ij}:

$$L_s = KL(P||Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}} \tag{5.13}$$

5.2.2 DSF Fusion

View-specific components are often neglected in consensus MVC methods. For example, previous correlation-based MVC only used single view (h_1^i or h_2^i) from correlated space for clustering, but did not consider small portion of view-specific components in each view [83]. To effectively use discriminative consensus and complementary view-specific information for a comprehensive representation, we propose a simple yet effective Deep Serial Feature-level (DSF) Fusion. Detailed reasoning underlying DSF fusion is provided and the performance is extensively validated in the Experiments. DSF fusion can be formularized as:

$$z^i = \begin{pmatrix} h_1^i \\ h_2^i \end{pmatrix} = \begin{pmatrix} f_1(X_1^i;\theta_1) \\ f_2(X_2^i;\theta_2) \end{pmatrix} \in \mathbf{R}^{2L*N} \tag{5.14}$$

where $h_1^i \in H_1$ and $h_2^i \in H_2$ are representations for different views of the same sample. After this feature-level fusion, Equation 5.11 in L_s for different views H_1 and H_2 can be integrated into one formula:

$$q'_{ij} = \frac{(1 + |z^i - \mu^j|^2)^{-1}}{\sum_{j'}(1 + |z^i - \mu^{j'}|^2)^{-1}} \tag{5.15}$$

where centroids $\mu^j = \left(\mu_1^j, \mu_2^j\right)^T$. This equation replaces Equation 5.11 and used by Equation 5.12 and 5.13 to calculate L_s.

The reasoning underlying the proposed DSF fusion can be elaborated from two perspectives, including fusion of consensus information and fusion of view-specific information. Firstly, previous studies found that fusion of correlated components from linear CCA led to elevation of clustering performance [225], [228]. Specifically, Sun *et al.* proposed two fusion schemes, namely Serial Feature Fusion (SFF) and Parallel Feature Fusion (PFF) for fusion of correlated representations and achieved superior performance on image recognition tasks [225], [228]. SFF is based union-vector while PFF is based on a complex vector [229], which are summarized below:

$$SFF = \begin{pmatrix} w_1^T X_1 \\ w_2^T X_2 \end{pmatrix} = \begin{pmatrix} w_1 & 0 \\ 0 & w_2 \end{pmatrix}^T \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \tag{5.16}$$

Algorithm 2 Optimization of SDC-MVC

```
Input: Multi-view data $X_1$, $X_2$, loss weight $\lambda$,
number of cluster $k$
Initialize:
$f_1$, $f_2 \leftarrow$ pre-trained DCCA model
$Z = [f_1(X_1), f_2(X_2)] \leftarrow$ information fusion
$\mu \leftarrow$ cluster centroids from k-means clustering
Optimize:
while not converged do
    $X_1'$, $X_2' \leftarrow$ a random batch of multi-view data
    $L_c \leftarrow$ via Equation 5.10
    $L_s \leftarrow$ via Equation 5.13
    Gradient descent on $\lambda L_c + L_s$, update $f_1, f_2$
end while
Output: $f_1, f_2$
```

$$PFF = w_1^T X_1 + w_2^T X_2 = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}^T \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \tag{5.17}$$

Inspired by this work, we extend SFF to fuse correlated deep features from the multi-view representation module. SFF is chosen instead of PFF as it achieved better performance than PFF in previous work [228].

Secondly, the fusion of view-specific components contributes to the clustering performance in MVC algorithms such as multi-kernel clustering [230], because view-specific information from different views is potentially complementary to each other. In conventional deep CCA, the representation of only one view was used for clustering, where the view-specific information in the other view was wasted. In our SDC-MVC, although multi-view representations h_1^i and h_2^i are highly correlated, view-specific components still exist and are potentially discriminative because clustering-related information is boosted by self-supervised loss L_s. Thus, the fusion of the discriminative view-specific information can assist with the clustering process and boosting the performance. To conclude, our proposed DSF fusion effectively handles the fusion of consensus components and view-specific components, with its effectiveness towards clustering explicitly verified by experiments.

5.2.3 SDC-MVC Optimization

To calculate the gradient of SDC loss with respect to representation network parameters $\frac{dL}{d\theta_1}$, $\frac{dL}{d\theta_2}$, we can first compute the gradient with respect to the output of representation networks $\frac{dL}{dh_1}$ and $\frac{dL}{dh_2}$, and then pass it back through the DNN using backpropagation. SDC loss gradient can be expressed as the sum of weighted two sub-loss gradients:

$$\frac{dL}{dh_1} = \lambda \frac{dL_c}{dh_1} + \frac{dL_s}{dh_1} \tag{5.18}$$

The gradient of Correlational Loss L_c with respect to h_1 can be computed as:

$$\begin{aligned} \frac{dL_c}{dh_1} &= \frac{1}{N}(2 * \nabla_{11} h_1 + \nabla_{12} f_2) \\ \text{where} \quad \nabla_{12} &= \hat{\Sigma_{11}}^{-1/2} U V' \hat{\Sigma_{22}}^{-1/2} \\ \nabla_{11} &= -\frac{1}{2} \hat{\Sigma_{11}}^{-1/2} U D U' \hat{\Sigma_{11}}^{-1/2} \end{aligned} \tag{5.19}$$

The gradient of Self-supervised Loss L_s with respect to h_1 can be computed as:

$$\begin{aligned} \frac{dL_s}{dh_1} =& 2 * \sum_j (1 + |h_1^i - \mu_1^j|^2)^{-1} \\ & * (p_{ij} - q_{ij})(h_1^i - \mu_1^j) \end{aligned} \tag{5.20}$$

The gradient of SDC loss with respect to H_2 can be calculated similarly. The entire optimization process is outlined in Algorithm 1.

5.3 Interpretation on Deep Correlational Fusion

To address the challenge of interpreting complex nonlinear association in the deep multimodal fusion model, we propose a cross-modal association (CA) score to provide a model-level interpretation on feature importance of multimodal inputs towards the highly correlated association. As shown in Fig 5.1, it is achieved by 1) establishing a CA matrix C_x, C_y to represent the association between highly correlated variates Z_x, Z_y with respect to their feature inputs X, Y respectively, and 2) summarizing the CA score for each individual features in X,

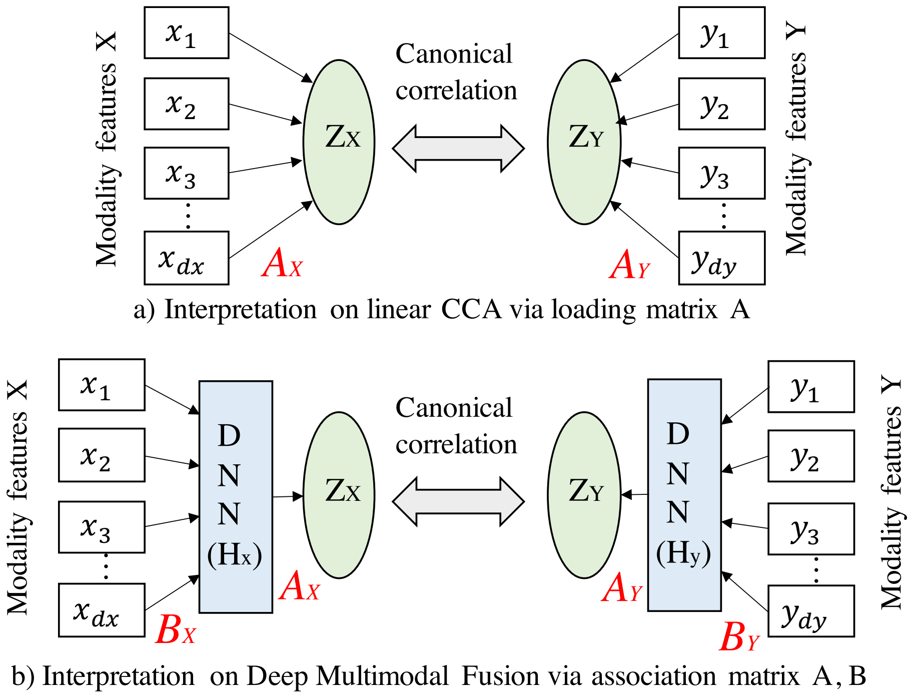


FIGURE 5.3. Interpreting multimodal correlational fusion models.

Y according to their contribution towards canonical variates Z_x, Z_y based on CA matrices and strength of cross-modal correlation between Z_x, Z_y. The difference between our interpretation module for non-linear association and previous methods for linear association is illustrated in Figure 5.3.

The computation of CA matrix C_x includes three steps. The first step is to compute CCA Association matrix $A_x \in \mathbb{R}^{d_o * d_o}$ of canonical variates Z_x with respect to multimodal representation H_x. The second step is to calculate Deep Association matrix $B_x \in \mathbb{R}^{d_x * d_o}$ of multimodal representation H_x with respect to input X. Lastly, CA matrix $C_x \in \mathbb{R}^{d_x * d_o}$ of canonical variates Z_x with respect to input X can be obtained based on CCA Association matrix A_x and Deep Association matrix B_x:

$$C_x = B_x A_x \tag{5.21}$$

To obtain CCA Association matrix A_x, we adopt a correlation-based approach by calculating canonical loading. Specifically, the canonical loading (also known as structure coefficients) measures bivariate Pearson correlation between the observed variable (H_x in our case) and canonical variates Z_x. The canonical loading reflects the variance that H_x shared with canonical variate Z_x, thus can be considered as relative contribution of each variable H_x in towards the canonical variate Z_x. The computation of A_x is formulated as:

$$A_x = Corr(Z_x, H_x) \tag{5.22}$$

The computation of Deep Association matrix B_x is based on integrated gradients [29]. Different from conventional integrated gradients computing contribution towards prediction outputs, we apply it directly to the output of each neuron in the output layer of $f_x(X)$ to assess the contribution towards multimodal representations H_x. The association of j^{th} neuron with respect to inputs X can be formulated as:

$$IG_j(X) = \sum_{i=0}^{n}(X_i - X_i') * \int_{\alpha=0}^{1} \frac{\partial f_x^j(X' + \alpha(X - X'))}{\partial X_i} d\alpha \tag{5.23}$$

where $X_i \in X$ is a sample, and α is a scaling coefficient. By concatenating IG_j for all neurons in the output layers of $f_x(X)$, we can obtain Deep Association matrix B_x.

$$B_x = [IG_1^T, IG_2^T, ..., IG_{d_o}^T] \tag{5.24}$$

Lastly, we compute the proposed CA scores using the obtained CA matrix C_x (from A_x and B_x). Inspired by the squared canonical loading in linear CCA, we compute importance scores for k^{th} input features considering both squared association value in CA matrix C_x as well as the strength of canonical variates correlated with each other:

$$score_k = \sum_{j=1}^{d_o} C_{x;jk}^2 * |r_j| \tag{5.25}$$

where r_j is Pearson correlation between j^{th} pairs of canonical variables. Similarly, we can compute CA scores for inputs Y.

5.4 Datasets and Implementation

5.4.1 Datasets

The supervised deep fusion network was validated on the diagnostic differentiation of MS and NMO, which are the most common causes of neurological disability in young people [184]. The clinical difficulty of differential diagnosis of these two diseases arises from their similar lesion appearance on MRI and overlapped clinical symptoms [143]. A dataset of 94 patients (including 66 MS and 28 NMO patients) was collected from Xuanwu Hospital, Capital Medical University. Each patient had two multi-parametric MRI imaging including T2 and T1-MPR, and four clinical factors including age, gender, disease duration and EDSS scores. For each MRI image modality, we adopted ROI-based radiomics quantification to extract 1118 features including intensity, texture, filter-based features. The imbalanced issue in the dataset was handled by Adaptive Synthetic (ADASYN) oversampling algorithm [231].

The unsupervised deep fusion framework was validated on three public multi-view datasets, including two relatively large datasets and one small dataset:

- **Noisy MNIST** [1] is a two-view challenging version of MNIST image dataset [232]. It was proposed by [83] and consisted of a rotation view and a noisy view of 70k grey-scale 28*28 digit images. The rotation view was obtained by randomly rotating the images with an angle from $[-\pi/4, \pi/4]$, after the pixels were rescaled to [-1, 1]. The noisy view was generated by selecting a random image of the same identity and then adding independent random noise sampled from [-1, 1] to it. The data was split to 50k/10k for training and testing, respectively.
- **XRMB Vowel** [2] is an audio dataset consisting of 273 acoustic features and 112 articulatory features, which has been pre-processed by [83]. It is a subset of XRMB dataset [233] containing all vowel utterances because most baselines were inefficient

[1] https://www2.cs.uic.edu/~vnoroozi/noisy-mnist/

[2] https://ttic.uchicago.edu/~klivescu/XRMB_data/full/README

to cluster large-scale datasets. The acoustic view was computed with melfrequency cepstral coefficients (MFCCs) over a 25 ms window while the articulatory view measures the horizontal and vertical displacement of eight pallets placed on the speakers' jaw, lips and tongue. The frames were split to around 100k/12k for training and testing, respectively.

- **Yale Face** [3] is a small face-recognition dataset, consisting of 165 images of 15 subjects. Each subject has 11 different face images, such as with/without glasses. We extracted two feature views from the original images: 3304 Local Binary Patterns (LBP) features and 6750 Gabor features. The data was split into 120/45 for training and testing, respectively.

5.4.2 Implementations

Supervised deep Fusion framework. The effectiveness of the proposed DMFusion framework was validated on three specific fusion tasks: 1) T2 imaging with clinical factors, 2) T1-MPR imaging with clinical factors, and 3) T2 imaging with T1-MPR imaging. The diagnostic performance of the fused representation was assessed with 10-fold cross-validation via Support Vector Machine (SVM). Four evaluation metrics, including AUC, Accuracy (ACC), Sensitivity (SEN) and Specificity (SPE) were used and reported as mean values with standard deviation. We compared the experimental results of the DMFusion with six state-of-the-art multimodal methods, including two radiomics methods (Lasso and RFS), two multi-branch deep learning methods (MAE [234], DCCAE [83]) and two CCA-based methods (Linear CCA and Deep CCA [82]). To verify the effectiveness of DMFusion Layer, we compared the fused representation with modality-specific representations.

In terms of hyperparameters, the parameters C in SVM for all methods were selected from [0.01, 0.1, 1]. The dimension of the output layer for all methods was set as four. For deep learning-based methods, the dimensionality of the intermediate layers was tuned from [256, 512, 1024]. Loss-balancing parameters λ_1, λ_2 and λ_3 for our method and DCCAE were tuned

[3] https://vismod.media.mit.edu/vismod/classes/mas622-00/datasets/

from [0.01, 0.1, 1]. The deep networks were trained with RMSprop method with 1e-3 learning rate and 20 epochs. L2 penalty was applied to weights of the network to reduce the risk of overfitting.

Unsupervised deep fusion framework. In this section, the implementation of SDC-MVC for three datasets is introduced. The whole network was deployed with Pytorch framework and was run with one GPU GeForce RTX 2080 Ti. In a fully unsupervised setting, it was not applicable to determine hyper-parameters by tuning with the supervision of labels, so we used commonly reported parameters for each dataset.

The multi-view representation module was implemented with one MLP neural network per view. Each MLP consisted of three hidden layers with 1024, 1500 and 2000 neurons for Noisy MNIST, XRMB Vowels and Yale Face datasets respectively. Activation layer was implemented by either Sigmoid units or ReLU depending on the commonly used setting for the dataset [83]. The multi-view representation was pretrained with DCCA objective to generate initialized clustering centroids with the help of k-means clustering. In terms of optimization, Stochastic Gradient Descent (SGD) with the learning rate ranging from 1e-4 to 5e-3 was used with a momentum of 0.9. The batch size was set as 256 for Noisy MNIST, XRMB Vowels following Xie's implementation, and full-batch optimization was performed on Yale because it is a small dataset. The risk of overfitting was handled by weight decay. As for comparison methods, the results for Noisy MNIST were acquired from the authors' reported performance while the methods for the other two datasets were implemented with the authors' codes.

5.4.3 Evaluations

For a comprehensive evaluation, clustering performance was assessed with three standard clustering metrics: clustering accuracy (ACC), normalized mutual information (NMI), and homogeneity (HOM). All metrics range from 0 to 1 and a higher value indicates better performance. Clustering ACC is measured by finding best matching m between true labels l_i

and predicted clustering labels c_i using Hungarian algorithm [235]. It is defined as:

$$ACC = \max_{m} \frac{\sum_{i=1}^{n} \mathbf{1}\{l_i = m(c_i)\}}{n} \tag{5.26}$$

NMI is a normalized version of the MI, measuring the similarity between two clustering, which is defined as:

$$NMI = \frac{I(l;c)}{max\{H(l), H(c)\}} \tag{5.27}$$

where $I(l,c)$ denotes mutual information between true label l and predicted clustering label c while H represents entropy. HOM measures a desirable objective of clustering assignment: each cluster only contains members of a single class, which is defined by:

$$HOM = 1 - H(l|c)H(l) \tag{5.28}$$

where $H(l|c)$ denotes the conditional entropy of the true labels given predicted clustering labels while $H(l)$ denotes the entropy of true labels.

5.5 Experimental Results and Discussion

5.5.1 Diagnostic Results of Supervised Fusion.

Table 5.1-5.2 shows the fusion results of imaging (T2 MRI and T1-MPR MRI) and non-imaging (clinical factors), while Table 5.3 shows the results of imaging-imaging fusion of two MRI sequences.

Fusion of T2-MRI and Non-imaging. In terms of fusing T2 imaging features and clinical factors, Table 5.1 shows that our DMFusion framework outperformed all other methods in terms of ACC, AUC and SEN. Particularly, our method outnumbered the state-of-the-art DCCAE method by a large margin in terms of AUC (0.82 vs 0.76) and SEN (0.75 vs 0.60). As for fusing T1-MPR and clinical factors, Table 5.2 demonstrates that our fusion framework outperformed all the other methods on ACC, AUC and SPE. The outstanding diagnostic

TABLE 5.1. Diagnostic results of the fusion of T2 images and clinical factors.

	ACC	AUC	SEN	SPE
Concat+Lasso	0.66 ± 0.11	0.73 ± 0.08	0.40 ± 0.21	0.77 ± 0.16
Concat+RFS	0.64 ± 0.06	0.71 ± 0.08	0.51 ± 0.23	0.70 ± 0.17
CCA	0.60 ± 0.06	0.57 ± 0.12	0.43 ± 0.18	0.67 ± 0.06
DCCA	0.73 ± 0.06	0.81 ± 0.07	0.60 ± 0.28	0.79 ± 0.13
MAE	0.69 ± 0.23	0.70 ± 0.23	0.64 ± 0.25	0.72 ± 0.28
DCCAE	0.74 ± 0.09	0.76 ± 0.09	0.60 ± 0.23	**0.81** ± **0.14**
DMFusion_T2	0.69 ± 0.04	0.82 ± 0.08	0.61 ± 0.26	0.73 ± 0.09
DMFusion_clinical	0.69 ± 0.08	0.72 ± 0.09	0.65 ± 0.20	0.72 ± 0.17
DMFusion	**0.77** ± **0.12**	**0.82** ± **0.09**	**0.75** ± **0.10**	0.78 ± 0.18

performance validated that the proposed DMFusion loss successfully extracted discriminative multimodal information.

Fusion of T1-MRI and non-imaging. By comparing the performance of variants of DMFusion in both fusion tasks, the effectiveness of DMFusion Layer was experimentally verified. In the first task, Table 5.1 shows although AUC of DMFusion_T2 was the same as DMFusion, fusion of T2 and clinical representations significantly boosted all other metrics (e.g., ACC from 0.69 to 0.77). For the second task, DMFusion outperformed its modality-specific representations in terms of ACC, AUC and SPE.

TABLE 5.2. Diagnostic results of the fusion of T1-MPR images and clinical factors.

	ACC	AUC	SEN	SPE
Concat+Lasso	0.71 ± 0.08	0.74 ± 0.13	0.61 ± 0.24	0.76 ± 0.20
Concat+RFS	0.64 ± 0.06	0.66 ± 0.17	0.49 ± 0.37	0.70 ± 0.16
CCA	0.69 ± 0.08	0.66 ± 0.18	0.69 ± 0.17	0.70 ± 0.09
DCCA	0.60 ± 0.18	0.73 ± 0.09	0.57 ± 0.35	0.62 ± 0.37
MAE	0.67 ± 0.16	0.71 ± 0.12	**0.75** ± **0.13**	0.64 ± 0.23
DCCAE	0.70 ± 0.15	0.77 ± 0.11	0.73 ± 0.22	0.70 ± 0.26
DMFusion_MPR	0.63 ± 0.11	0.67 ± 0.18	0.57 ± 0.34	0.65 ± 0.15
DMFusion_clinical	0.69 ± 0.11	0.69 ± 0.07	0.61 ± 0.24	0.73 ± 0.07
DMFusion	**0.73** ± **0.10**	**0.79** ± **0.08**	0.59 ± 0.18	**0.79** ± **0.15**

TABLE 5.3. Diagnostic results of the fusion of T1-MPR images and clinical factors.

	ACC	AUC	SEN	SPE
Concat+Lasso	0.72 ± 0.08	0.71 ± 0.13	0.54 ± 0.28	**0.81 ± 0.17**
Concat+RFS	0.75 ± 0.05	0.74 ± 0.16	0.67 ± 0.26	0.77 ± 0.17
CCA	0.65 ± 0.12	0.70 ± 0.16	0.49 ± 0.30	0.72 ± 0.18
DCCA	0.73 ± 0.06	0.85 ± 0.09	0.72 ± 0.31	0.75 ± 0.16
MAE	0.65 ± 0.18	0.63 ± 0.27	0.47 ± 0.37	0.74 ± 0.31
DCCAE	0.67 ± 0.13	0.86 ± 0.12	0.57 ± 0.37	0.72 ± 0.28
DMFusion	**0.75 ± 0.04**	**0.88 ± 0.06**	**0.79 ± 0.24**	0.75 ± 0.14

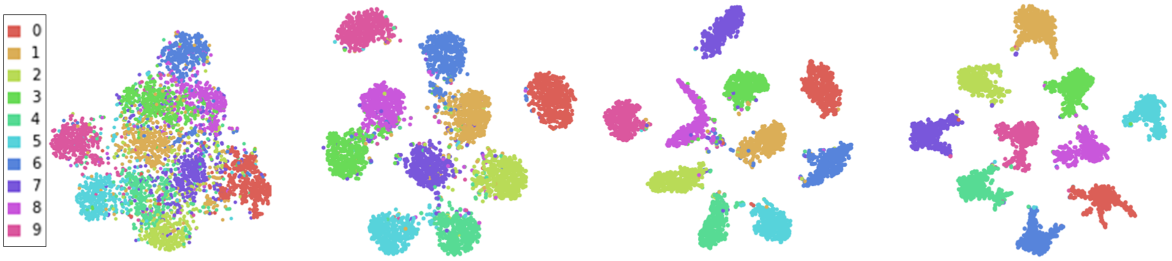
Imaging-imaging fusion of T2 and T1. In addition to the fusion of imaging and non-imaging data, our model also successfully handled the feature fusion between different imaging modalities. Specifically, Table 5.3 shows that our DMFusion framework achieved ACC 0.75 ± 0.04, AUC 0.88 ± 0.06, SEN 0.79 ± 0.24 and SPE 0.75 ± 0.14 on fusing T2 images and T1-MPR images, outperforming all other methods on ACC, AUC and SEN. Besides, our model shows excellent stability in different train-test split during the 10-fold cross-validation, which is supported by the lowest standard deviation in all evaluation metrics. Among the three fusion tasks, the highest AUC (0.88 ± 0.06) for the differential diagnosis of MS and NMO was achieved by our DMFusion framework by fusing T2 and T1-MPR images.

5.5.2 Clustering Results of Unsupervised Fusion

Compared with 6 SOTA correlational MVC methods, Table 5.4 shows the quantitative clustering results on the three datasets. The results demonstrate that SDC-MVC outperforms the compared on all evaluation metrics. In specific, our model achieved ACC 98.0% on **Noisy MNIST**, outperforming all the SOTA correlational MVC methods, even a supervised VCCA method. In addition, our methods achieved a large margin on **XRMB Vowel** in terms of NMI and HOM. Lastly, the competitive performance on **Yale Face** shows SDC-MVC handles not only large datasets but also small datasets.

	Noisy MNIST		XRMB Vowels			Yale Faces		
Method	ACC	NMI	ACC	NMI	HOM	ACC	NMI	HOM
CCA	72.9*	56.0*	51.2	36.7	37.4	57.8	74.2	72.5
FKCCA	94.7*	87.3*	61.1	55.6	58.7	55.5	72.6	71.4
MVAE	64.0*	69.0*	56.4	49.3	58.7	53.3	71.3	69.3
DCCA	97.0*	92.0*	66.7	54.2	46.7	60.0	74.4	72.2
DCCAE	97.5*	93.4*	73.5	55.0	53.7	62.2	76.1	75.2
VCCA+SVM	97.6$^+$	-	-	-	-	-	-	-
SDC-MVC_view1	95.9	89.7	51.2	41.3	40.9	62.2	75.9	78.4
SDC-MVC_view2	83.1	79.6	68.8	52.9	51.9	42.2	64.3	68.7
SDC-MVC_net	**98.0**	**94.6**	**75.8**	**68.8**	**71.3**	**64.4**	**77.6**	**80.1**

TABLE 5.4. Performance of unsupervised MVC algorithms on three public datasets.



(a) CCA (b) KCCA (c) DCCAE (d) SDC-MVC

FIGURE 5.4. t-SNE plot of multi-view representation acquired by different multi-view learning methods on the testing set of noisy MNIST dataset.

To visualize the effectiveness of SDC-MVC, we ploted the feature embeddings with t-SNE method and showed the images with the most confident clustering assignment on the test set of Noisy MNIST. In t-SNE plots (Figure 5.4), SDC-MVC feature embeddings achieved better homogeneity compared with other MVC methods.

Contribution of self-supervised loss. To verify the contribution of self-supervised loss L_s, we visualized the relationship between the assignment confidence and its influence on refining the representation in the backward path. Figure 5.5 shows that samples with confidence [0.5, 0.8] had more influence on the refinement of representation. Samples with confidence over 0.8 had little space to further contribute to the refining process, while samples with confidence

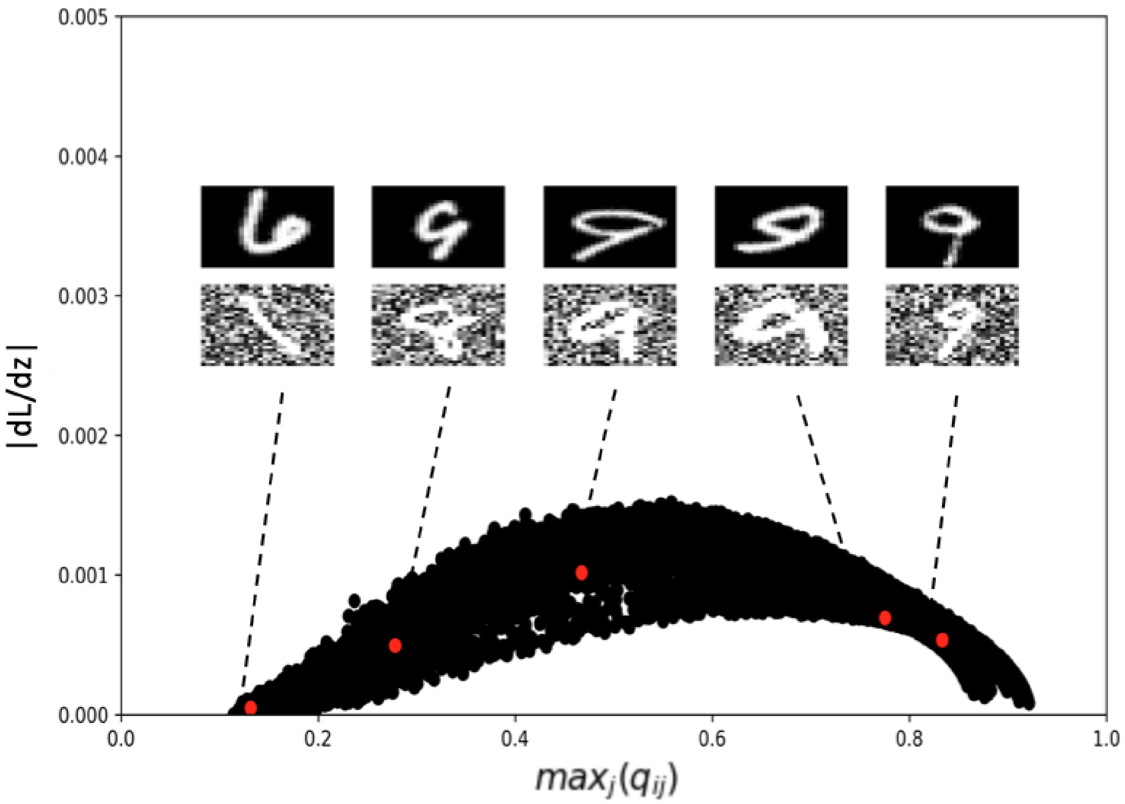
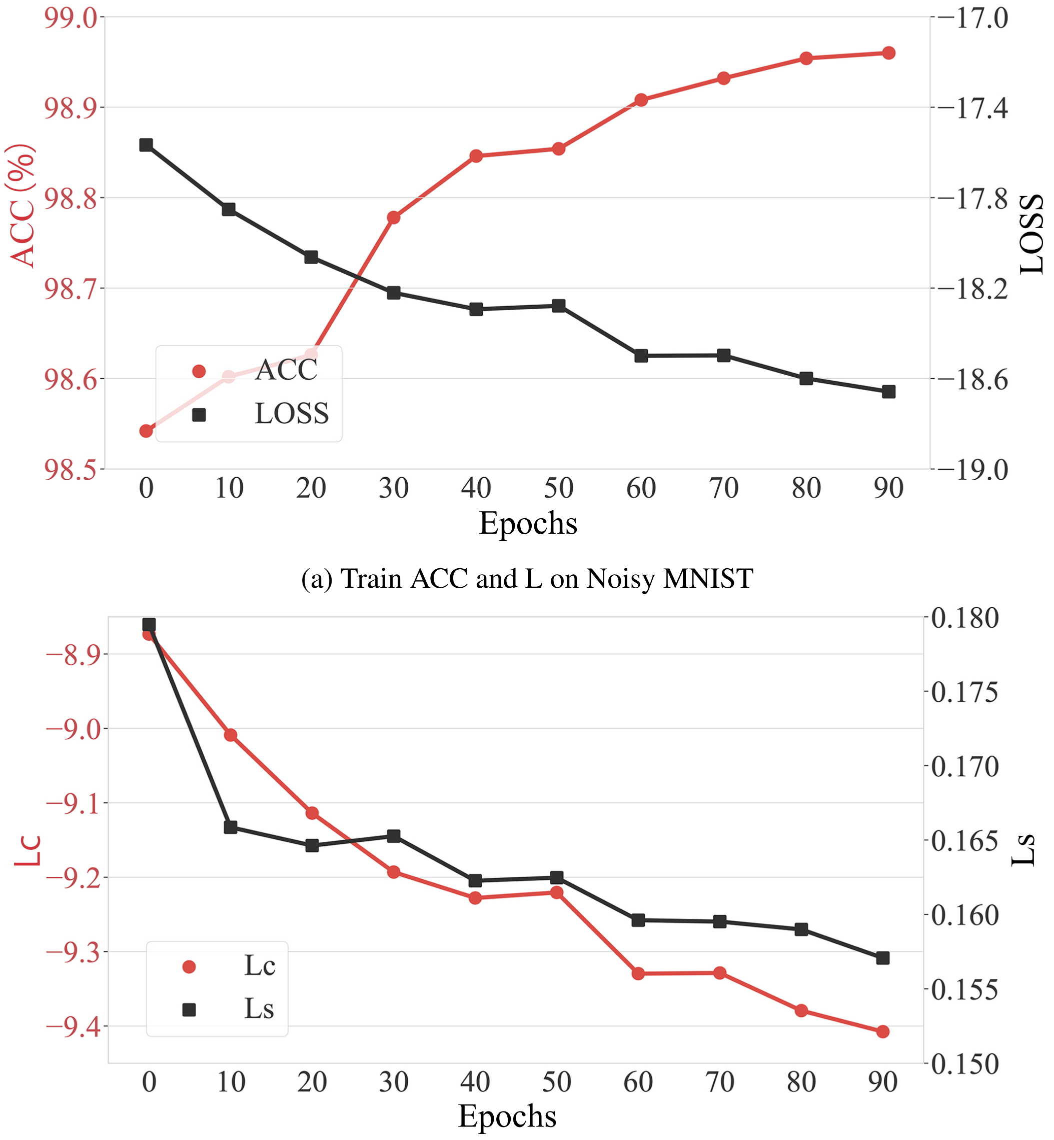


FIGURE 5.5. Hypothesis validation of Self-supervised Loss. Figure shows the impact of different assignment confidence on the gradient of loss L with respect to representation z. Displayed sample images were selected from different confident range $[m, m + 0.2]$ where $m \in [0, 0.2, 0.4, 0.6, 0.8]$.

below 0.5 are likely to be wrongly classified as shown in the images of digit number in Figure 5.5.

Contribution of DSF fusion. The effectiveness of the DSF fusion was verified experimentally, by comparing the performance of SDC-MVC, SDC-MVC_view1, SDC-MVC_view2. As shown in Table 5.4, SDC-MVC outperforms both SDC-MVC_view1 and SDC-MVC_view2 on three datasets in terms of all evaluation metrics. * result in Table 5.4 was acquired from [83], while $^+$ result from [236]. This indicates that the proposed DSF fusion effectively integrated discriminative consensus and view-specific information.



(a) Train ACC and L on Noisy MNIST

(b) L_c and L_s on Noisy MNIST

FIGURE 5.6. Loss convergence analysis. (a) Convergence analysis of train ACC (left vertical axis) and loss L (right vertical axis). (b) Convergence analysis of L_c (left vertical axis) and L_s (right vertical axis) in loss L. The horizontal axis for both subfigures is the number of epochs.

Analysis of parameter and convergence. In this section, we present a convergence analysis followed by parameter analysis on the influence of the loss-balancing parameter λ. In terms of convergence analysis, Figure 5.6a illustrates that the training accuracy of clustering growed steadily with the decrease of the SDC loss. The loss declines with a relatively higher speed until 70 epochs and then starts to approach convergence. To dive deeper into our proposed

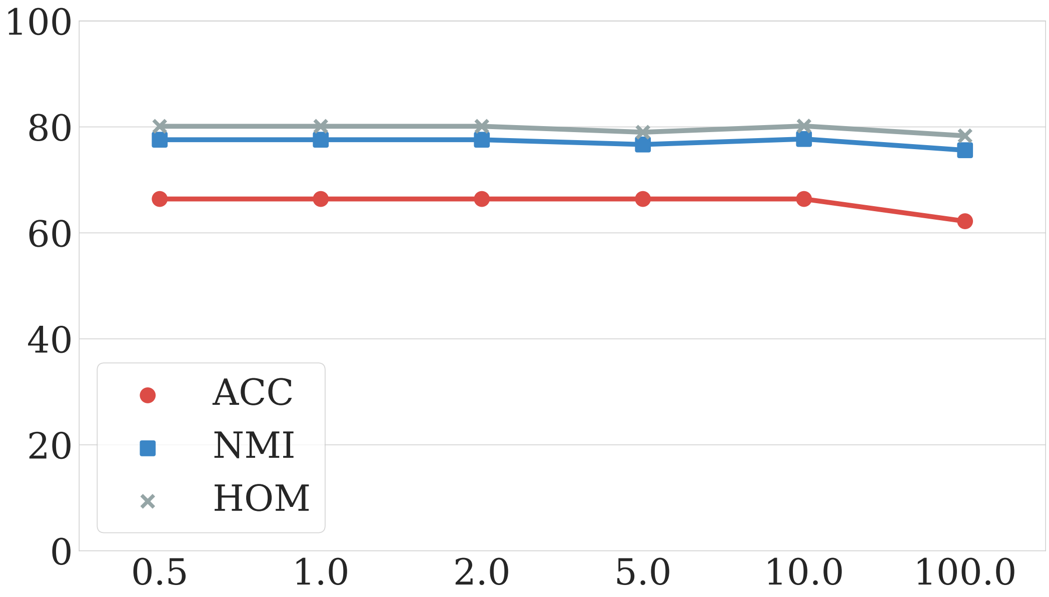


FIGURE 5.7. Parameter analysis of loss balancing parameter λ on Yale.

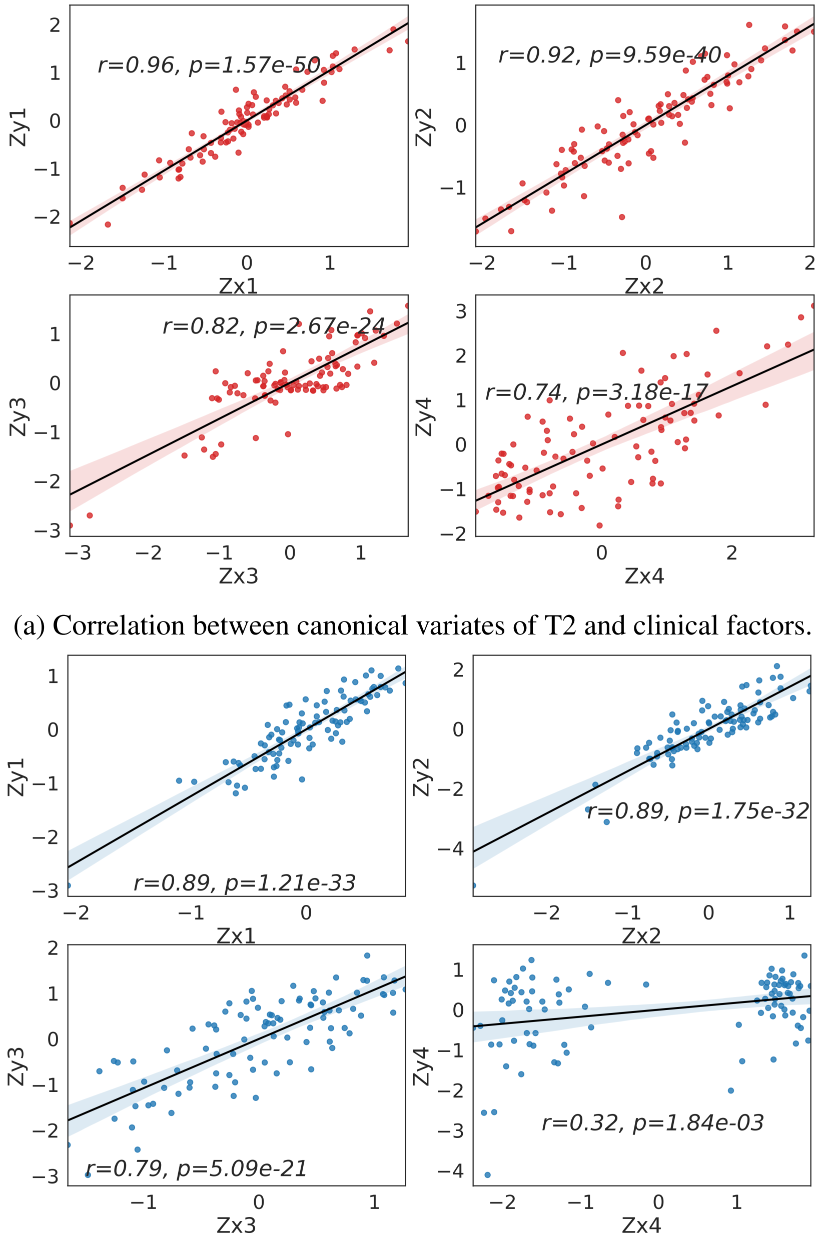
SDC loss L, we further investigated the convergence of L_c and L_s inside L. Figure 5.6b shows that both elements of loss L converged gracefully during the training process.

The parameter analysis of λ was conducted on the dataset **Yale Face**. Figure 5.7 shows that NMI and HOM remain relatively stable until λ reaches 100 and ACC remains almost unchanged from 0 to 10.

5.5.3 Interpretation on Deep Multimodal Fusion

We illustrate the interpretation results of our DMFnet based on two imaging-non-imaging fusion tasks. We firstly evaluated the value of cross-modal correlation that learned by our DMFnet, and then uncover the cross-modal association of different modalities towards the correlated space using the proposed CA scores and CA matrices.

Interpreting cross-modal correlation. Figure 5.8a visualized cross-modal Pearson correlation between deep canonical variates projected from T2 images and clinical factors. All four pairs of deep canonical variates achieved statistically significant correlation ($p < 0.05$)



(a) Correlation between canonical variates of T2 and clinical factors.

(b) Correlation between canonical variates of T1-MPR and clinical factors.

FIGURE 5.8. Cross-modal correlation between imaging and non-imaging.

with high correlation coefficients ranging from 0.74 to 0.96. Similarly, all pairs of canonical variates projected from T1-MPR images and clinical factors achieved statistical significant correlation ($p < 0.05$), in which three of four pairs with correlation coefficients higher than

0.79. The results indicate that our DMF successfully captured the consensus association between different modalities.

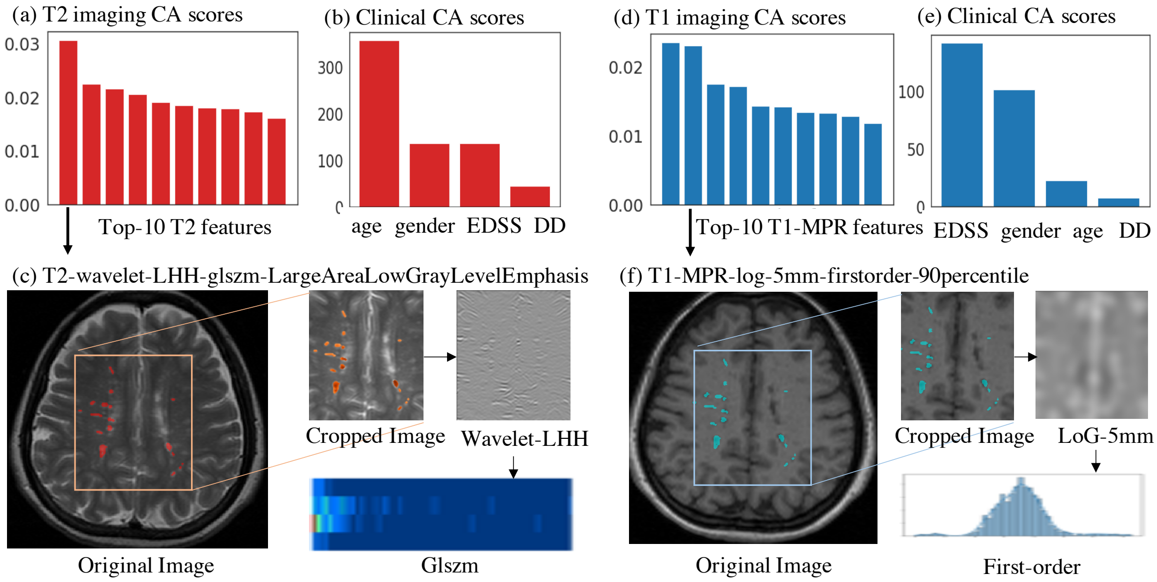
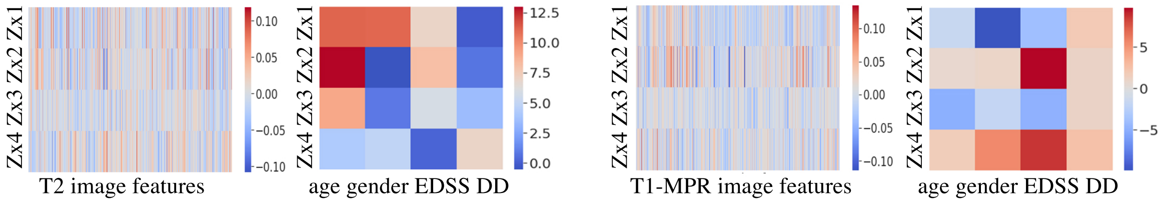


FIGURE 5.9. CA scores for input imaging and non-imaging features in DM-Fusion network.

Interpreting cross-modal association scores. Cross-modal association scores were computed to interpret how multimodal input features contributed to the consensus associations. Figure 5.9a-5.9b illustrate the CA scores of top-10 T2 images features and clinical factors towards the correlated space respectively. In specific, the imaging feature T2-wavelet-LHH-glszm-LargeAreaLowGrayLevelEmphasis and clinical factor age contributed most towards the nonlinear correlational consensus. Figure 5.9c visualizes the extraction of the wavelet features to aid with interpretation. Similarly, Figure 5.9d-5.9e show that one T1-wavelet feature and one T1-log feature contributed most towards nonlinear correlation with EDSS and gender. We also visualizes the extraction of imaging T1-log-feature in Figure 5.9f to further facilitate the understanding of fusion mechanism.

Interpreting cross-modal association matrix. To gain insights into CA scores, we further plotted the heatmap of CA matrix to reveal the mechanism of the proposed CA matrix. Figure 5.10a shows the association between all T2 image features, clinical factors, and their



(a) CA matrix for fusion of T2 and clinical factors (b) CA matrix for fusion of T1-MPR and clinical factors

FIGURE 5.10. CA matrices for input imaging and non-imaging features in DMFusion network.

corresponding canonical variates. From the clinical CA matrix, we found that the high CA score of age feature was dominantly attributed to the association of age and zy2 canonical variate. Figure 5.10b shows the high CA score of EDSS was primarily contributed by its association with zy2 and zy4 canonical variables.

5.6 Chapter Summary

In information-level, we propose a new Interpretable Deep Correlational Fusion framework to optimize multimodal representations with inter-modal consensus information and intra-modal discriminative information in both supervised and unsupervised settings. To interpret the nonlinear assoication of input modalities, we propose a new CA score to quantify the feature importance towards correlated association in deep networks. We validated our framework on the differential diagnosis task of two clinically challenging demyelinating diseases (MS vs NMO) as well as on three public datasets, outperforming six state-of-the-art methods. The results show that our framework can assist clinicians with more accurate diagnostic decisions and uncovering cross-modal associations during exploring the disease mechanism.

CHAPTER 6

Knowledge-level Fusion: Dynamic Topology Analysis Framework with Domain Knowledge for Spatial Lesion Pattern on MRI

Knowledge distillation from multi-focused regions is a challenging topic, because each focused regions may present heterogeneous information hindering the fusion process. Quantitatively analyzing the spatial patterns of multifocal lesions (an example of multi-focused regions) on clinical MRI is an important step towards a better understanding of the disease and for precision medicine. However, it is which is yet to be properly explored by feature engineering and deep learning methods. Network science addresses this issue by explicitly modeling the inter-lesion topology. However, the construction of the informative graph with optimal edge sparsity and quantification of community graph structures are the current challenges in network science.

In this chapter, we address these challenges with a novel Dynamic Topology Analysis (DTA) framework on the basis of persistent homology, aiming to investigate the predictive values of global geometry and local clusters of multifocal lesions. Firstly, Dynamic Hierarchical Network is proposed to construct informative global and community-level topology over multi-scale network from sparse to dense. Multi-scale global topology is constructed with a nested sequence of Rips complexes, from which a new K-simplex Filtration is designed to generate a higher-level topological abstraction for community identification based on the connectivity of k-simplices in the Rips Complex. Secondly, to quantify multi-scale community structures, we design a new Decomposed Community Persistence algorithm to track the dynamic evolution of communities, and then summarize the evolutionary communities incorporated with a customizable descriptor. The quantified community features are encapsulated with global geometric invariants for topological pattern analysis.

This chapter is organized as follows: Firstly, more detailed background description on the fusion of multi-focal lesions and the overall DTA framework are introduced in Section 6.1. Then, we present each component of Dynamic Topology Analysis framework in different sections, including Dynamic Hierarchical Network Construction (Section 6.2), Dynamic Topology Quantification (Section 6.3)m and Topological Pattern Analysis (Section 6.4). The implementations and results are summarized in Section 6.5-6.6 and discussed in Section 6.7.

6.1 Problem Description and Overall Framework

Clinical challenges. MS is a typical and the most prevalent multifocal demyelinating disease in CNS, affecting 2.3 million young people globally [237]. This currently incurable disease [238] causes severe, non-traumatic and enduring physical and cognitive disability in patients, which is pathologically characterized by multiple white matter lesions. MRI is a fundamental imaging technique for the identification of demyelinating multifocal lesions, supporting clinical diagnosis, and monitoring the progression of MS. However, there is a long-standing discrepancy or the clinico-radiological paradox [239], [240] between common MRI markers (such as lesion volume) and clinical disability. This paradox may arise from 1) the disparate spatial pattern of lesions in CNS across patients [241] and 2) pathological heterogeneity of lesions in the same patient [242]. More recently, clinical research suggests the spatial distribution of MS lesions have neuropathologic [243], diagnostic [175], [244] and prognostic [245] associations. Thereby, a comprehensive understanding of the imaging features and patterns, in particular the spatial relationship of multifocal lesions, may potentially contribute to breaking the paradox. Thus, quantitatively profiling the heterogeneity of multifocal lesions is in demand to better understand the underlying collective pathological process to support objective and more effective clinical decision making.

Importance of topological patterns. Such inter-lesion spatial patterns are clinically observable in multifocal demyelinating brain diseases (such as global lesion cycles or local lesion clusters) and have potential clinical implications [246], [247]. For example, the global cycle pattern (a group of lesions around a hole) could indicate the incidence of normal-appearing white matter lesions inside the circle of lesions, which is however not intuitively visible on conventional MRI used in the clinical routine [248], [249]. For another example, local lesion clusters may indicate the potentially actively growing lesion regions [250], [251]. Thus, to quantitatively characterise global and local lesion topology patterns, a graph-based topological profiling tool for multifocal lesions is in high demand.

Limitation of feature engineering and deep learning. Quantitatively characterizing the imaging features of single lesions by using feature engineering or deep learning, has been widely investigated in more recent years. For multifocal lesions, the current feature engineering methods often combine multiple lesion masks as a single entity before mining the imaging patterns via texture features [8], [252] or clinical MRI markers [253], [254] for diagnostic or prognostic tasks. However, in such a method, the patterns of individual lesions are averaged and therefore the contribution of an individual lesion is neglected. Deep learning provides a data-driven approach to learn potential MRI predictors for multifocal lesions in MS [31], [32]; however, the extracted deep features are generally difficult to interpret, hardly biologically meaningful [18] and usually requires a large amount of data to train the network. More importantly, both feature engineering and deep learning methods neglect the importance of the inter-lesion spatial relationship of multifocal lesions. Thus, a topological profiling tool for multifocal lesions is in high demand for systematic analysis of lesion spatial patterns and yet missing in current feature engineering and deep learning methods.

In this chapter, we aim to comprehensively investigate the spatial patterns of multifocal MS lesions, including the global geometrical structure and local lesion clusters. To simultaneously address two challenges in network science including graph construction and topology quantification, we propose a DTA framework based on persistent homology. Firstly, to bypass the challenge of topology construction with the optimal scale, we propose a Dynamic Hierarchical Network to encode both global-level geometrical structure and community-level lesion proximity in an encapsulated dynamic network. Based on Global-level Network (G-Net) constructed by Rips complex, we propose a new K-simplex Filtration to create a high-level abstraction of G-Net using the connectivity of k-simplices for community identification (Community-level Network). Secondly, to incorporate community heterogeneity into the quantification of dynamic community topology, we propose a novel Decomposed Community Persistence algorithm based on union-find data structure to track the evolutionary communities at fine-grained scales. The tracked dynamic communities incorporated with lesion attributes are then summarized by the designed Adaptive Community Profile that is equipped with a customizable community descriptor. Lastly, the quantified community dynamics is fused with global-level geometrical invariants for subsequent topological pattern analysis and modeling.

The overall framework of DTA is illustrated in Fig 6.1, DTA framework is composed of three modules, including a) Dynamic Hierarchical Network Construction, b) Dynamic Topology Quantification, and c) Topological Pattern Analysis.

6.2 Dynamic Hierarchical Network Construction

Hierarchical Network encodes multi-scale dynamics of both global-level geometrical structure and community-level lesion proximity to capture the informative topology. As illustrated in Figure 6.1a, multifocal lesions on MRI are firstly transformed into a point cloud to construct Global-level Network (G-Net), from which we construct a new Community-level Network (C-Net) for community structure identification.

6.2.1 Point Cloud Generation from Multifocal Lesion Volumes

From multifocal lesion volumes on MRI, the lesion cloud is generated by representing individual volumetric lesions in the lesion mask with their corresponding 3-dimensional

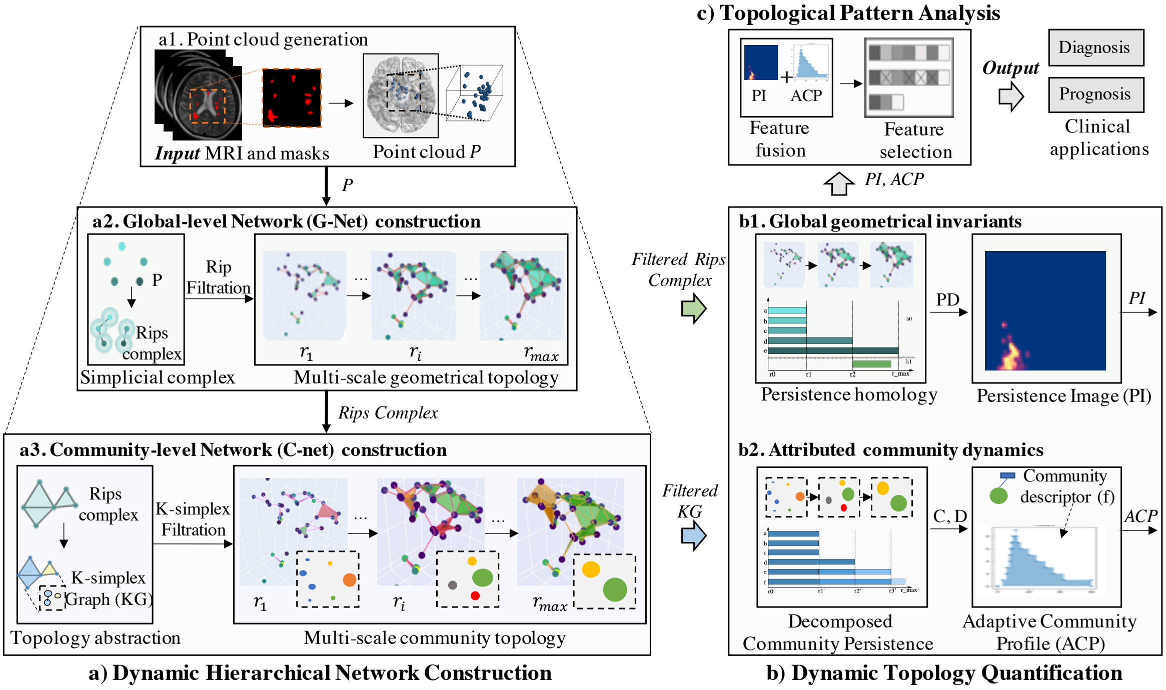

c) Topological Pattern Analysis
PI
ACP
Feature fusion
Feature selection
Output
Diagnosis
Prognosis
Clinical applications
PI, ACP
a1. Point cloud generation
Input MRI and masks
Point cloud P
P
a2. Global-level Network (G-Net) construction
P
Rips complex
Rip Filtration
Simplicial complex
Multi-scale geometrical topology
r_1
r_i
r_{max}
Rips Complex
a3. Community-level Network (C-net) construction
Rips complex
K-simplex Filtration
K-simplex Graph (KG)
Topology abstraction
Multi-scale community topology
Filtered Rips Complex
Filtered KG
b1. Global geometrical invariants
PD
PI
Persistence homology
Persistence Image (PI)
b2. Attributed community dynamics
C, D
Community descriptor (f)
ACP
Decomposed Community Persistence
Adaptive Community Profile (ACP)
a) Dynamic Hierarchical Network Construction
b) Dynamic Topology Quantification

FIGURE 6.1. The proposed DTA framework consists of three major modules. **a) Dynamic Hierarchical Network Construction:** From input MRI and lesion masks, lesion point cloud is generated to construct dynamic Global-level Network using Rips Filtration and construct dynamic Community-level Network using K-simplex Filtration. **b) Dynamic Topology Quantification:** Global-level Network is quantified as Persistence Image based on persistence homology to measure the global geometric invariance of homological objects (such as 1-dimensional holes and 0-dimensional connected components). In the meantime, Community-level Network is quantified as Adaptive Community Profile to measure the statistics about community lesion volume and lesion density during the dynamic graph evolution. **c) Topological Pattern Analysis:** Persistence Image and Adaptive Community Profile are concatenated as a feature pool, from which informative, non-redundant and highly relevant topological features are selected. The selected features are used to construct machine learning models for clinical applications.

centroids, same as in the studies [255], [256]. These lesion centroid points approximate the lesion topology and form a point cloud $P = \{p_i \mid p_i = (x_i, y_i, z_i) \in \mathbb{R}^3, p_i = centroid(V_i)\}$, where $\mathbb{R}^3$ denotes 3D physical space and p_i is the centroid of the corresponding lesion volume V_i. To convert the volumetric lesions into point cloud, we firstly separate individual lesions using disconnected component labeling based on pixel connectivity [257], and then compute the centroid of each corresponding lesion as the corresponding point of the cloud. As the

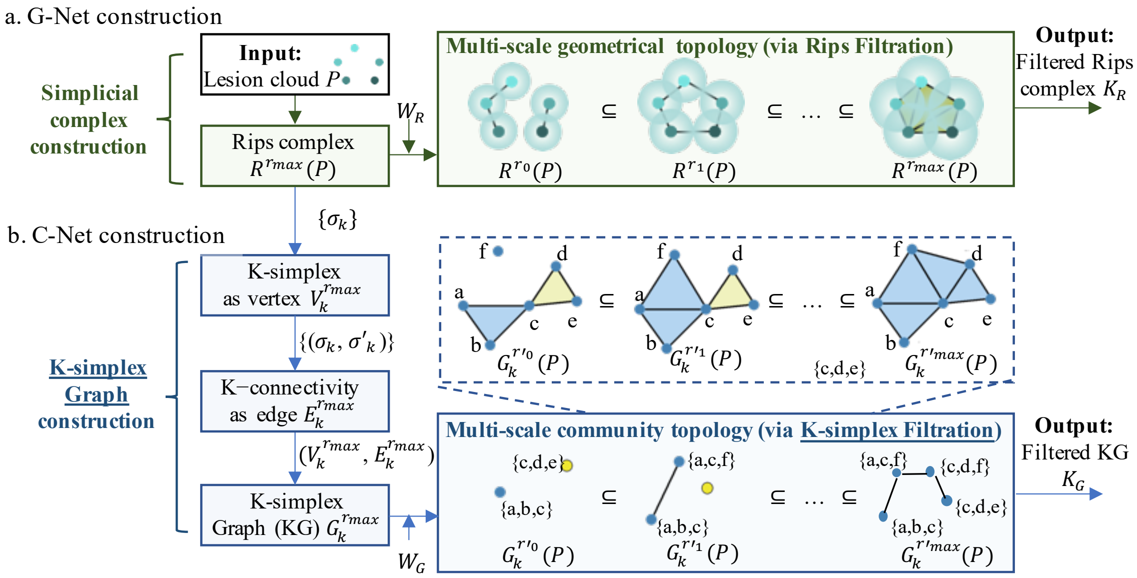

a. G-Net construction
Simplicial complex construction
Input: Lesion cloud P
Rips complex $R^{r_{max}}(P)$
W_R
Multi-scale geometrical topology (via Rips Filtration)
$R^{r_0}(P)$ ⊆ $R^{r_1}(P)$ ⊆ … ⊆ $R^{r_{max}}(P)$
Output: Filtered Rips complex K_R
$\{\sigma_k\}$
b. C-Net construction
K-simplex Graph construction
K-simplex as vertex $V_k^{r_{max}}$
$\{(\sigma_k, \sigma'_k)\}$
K−connectivity as edge $E_k^{r_{max}}$
$(V_k^{r_{max}}, E_k^{r_{max}})$
K-simplex Graph (KG) $G_k^{r_{max}}$
W_G
f d a c e b
$G_k^{r'_0}(P)$ ⊆ $G_k^{r'_1}(P)$ ⊆ … ⊆ $G_k^{r'_{max}}(P)$
{c,d,e}
Multi-scale community topology (via K-simplex Filtration)
{c,d,e} {a,b,c} {a,c,f} {a,b,c} {a,c,f} {c,d,f} {c,d,e} {a,b,c}
Output: Filtered KG K_G

FIGURE 6.2. Dynamic Hierarchical Network Construction. (a) Global-level Network (G-Net) construction contains two submodules, including simplicial complex construction and Rips Filtration. (b) Community-level Network (C-Net) construction has two submodules, including K-simplex Graph Construction, and K-simplex Filtration.

resolution of clinical MRI may vary across different patients, we further transform the point cloud from MRI space p_{mr} to physical space p for uniform comparison using the equation:

$$p = DSp_{mr} + O \tag{6.1}$$

where the direction matrix D, spacing S, and origin coordinate O are obtained from MRI meta-information.

6.2.2 Global-level Network (G-Net)

To encode dynamic geometrical structure in G-Net, we firstly construct a simplicial complex from the point cloud and then expand it to a multi-scale topology via Rips Filtration (Fig 6.2a). The theoretical motivations using the simplicial complex instead of the conventional graph are two-folds. Firstly, simplicial complex models the potential higher-order group-wise interaction among multiple lesions, while in contrast, the conventional graphs only capture pair-wise interaction between a lesion pair [258]. Secondly, simplicial complex, equipped with filtration techniques, naturally generates multi-scale graphs in contrast to conventional fixed-scale graphs [105], thus can be used to mine dynamic patterns in the evolutionary networks and bypass the challenge of determining the optimal sparsity scale of the graph edges. G-Net also serves as the base for dynamic community identification in Section 6.2.3. Specifically, to establish the connectivity of the simplicial complex, we use Vietoris-Rips complex (Rips complex) [113] that is a major algebra representation in persistent homology. Given a point cloud $P = \{p_1, p_2..., p_n\} \subset \mathbb{R}^3$, Rips complex $R^r(P)$ at scale $r > 0$ is defined as:

$$R^r(P) = \{\sigma \subseteq P \mid d(u, v) \leq 2r, \forall u \neq v \in \sigma\} \tag{6.2}$$

where d is the Euclidean distance and r is a scale factor represented as the radius of a Euclidean ball $\mathbb{B}^r(p_i)$ centered at a point p_i. More specifically, points $\{p_1, p_2..., p_n\} \in P$ span a k-simplex σ if and only if the Euclidean balls $\mathbb{B}^r(p)$ have pairwise intersection. With the Rips complex, global topological connections can be modelled at a scale r.

In the second step, Rips complex is expanded to multi-scale topology to encode geometry dynamics using Rips Filtration [259], which generates a nested family of Rips complexes (filtered Rips Complex K_R). Practically, as illustrated in Fig 6.2a, K_R is calculated by firstly computing Rips complex at a maximum scale r_{max} and then extract Rips sub-complex at a lower scale $r \leq r_{max}$. To this aim, a weight function W_R is defined for each simplex σ to represent the minimum scale r of a simplex σ when it is generated in the Rips complex. Given $\sigma \in R^{r_{max}}(P)$, the discrete weight function $W_R : R^{r_{max}}(P) \to \mathbb{R}$ is defined as:

$$W_R(\sigma) = \begin{cases} 0, & \text{dim}(\sigma) \leq 0 \\ d(u, v), & \sigma = \{u, v\} \\ \max_{N \subset \sigma} W_R(N), & \text{otherwise} \end{cases} \tag{6.3}$$

The defined weight W_R of a simplex equals to the maximum of the weights of all its edges. Then, the filtered Rips complex K_R can be computed with the tuple $(R^{r_{max}}(P), W_R)$, representing a series of nested Rips complexes:

$$\emptyset = R^{r_0}(P) \subseteq R^{r_1}(P) \subseteq ... \subseteq R^{r_{max}}(P) \tag{6.4}$$

where $R^r(P) = \{\sigma \mid W_R(\sigma) \leq r, \sigma \in R^{r_{max}}(P)\}$ and $r_0 \leq r_1 \leq ... \leq r_{max}$. The filtered Rips complex K_R is the required input of persistence homology algorithm to calculate global-level geometric invariants and will be illustrated in Section 6.3.1.

6.2.3 Community-level Network (C-Net)

To further capture the spatial patterns of lesion clusters, we propose a C-Net as illustrated in Figure 6.2b. Firstly, we design a K-simplex Graph as a high-level topological abstraction of global simplicial complex, from which dynamic community structures are encoded via the proposed K-simplex Filtration.

The definition of community topology is largely application dependent [260], [261]. As clinically multifocal lesions on MRI show patterns of densely clustered groups [246], in which one lesion may pathologically affect more than one lesion community, we model the lesion community graph based on overlapping communities. Accordingly, we define k-simplex community to characterise tightly connected k-simplices in Rips Complex from G-Net, and then extend it to multi-scale community topology using K-simplex Filtration. Specifically, we firstly define K-simplex Community and its connectivity (the edge connecting simplices) as:

DEFINITION 1. K-simplex Community is a maximal union of k-simplices that are pairwise connected with k-simplex connectivity. The k-simplex connectivity is defined as two k-simplices sharing a (k-1)-simplex.

For example, two filled triangles (2-simplices) are connected if they share a common edge (1-simplex). Notably, 0-simplex community is a singleton community as 0-simplex connectivity is $\emptyset$. We include 0-simplex community in the definition of k-simplex community because isolated lesions are also biological meaningful patterns [112]. Our definition of k-simplex Community is based on one of the essential principles of graph community that its members should be reachable through the well-connected subset of nodes (also known as community connectivity) [262]. When the community connectivity is defined as sharing fully connected k nodes, as in one of mainstream community detection methods named Clique Expansion [262]–[265], two communities sharing k-1 nodes would be regarded as two separate communities due to the lack of sufficiently strong community connectivity.

With the aim to identify the defined k-simplex community, a K-simplex Graph $G_k = (V_k, E_k)$ is designed to encode k-simplex connectivity in a high-level abstraction of global topology. This is achieved by simplifying k-simplices σ_k as vertices V_k and its connectivity (σ_k, σ_k') as E_k. As illustrated in Fig 6.2b, the vertices V_k are k-simplices σ_k extracted from Rips Complex $R^r(P)$ at scale r:

$$V_k^r = \{\sigma_k \mid \sigma_k \in R^r(P), dim(\sigma_k) = k\} \tag{6.5}$$

The edges E_k are computed based on the definition of k-simplex connectivity (σ_k, σ_k'):

$$E_k^r = \{(\sigma_k, \sigma_k') \mid \sigma_k \text{ and } \sigma_k' \text{ are connected}\} \tag{6.6}$$

To encode the community topology at dynamic scales, in the next step, we expand the K-simplex Graph to a multi-scale network via the proposed K-simplex Filtration. Specifically, similar to Rips Filtration, we firstly compute K-simplex Graph $G_k^{r_{max}}$ at the max scale r_{max} and then define a weighting function $W_G : G_k^r \to \mathbb{R}$ for vertices V_k^r and edges E_k^r to generate the nested sequence of graphs (filtered K-simplex Graph K_G). The weight function W_v for vertices V_k^r is defined as:

$$W_v(\sigma_k) = W_R(\sigma_k) = \max_{N \subset \sigma_k} W_R(N) \tag{6.7}$$

where $W_R(\sigma_k)$ is the weight function defined for the filtered Rips complex in Equation 6.3. The weight function W_e for the edges E_k of K-simplex Graph is defined as:

$$W_e((\sigma_k, \sigma_k')) = \max(W_v(\sigma_k), W_v(\sigma_k')) \tag{6.8}$$

With K-simplex Graph $G_k^{r_{max}} = (V_k^{r_{max}}, E_k^{r_{max}})$ and its weighting functions $W_G = (W_v, W_e)$, we can compute K-simplex Filtration as:

$$\emptyset = G_k^{r_0} \subseteq G_k^{r_1} \subseteq ... \subseteq G_k^{r_{max}} \tag{6.9}$$

where $G_k^r = \{\hat{\sigma} \mid W_G(\hat{\sigma}) \leq r, \hat{\sigma} \in G_k^{r_{max}}\}$ and $r_0 \leq r_1 \leq ...r_{max}$. An illustration of K-simplex Filtration is shown in Figure 6.2b. The filtered K-simplex Graph $K_G = (G_k^{r_{max}}, W_G)$ is the required input of Decomposed Community Persistence algorithm for computing community topological features, which will be illustrated in Section 6.3.2.

6.3 Dynamic Topology Quantification

After the construction of Dynamic Hierarchical Network, the next challenge is to quantify dynamic spatial patterns of the multi-scale network, particularly community topology. Firstly, we quantify geometrical structure in G-Net as homological invariants based on persistence homology, and then quantify the characteristics of clustered lesions in C-Net as attributed community dynamics based on community topology tracking and incorporation of lesion attributes.

6.3.1 Global Geometrical Invariants

Multi-scale global topology is quantified as geometrical invariants by homology tracking and persistence vectorization. From the filtered Rips Complex K_R, we firstly compute Persistence Diagram $B = (T_b, T_d)$ and record the persistence of homological components with their birth time T_b and death time T_d. The persistence of the homology ($T_d - T_b$) is defined as the difference between the birth time and death time, which captures the lifetime

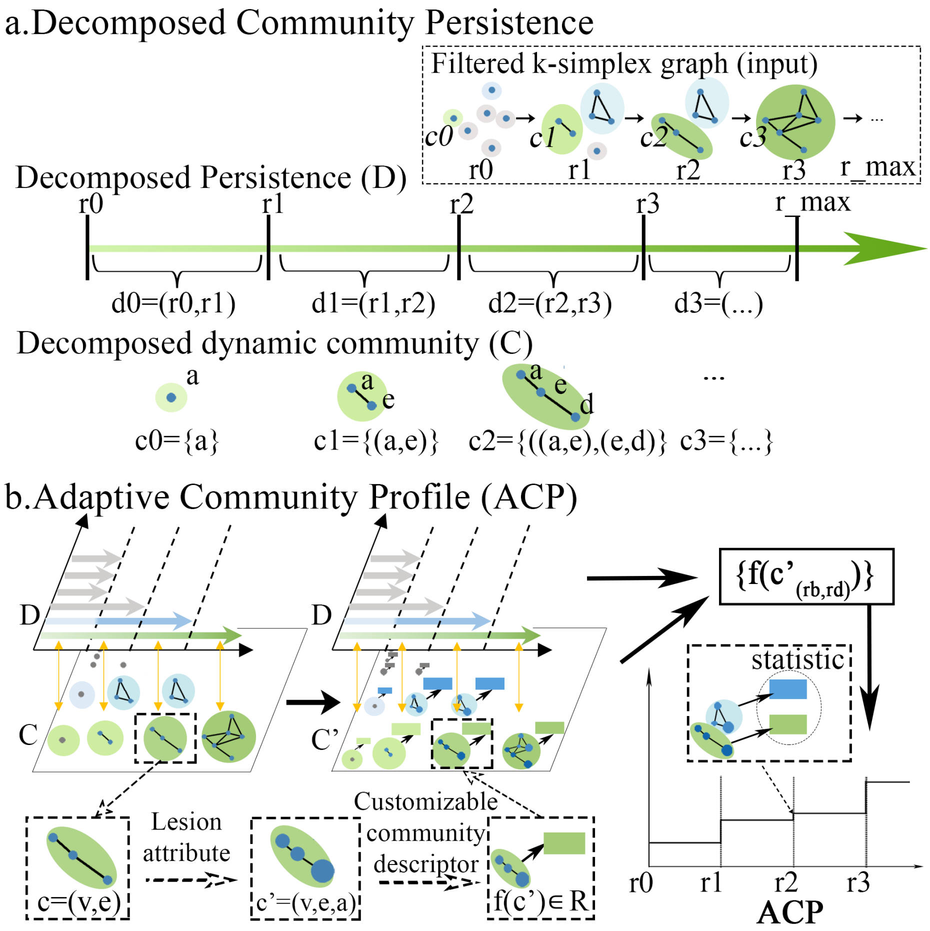


FIGURE 6.3. Multi-scale community topology quantification. (a) Illustration of Decomposed Community Persistence algorithm. Decomposed Persistence $D = \{d_1, d_2, ...\}$ and decomposed dynamic communities $C = \{c_1, c_2, ...\}$ are outputs of the algorithm. (b) The workflow for Adaptive Community Profile.

of the homological object. As the multi-set form of PD is not compatible with machine learning models, we vectorize the diagram using Persistence Image. Vector representation of PD is adopted over the kernel-based representation for easy fusion with our vector-based community features and compatibility with a broader range of machine learning algorithms. Given a Persistence Diagram $B = (T_b, T_d)$, Persistence Image [121] is a discretization of the persistence surface $\rho_B : \mathbb{R}^2 \to \mathbb{R}$, which is generated from the weighted sum of Gaussian centered at the points of a rotated persistence diagram $B' = (T_b, T_p) = (T_b, T_d - T_b)$. For any $z \in \mathbb{R}^2$,

$$\rho_B(z) = \sum_{u \in B'} \alpha(u)\Phi_u(z) \tag{6.10}$$

where $\alpha : \mathbb{R}^2 \to \mathbb{R}$ is a non-negative weight function that only depends on the persistence T_p and $\Phi_u(z) = \frac{1}{2\pi\tau^2} e^{-\frac{1}{2\tau^2}\|z-u\|^2}$ is the normalized Gaussian. τ and u denote standard deviation and mean of persistence pairs in Persistence Diagram B, respectively.

6.3.2 Attributed Community Dynamics

Multi-scale community topology is quantified as attributed community dynamics, aiming to summarize the evolutionary communities in terms of community shapes and lesion attributes. Different from the previous approach that only tracks the number of communities across scales [266], the strength of our quantification is that community heterogeneity (such as community shape and volume) is incorporated into the quantification process. To achieve this, as illustrated in Figure 6.3, we propose a novel *Decomposed Community Persistence* algorithm to track the community evolution for the incorporation of lesion attributes, and then depict the characteristics of these communities over scales by a designed *Adaptive Community Profile* equipped with a customizable descriptor.

To timely incorporate lesion attributes into dynamic communities, Decomposed Community Persistence algorithm is designed to track every change of community members at fine-grained scales during the topological evolution. Different from the previous persistence algorithm [266] that only tracks major merges of communities, we define Decomposed Persistence of dynamic communities as:

DEFINITION 2. Decomposed Persistence (r_b, r_d) is defined as the life span of a k-simplex community $C_{(r_b,r_d)}$ with the constant community members. $C_{(r_b,r_d)}$ is created at the scale r_b and ended at the scale r_d, at which new members are added in the community.

K-simplex community with Decomposed Persistence $c_{(r_b,r_d)}$ is denoted as Decomposed Dynamic Community. Figure 6.3a shows examples of decomposed dynamic communities such as c_0 with its Decomposed Persistence (r_0, r_1), and c_1 with the persistence (r_1, r_2).

Algorithm 3 Decomposed Community Persistence Algorithm

```
Input: K-simplex Graph $G_k = (V_k, E_k)$, and its weighting functions $W_G = \{W_e, W_v\}$.
Initialize:
UF ← ∅ {Empty Union-Find Structure}
D ← ∅ {Empty Decomposed Persistence}
C ← ∅ {Empty Decomposed Dynamic Communities}
Sort edges in ascending order of its weight
for every edge $(u, v) \in$ edges $E_k$ do
    if k>0 then
        $r_u \leftarrow$ UF.find(u), $r_v \leftarrow$ UF.find(v)
        $c_u \leftarrow$ UF.connected_components($r_u$)
        $c_v \leftarrow$ UF.connected_components($r_v$)
        if $W_v(r_u) < W_v(r_v)$ then {Merge the older into the newer}
            UF.union($r_v, r_u$)
        else
            UF.union($r_u, r_v$)
        end if
        D ← D ∪ $(W_v(r_u), W_e(u, v)) \cup (W_v(r_v), W_e(u, v))$
        C ← C ∪ $c_u \cup c_v$
    else if k=0 then
        if u∉C then
            D = D ∪ $(W_v(u), W_e(u, v))$
            C = C ∪ u
        else if v∉C then
            D = D ∪ $(W_v(v), W_e(u, v))$
            C = C ∪ v
        end if
    end if
end for
Output: C, D
```

Based on the decomposed persistence, we propose a *decomposed community persistence* algorithm to track and record the evolutionary changes of members in k-simplex communities $c \in C$ and its decomposed persistence $d = (r_b, r_d) \in D$, by tracking the k-simplex connectivity in the filtered K-simplex Graph $K_G = (G_k^{r_{max}}, W_c)$ based on a Union-Find data structure [267]. Specifically, as summarized in Algorithm 3, we traverse the k-simplex connectivity (edges) in ascending order of its weight and then record the status of the connected components in the union-find structure for identification of the k-simplex community C. To obtain the decomposed persistence $d = (r_b, r_d) \in D$ for these communities, we regard the newly added component at each merge as roots in the union-find structure and its associated weight as birth time r_b. The death time r_d can be obtained with weight at the next merge of this community. As singleton communities (k=0) are not associated with k-connectivity $E_0^r = \emptyset$, we explicitly set $E_0^r = E_1^r$ for the computation of the decomposed persistence of singleton communities, which is also summarized in Algorithm 3. In our study, we tracked the 0-simplex and 1-simplex communities C and their decomposed persistence D for further quantification and analysis.

In the second step, we propose Adaptive Community Profile to quantify heterogeneous community characteristics at difference scales. In the previous approach, Rieck *et al.* [266] designed a persistence indicator function $\mathbb{I}_D : \mathbb{R} \longrightarrow \mathbb{N}$ to summarize the number of communities at different scales, which was formularized as below:

$$r \longrightarrow card\{(r_b, r_d) \in D \mid r \in (r_b, r_d)\} \tag{6.11}$$

As persistence indicator function neglects the heterogeneity of lesion communities, we improve the previous formulation by incorporating the attributes of individual lesions and a descriptor function. Specifically, as illustrated in Figure 6.3b, we firstly incorporate lesion attributes a (such as lesion volume) with each lesion in k-simplex community $c_{(r_b,r_d)}$ to obtain the attributed community c'. Then, a customized community descriptor $f : C' \rightarrow \mathbb{R}$ is used to map the community to the community attribute. The descriptor could be a clinical descriptor f_a (such as community lesion volume), but also a topological descriptor f_t to describe lesion shapes (such as community clustering coefficient). The clustering coefficient [110] is a graph measurement of edge density. Lastly, we capture the heterogeneity of these communities with descriptive statistics (such as standard deviation, skewness) followed by discretization for further analysis. The proposed Adaptive Community Profile is summarized as below:

$$\begin{aligned} &\mathbb{P}_D : \mathbb{R} \longrightarrow \mathbb{N} \\ &\qquad r \longrightarrow stats\{\ f(c'_{(r_b,r_d)}) \mid r \in (r_b, r_d)\} \end{aligned} \tag{6.12}$$

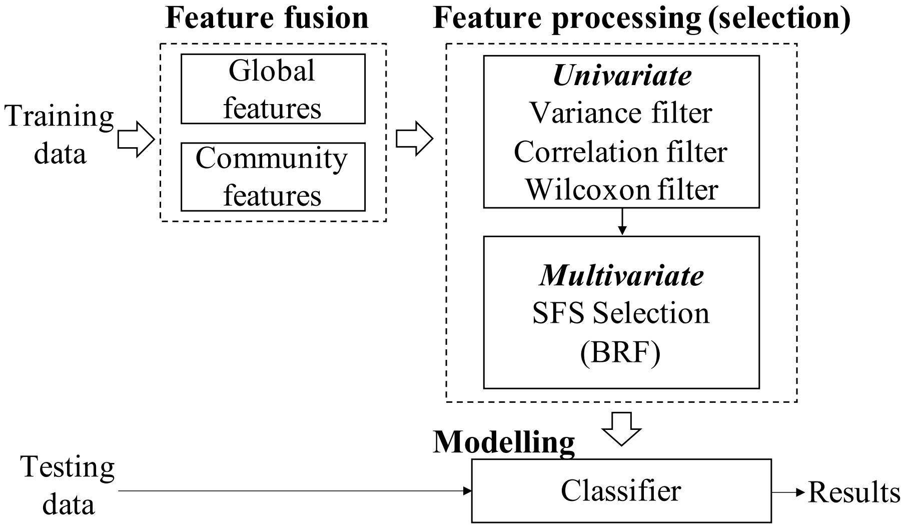


FIGURE 6.4. Topological pattern analysis.

6.4 Topological Pattern Analysis

From the quantified multi-level topological features, topological pattern analysis is performed to select informative, non-redundant and highly-relevant topological features through feature fusion and selection. The process of topological pattern analysis is illustrated in Figure 6.4. Specifically, global-level Persistence Images and community-level Adaptive Community Profile features are firstly integrated with a feature-level fusion. From the fused feature pool, we design a feature selection method to obtain the final topologial feature set with four steps: 1) Informative features are preserved by applying a variance filtering to remove zero-variance features; 2) redundant features are removed by identifying perfectly correlated feature pairs and randomly removing one of them; 3) To select task-relevant features, we perform Wilcoxon filtering to select features with significant distribution difference in two target classes; 4) lastly, the final feature set is selected via a multivariate sequential forward selection. During the feature processing, Balanced Random Forest (BRF) [161] is adopted in the multivariate SFS to select the final discriminative feature set. SFS is a wrapper feature selection algorithm, searching for the optimal feature combination based on the predictive performance from the classifier. In the subsequent modelling, the processed features are validated on two machine learning classifiers, including BRF and Cost-effective Support

Vector Machine (CE-SVM) [268]. The selected topological features are subsequently modeled with machine learning algorithms for diagnostic or prognostic applications.

6.5 Experiments and Implementations

The proposed DTA framework was evaluated on both diagnostic and prognostic tasks for multifocal diseases on two independent datasets. The experiments were comprehensively evaluated using seven evaluation metrics, including ROC_AUC, balanced accuracy (BAC), f1 score (F1), area under the precision recall curve (PR_AUC), ACC, SEN, and SPE, to avoid the inflated performance on imbalanced datasets. BAC [269] is defined as the arithmetic mean of sensitivity (true positive rate) and specificity (true negative rate), which is formulated as:

$$BAC = \frac{1}{2}(SEN + SPE) = \frac{1}{2}(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}) \tag{6.13}$$

where SEN and SPE are sensitivity and specificity respectively, and TP, TN, FP, FN denote True Positive, True Negative, False Positive and False Negative respectively.

The results were compared with seven state-of-the-art persistent homology methods, including (a) three global kernel-based methods, namely PSSKernel [116], PWGkernel [117], PFKernel [118], and (b) three global vectorisation-based methods, including Persistent Landscape [120], Persistence Image [121], Betti curve [119] and (c) one community vectorisation-based method Persistent Indicator [266]. Furthermore, our DTA framework was also compared with the reported performance of six state-of-the-art feature engineering and deep learning methods, including three methods [8], [31], [270] on the tasks of differential diagnosis and three methods [271]–[273] on the prognosis for multifocal diseases.

6.5.1 Datasets

We evaluated the proposed DTA framework on two independent datasets for diagnostic and prognostic tasks respectively, which were collected from Xuanwu Hospital, Capital Medical

University. The clinical demographics of both datasets are summarized in Table 6.1. The first dataset was collected to evaluate the value of lesion spatial pattern on brain MRI for the differential diagnosis of MS from NMO. This dataset contains T2 brain MRI of 97 patients, including 66 MS and 31 NMO patients. The diagnosis of MS and NMO was based on respective diagnostic criteria [155], [156]. As there is no public-recognized segmentation tool for NMO lesions, the marking of hyperintense brain lesions of both MS and NMO was performed by a neuroradiologist with more than nine years of experience and validated by a senior neuroradiologist, who had more than 20 years of experience.

The second dataset was collected to evaluate the value of lesion topology on brain MRI for disability progression prediction of MS patients. In this dataset, we collected T2-Flair brain MRI from 144 MS patients for prediction of disease progression, in which 90 patients with follow-up progression and 54 patients without progression. The algorithm was run on one baseline MR scan for predicting the follow-up disease progression. The follow-up disease progression was measured by the difference between the baseline EDSS and follow-up EDSS, according to the prognostic criteria defined in [274]. MS lesions on T2-Flair were automatically segmented by the lesion prediction algorithm [275] in the LST toolbox version 3.0.0, and then revised by a neuroradiologist with over nine years of experience. P-values for age, disease duration, EDSS and follow-up time is calculated through t-test, while p-value for sex is calculated through chi-squared test.

6.5.2 Implementations

In terms of implementation of our method, each lesion was separated from each other using disconnected component labeling [257] based on six-connectivity in Lesion Cloud Generation. In global topology quantification, global geometrical invariants were vectorized as Persistence Image with the resolution of 20*20. The dimension of each Persistence Image (global topology descriptor) was 400. In community topology quantification, we implemented two Adaptive Community Profiles with different community descriptors f. The first descriptor f_a was the community lesion volume (abbreviated as cmnt_lv) to capture lesion heterogeneity,

TABLE 6.1. Clinical Demographics of multifocal patients.

Dataset1: MS diagnosis (N=97)			
	MS (N=66)	NMO (N=31)	p-value
Age, year	36.5±10.2	40.5±12.0	0.119
Female/male	45/21	24/07	0.486
Disease duration	52.2±52.7	64.1±57.3	0.333
EDSS	2.9±1.5	4.3±1.6	<0.001
Dataset2: MS prognosis (N=144)			
	NP (N=54)	P (N=90)	p-value
Age, year	34.9 ±10.1	37.7 ±10.6	0.125
Female/male	35/19	60/30	0.964
Disease duration	9.5±11.1	13±12.1	0.082
EDSS_0	1.2±1.0	1.4±0.7	0.192
EDSS_1	1.6±0.7	3.3±1.2	<0.001
Follow-up	1.7±0.9	2.0±1.0	0.097

while the second descriptor f_t was community clustering coefficient (abbreviated as cmnt_cc) to measure lesion topological shape. Community statistic measures ($stats$) used in Adaptive Community Profile included skewness, kurtosis, max, standard deviation. The dimension of each Adaptive Community Profile (community topology descriptor) was 60. For classification, we evaluated our DTA framework on two different classifiers that are capable of handling imbalanced datasets, including BRF classifier [161] and CE-SVM [268] with the linear kernel.

For the implementation of comparison methods, vectorization-based methods were evaluated with both CE-SVM and BRF while kernel-based methods were only evaluated with CE-SVM as they are not compatible with non-kernel classifiers such as BRF. SVM was equipped with linear kernel and one-norm regularization for vectorization-based methods to reduce the feature dimension [276]. Comparison methods evaluated with BRF used the same feature selection pipeline as ours for uniform comparison of topology quantification. To ensure the fair comparison, all vectorisation-based comparison algorithms used 1) the same lesion point cloud from MRI images, 2) the same feature processing pipeline and 3) the same classifier as our method; and all kernel-based comparison methods were 1) applied on the same lesion point cloud, and 2) fitted on the same kernel-based classifier since feature processing was not needed for kernel-based methods.

For experimental settings, stratified independent validation was used to evaluate the performance of the proposed method. Hold-out technique is used to randomly split the dataset into independent training and testing data. Specifically, both datasets were split into training and testing sets using a stratified random split with a ratio of 7:3. Homological dimension was selected from [h0, h1]. In BRF, the n_estimator was tuned from [10, 50, 100, 500, 1000]. In SVM, parameter C was tuned from [0.1,1,10,100,1000] and class weights were tuned from [5:3, 6:2, 7:1, 'balanced'] for the diagnostic dataset and [6:4, 7:3, 8:2, 9:1, 'balanced'] for the prognostic dataset. Hyperparameter tuning was performed with grid search and three-fold cross-validation on the training dataset.

6.6 Experimental Results

6.6.1 Diagnostic Performance and Case Study

Table 6.2 shows the performance comparison of the proposed DTA framework on the task of differential diagnosis of MS and NMO with seven state-of-the-art persistent homology methods. Our DTA framework based on BRF classifier yielded the best performance on ROC_AUC (0.875), BAC (0.850), F1 score (0.865), PR_AUC (0.930) and ACC (0.833). Our method based on SVM also outperformed all other methods on ROC_AUC (0.765). Although our method only achieved the second highest on SEN and SPE, our model demonstrated less bias towards the minority or majority class and the better balance in results, as supported by the performance on ROC_AUC, balanced accuracy, f1 score, PR_AUC. These results indicate that the spatial patterns captured by our method achieved high diagnostic performance for differentiating MS from NMO, ourperforming other multi-scale topological quantification methods.

In Figure 6.5, we provided a case study on the differential diagnosis of MS and NMO, to illustrate the effectiveness of the proposed DTA framework. Two challenging cases (including one MS and one NMO) with similar lesion volume were selected from the test set, and both

TABLE 6.2. Performance of DTA framework on differential diagnosis of MS and NMO, compared with state-of-the-art persistent homology methods.

Method	Classifer	ROC_AUC	BAC	F1	PR_AUC	ACC	SEN	SPE
PSSKernel [116]	CE-SVM	0.762	0.700	0.706	0.787	0.667	0.600	0.800
PFKernel [118]	CE-SVM	0.705	0.675	0.600	0.835	0.600	0.450	0.900
PWGKernel [117]	CE-SVM	0.762	0.725	0.687	0.787	0.667	0.550	0.900
BettiCurve [119]	CE-SVM	0.758	0.650	0.818	0.860	0.733	**0.900**	0.400
PersistenceImage [121]	CE-SVM	0.670	0.700	0.571	0.866	0.600	0.400	**1.000**
PersistenceLandscape [120]	CE-SVM	0.668	0.625	0.703	0.747	0.633	0.650	0.600
CommunityIndicator [266]	CE-SVM	0.555	0.550	0.588	0.719	0.533	0.500	0.600
PersistenceLandscape [120]	BRF	0.745	0.750	0.778	0.889	0.733	0.700	0.800
BettiCurve [119]	BRF	0.555	0.550	0.320	0.777	0.433	0.200	0.900
PersistenceImage [121]	BRF	0.720	0.675	0.600	0.888	0.600	0.450	0.900
Proposed DTA	CE-SVM	0.765	0.700	0.757	0.861	0.700	0.700	0.700
Proposed DTA	BRF	**0.875**	**0.850**	**0.865**	**0.930**	**0.833**	0.800	0.900

were successfully classified by our method. Figure 6.5b shows lesion communities of MS and NMO across different scales, in which scales 10 and 19 were important scales selected by our algorithm. At scale 10, MS showed the spatial pattern of scattered lesions with a few small communities while NMO showed the pattern of a densely connected community. At scale 19, MS showed the spatial pattern of a large chain-shaped community, while NMO showed the pattern of a ball-shaped community. The patterns regarding community shapes and community volumes were quantified in our Adaptive Community Profile (ACP) in Figure 6.5c. Specifically, ACP cmnt_lv_max shows that MS patients tended to have lesion clusters with larger volume than NMO in the two cases as well as in the studied population. Similarly, ACP cmnt_cc_max shows MS tended to have more sparsely connected communities (lower clustering coefficient) than NMO.

For the diagnostic task, the interpretation of the quantified community features was visualised in Figure 6.5d. The right-half plane of Figure 6.5d showed lesion communities for the MS and NMO patients at the important scale r=10 and their corresponding abstracted graphs. At this scale, the abstract graph showed the NMO patients had more densely connected communities compared with the MS patients, which was successfully quantified by Adaptive Community Profile (clustering coefficient). Similarly, the left-half plane of Figure 6.5d showed lesion

TABLE 6.3. Classification performance of DTA framework on prognostic prediction of MS, compared with state-of-the-art persistent homology methods.

Method	**Classifier**	**ROC_AUC**	**BAC**	**F1**	**PR_AUC**	**ACC**	**SEN**	**SPE**
PSSKernel [116]	CE-SVM	0.619	0.620	0.653	0.713	0.614	0.593	0.647
PFKernel [118]	CE-SVM	0.575	0.561	0.627	0.690	0.568	0.593	0.529
PWGKernel [117]	CE-SVM	0.625	0.620	0.653	0.722	0.614	0.593	0.647
BettiCurve [119]	CE-SVM	0.454	0.410	0.458	0.663	0.409	0.407	0.412
PersistenceImage [121]	CE-SVM	0.636	0.583	0.596	0.758	0.568	0.519	0.647
PersistenceLandscape [120]	CE-SVM	0.664	0.569	0.667	0.812	0.591	0.667	0.471
CommunityIndicator [266]	CE-SVM	0.619	0.542	0.600	0.718	0.545	0.556	0.529
PersistenceLandscape [120]	BRF	0.598	0.539	0.655	0.557	0.568	0.667	0.412
BettiCurve [119]	BRF	0.578	0.594	0.578	0.713	0.568	0.481	0.706
PersistenceImage [121]	BRF	0.658	0.598	0.679	0.775	0.614	0.667	0.529
CommunityIndicator [266]	BRF	0.611	0.612	0.609	0.678	0.591	0.519	0.706
Proposed DTA	CE-SVM	0.752	**0.716**	**0.735**	0.865	**0.705**	**0.667**	**0.765**
Proposed DTA	BRF	**0.767**	0.686	0.720	**0.872**	0.682	**0.667**	0.706

communities of MS and NMO patients at the important scale r=19 and their abstracted graph. At this scale, the abstracted graph showed the MS patient had a community with larger community lesion volume compared with the NMO patient, which was successfully quantified by Adaptive Community Profile (lesion volume).

6.6.2 Prognostic Performance and Case Study

Table 6.3 shows the comparative performance of the proposed DTA framework with other state-of-the-art persistent homology methods on the task of prognostic prediction of MS. Our method based on SVM classifier outperformed all other methods on all metrics with ROC_AUC (0.752), BAC (0.716), F1 score (0.735), PR_AUC (0.865), ACC (0.705), SEN (0.667) and SPE (0.765). Our method based on BRF also outperformed other methods on all metrics with ROC_AUC (0.767), BAC (0.686), F1 score (0.720), PR_AUC (0.875), SEN (0.667), SPE (0.706) and accuracy (0.682). The results indicate that our method successfully captured lesion spatial patterns for predicting disability progression and achieved high prognostic performance compared with other multi-scale topological quantification methods.

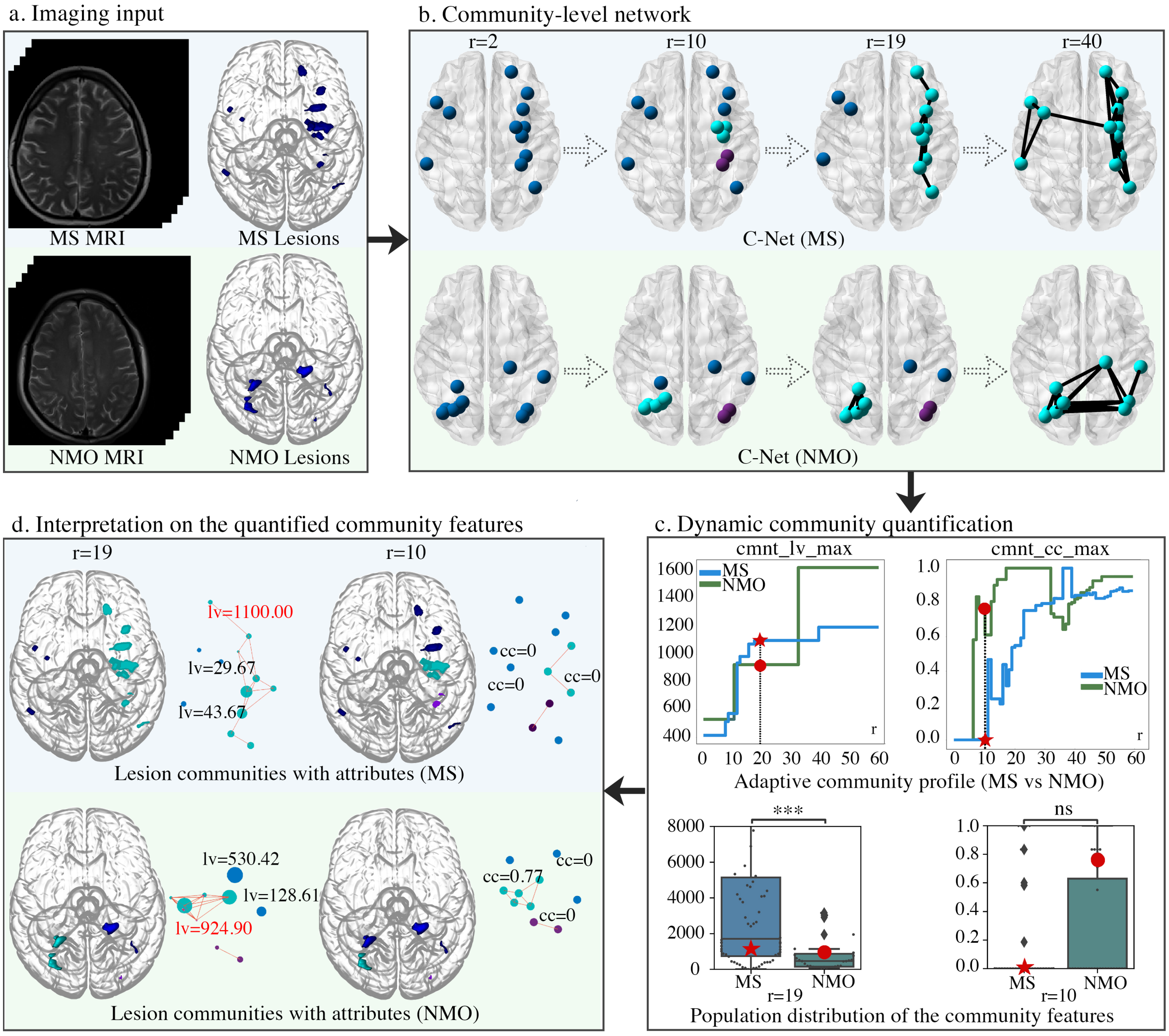

a. Imaging input
MS MRI
MS Lesions
NMO MRI
NMO Lesions
b. Community-level network
r=2
r=10
r=19
r=40
C-Net (MS)
C-Net (NMO)
c. Dynamic community quantification
cmnt_lv_max
cmnt_cc_max
MS
NMO
1600
1400
1200
1000
800
600
400
0 10 20 30 40 50 60
r
1.0
0.8
0.6
0.4
0.2
0.0
Adaptive community profile (MS vs NMO)
***
ns
8000
6000
4000
2000
0
MS
NMO
r=19
r=10
Population distribution of the community features
d. Interpretation on the quantified community features
lv=1100.00
lv=29.67
lv=43.67
cc=0
Lesion communities with attributes (MS)
lv=530.42
lv=128.61
lv=924.90
cc=0.77
Lesion communities with attributes (NMO)

FIGURE 6.5. Case study on the differential diagnosis of MS and NMO. (a) Input MRI and 3D lesion visualization. (b) Visualization of Community-level Network (C-Net). (c) Visualization of dynamic community quantification. Two Adaptive Community Profiles include one with community volume descriptor (cmnt_lv_max) and the other with community shape descriptor (cmnt_cc_max). The red star and circle denote the selected scale. The box plots show the population distribution of the community features between MS and NMO patients. (d) Visualization of connectivity and attributes of lesion communities for interpretation of community features. Abbreviations: MS=multiple sclerosis; NMO=neuromyelitis optica; C-Net=Community-level Network; cmnt_lv=community lesion volume; cmnt_cc=community clustering coefficient.

In Figure 6.6, we provided a case study on the prognostic prediction of MS, to illustrate the effectiveness of the proposed DTA framework. Two challenging cases with similar lesion volume, in which one with progression (P) and one with non-progression (NP), were selected from the test set, and both were successfully predicted by our method. Figure 6.6b shows lesion communities of P and NP across different scales, in which scale 14 and 17 were important scales selected by our algorithm. At scale 14, P case showed the spatial pattern of a relatively large community, while NP showed the spatial pattern of scattered small communities. At the scale 17, P case showed the spatial pattern of a more densely connected large community, while NP showed the pattern of several small sparsely connected communities. The patterns regarding community shapes and community volumes were quantified in our ACP in Figure 6.6c. Specifically, ACP cmnt_lv_max shows that P patients tended to have larger lesion communities than NP at scale 14, in the two cases as well as in the studied population. Similarly, ACP cmnt_cc_std shows NP tended to have more sparsely connected communities at scale 17 (more communities with 0 clustering co-efficient) than P. To interpret the quantified characteristics of communities, we visualize the topological connection and lesion attributes of the communities at the important scale r=14 and r=17 in Figure 6.6d.

For the prognostic task, the interpretation of the quantified community features was visualised in Figure 6.6d. The right-half plane of Figure 6.6d showed lesion communities for the Progression patient (P) and Non-progression patient (NP) at the important scale r=17 and their corresponding abstracted graphs. At this scale, the abstracted graph showed the P patient had more densely connected communities compared with the NP patients, which was successfully quantified by Adaptive Community Profile (clustering coefficient). Similarly, the left-half plane of Figure 6.6d showed lesion communities of MS and NMO patients at the important scale r=14 and their abstracted graph. At this scale, the abstracted graph showed the P patient had a community with larger community lesion volume compared with the NP patient, which was successfully quantified by Adaptive Community Profile (lesion volume).

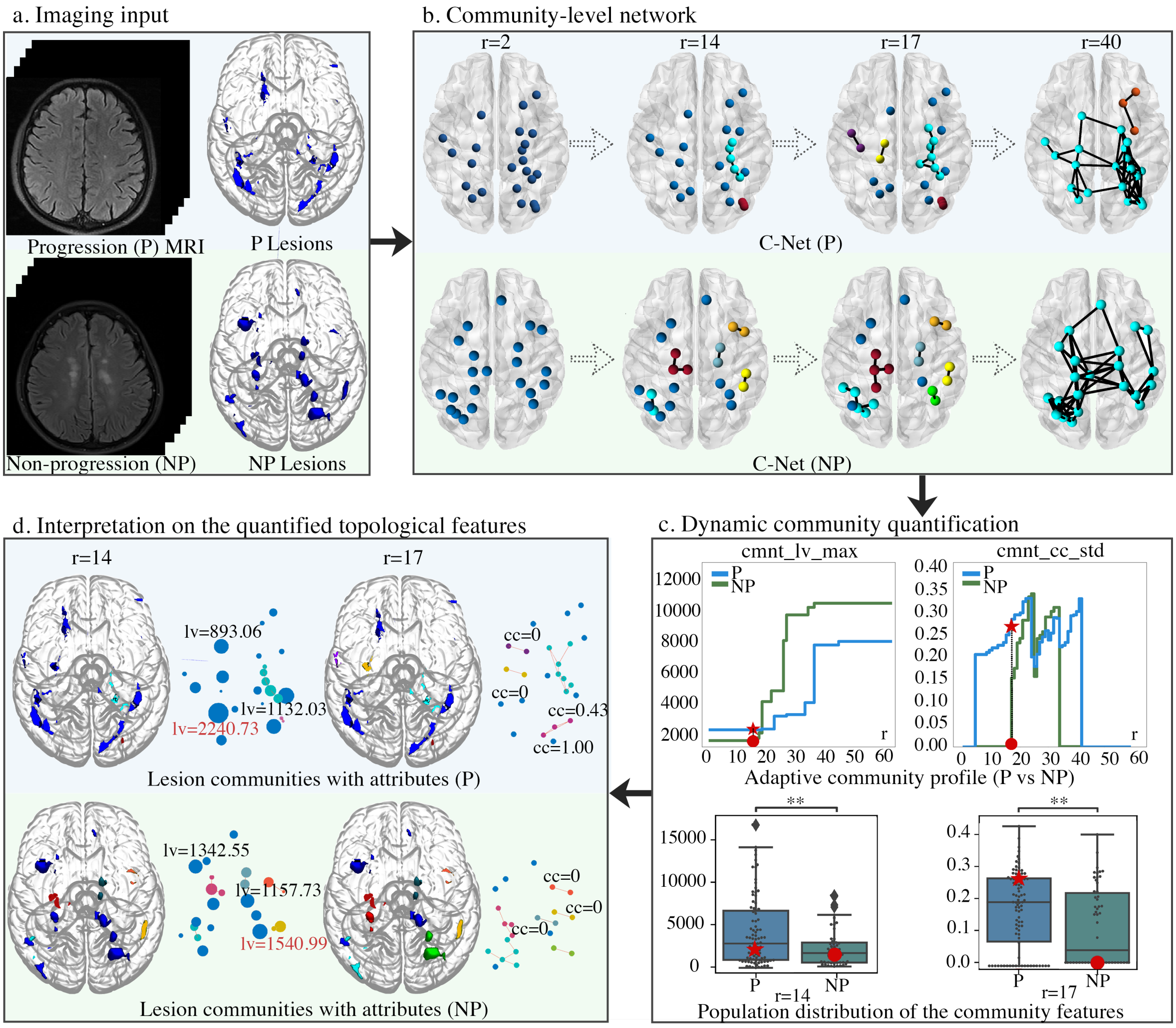


FIGURE 6.6. Case study on the prognostic prediction of MS. (a) Input MRI and 3D lesion visualization. (b) Visualization of Community-level Network (C-Net). (c) Visualization of dynamic community quantification. Two Adaptive Community Profiles include one with community volume descriptor (cmnt_lv_max) and the other with community shape descriptor (cmnt_cc_std). The red star and circle denote the selected scale. The box plots show the population distribution of the community features between P and NP patients. (d) Visualization of connectivity and attributes of lesion communities for interpretation of community features. Abbreviations: P = Progression Patient; NP=Non-Progression Patient; C-Net=Community-level Network; cmnt_lv=community lesion volume; cmnt_cc=community clustering coefficient.

6.6.3 Comparison With Feature Engineering and Deep Learning

Table 6.4 shows the comparison of DTA with the reported performance of feature engineering and deep learning methods on the metrics including ROC_AUC, accuracy, sensitivity and specificity. N/a indicates the metric was not reported. Specifically, for the differential diagnosis of MS and NMO, Table 6.4a shows that our proposed DTA method achieved the highest ROC_AUC 0.875, accuracy 0.833 and specificity 0.900, outperforming feature engineerings [8], [270] and deep learning methods [31]. For the prognostic prediction of MS, Table 6.4b shows that our proposed DTA method achieved the highest ROC_AUC 0.752, ACC 0.705, specificity 0.765, compared with feature engineering [271], [272] and deep learning methods [273].

TABLE 6.4. Comparison of the proposed method with six state-of-the-art computational radiology methods on multifocal diseases for diagnostic and prognostic tasks.

(a) Differential diagnosis of MS and NMO

Author(year)	ROC_AUC	ACC	SEN	SPE
Eshaghi *et al.* [270]	n/a	0.800	0.850	0.760
Liu *et al.* [8]	0.712	n/a	n/a	n/a
Yoo *et al.* [31]	0.801	0.813	0.850	0.750
Proposed DTA	**0.875**	**0.833**	0.800	**0.900**

(b) Prognositic prediction of MS

Author(year)	ROC_AUC	ACC	SEN	SPE
Zhao *et al.* [271]	n/a	0.690	0.720	0.670
Law *et al.* [272]	0.618	n/a	0.583	0.324
Tousignant *et al.* [273]	0.701	n/a	n/a	n/a
Proposed DTA	**0.752**	**0.705**	0.667	**0.765**

6.6.4 Comparison with Graph Classification Methods

We compared our method with three state-of-the-art graph learning methods including graph kernel (Propagation Kernel [277]), graph neural networks (Graph Convolutional Network [278]), and community detection (Community Indicator [266]). The assessment was conducted on the diagnostic prediction of MS and NMO, evaluated using CE-SVM classifier with the same split of training and testing data in the independent validation. For implementation of graph construction (Table 6.5), both propagation kernel and GCN used KNN method on Euclidean distance of points to construct a fixed-scale graph [277], while Community Indicator used its own k-clique graph to construct a multi-scale graph [266]. For the parameter tuning, Propagation Kernel and Community Indicator used CE-SVM classifier with parameter C tuned from [0.1, 1, 10, 100, 1000] and class weights tuned from [5:3, 6:2, 7:1, 'balanced']. For the parameters of GCN, Adam optimiser was used with learning rate tuned from [0.0001, 0.001, 0.01, 0.1] and hidden channels tuned from [128, 256, 512]. The results in Figure 6.7 showed that our DTA framework outperformed the 3 graph learning methods in terms of AUC, ACC and SEN. A corresponding discussion on the experimental comparison with graph learning is summarised in Section 6.7.1.

TABLE 6.5. Implementation details for graph learning methods.

Method	**Implementations**
Propagation Kernel	Graph construction: KNN with k=3, Evaluation: CE-SVM, C $\in$ [0.1, 1, 10, 100, 1000], class weight $\in$ [5:3, 6:2, 7:1, 'balanced'].
GCN	Graph construction: KNN with k=3, Optimiser: Adam, learning rate $\in$ [0.0001, 0.001, 0.01, 0.1], hidden channel $\in$ [128, 256, 512].
Community Indicator	Graph construction: k-clique graph, Evaluation: CE-SVM, C $\in$ [0.1, 1, 10, 100, 1000], class weight $\in$ [5:3, 6:2, 7:1, 'balanced'].
Our DTA	Evaluation: CE-SVM, C $\in$ [0.1, 1, 10, 100, 1000], class weight $\in$ [5:3, 6:2, 7:1, 'balanced'May].

6.6.5 Importance Analysis of Topological Features

To investigate the feature importance of global and community topological features in the final feature set, we assessed the impurity-based feature importance and univariate predictive ability based on BRF model. Univariate analysis is to separately investigate the predictive performance of each feature in the final feature set. For the diagnostic task, Figure 6.8

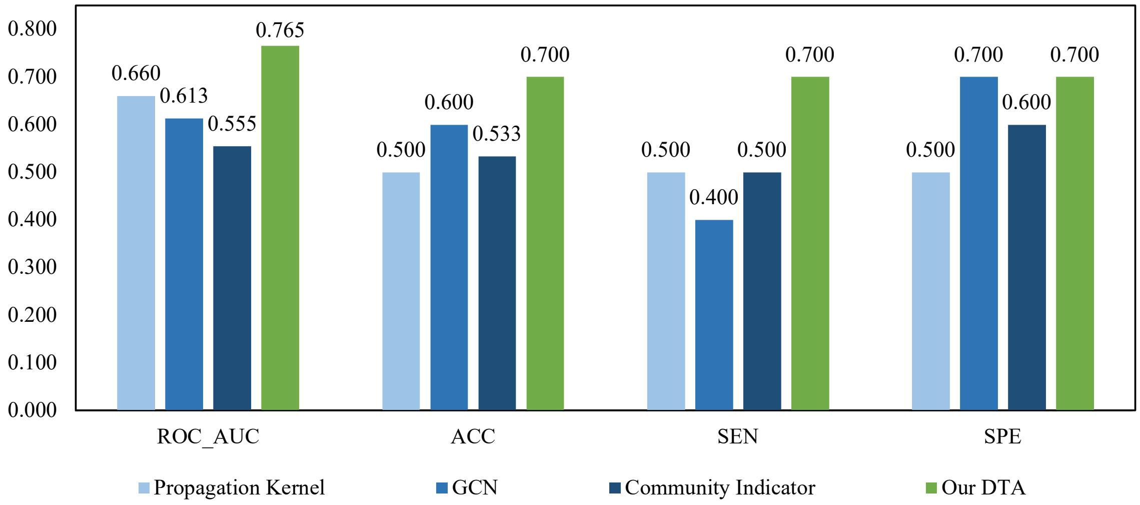


FIGURE 6.7. Comparison experiments with the state-of-the-art graph learning methods on the diagnostic task.

shows that only six community features and no global features were selected, including three features with lesion volume descriptor and the other three features with clustering coefficient descriptor. Figure 6.8 also shows that community features based on lesion volume yielded higher importance (28.9%, 27.5% and 21.7%) compared with community features based on clustering coefficient (9.9%, 7.6%, and 4.4%). This finding was consistent with the univariate analysis, in which the highest ROC_AUC was achieved by three community features based on lesion volume (0.805, 0.775 and 0.770).

For the prognostic task, Figure 6.9 shows that the selected nine topological features consisted of seven community features and two global features. In terms of feature importance, community features collectively made up 74.7% feature importance compared with 25.3% occupied by global features. As for univariate analysis, the three highest performance was all achieved by community features (0.773, 0.705 and 0.692).

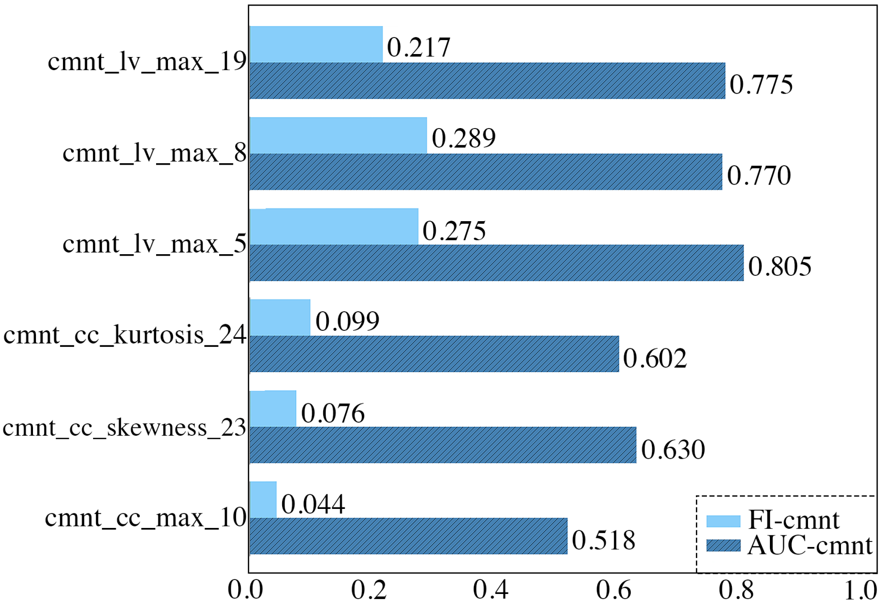


FIGURE 6.8. Feature importance and univariate results for the diagnostic task. Abbreviations: cmnt_lv=community lesion volume; cmnt_cc=community clustering coefficient; pi=persistence image; FI=Feature Importance; AUC=Area Under the receiver operating characteristic Curve.

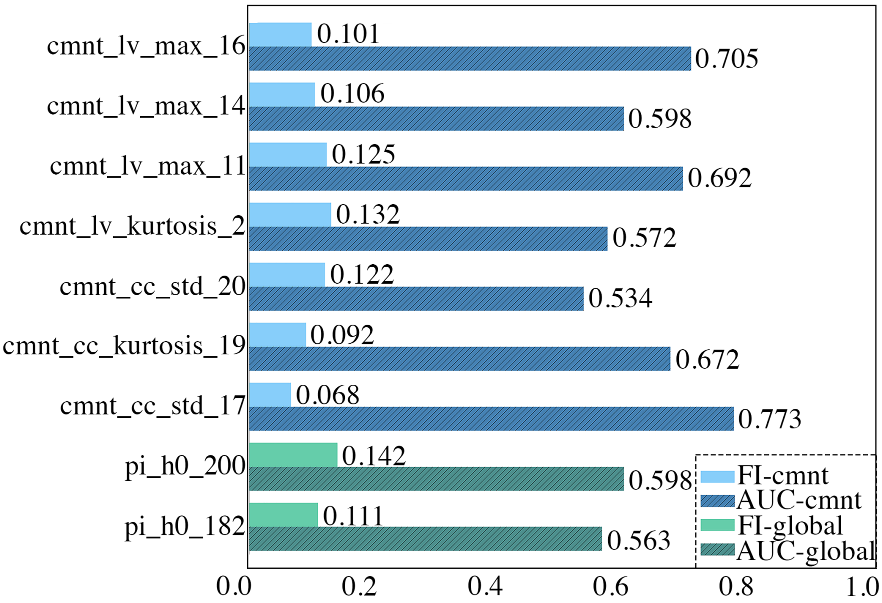


FIGURE 6.9. Feature importance and univariate results for the prognostic task.

6.6.6 Generalizability Analysis with Leave-one-out Validation

We conducted leave-one-out validation for both diagnostic and prognostic tasks to validate the generalizability of our DTA framework. As shown in Figure 6.10, the experimental results of leave-one-out validation of our method showed comparable results as hold-out independent validation for both diagnostic and prognostic tasks, which were not statistically significant ($p > 0.05$) with McNemar's test [279]. Specifically, for the prognostic task, leave-one-out validation achieved ACC 0.694 compared with hold-out validation (ACC 0.682), with a slight increase of 1.76% (p=0.25). Similarly, for the diagnostic task, our method with leave-one-out validation achieved ACC 0.804 compared with independent hold-out validation ACC 0.833, with a slight decrease of 3.48% (p=1.00).

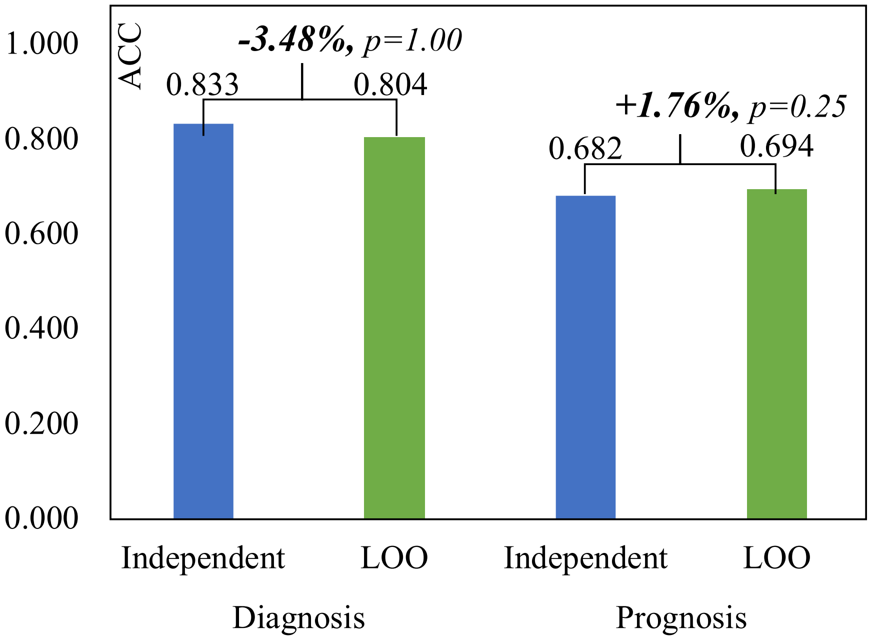


FIGURE 6.10. Comparison of leave-one-out validation (LOO) and independent hold-out validation.

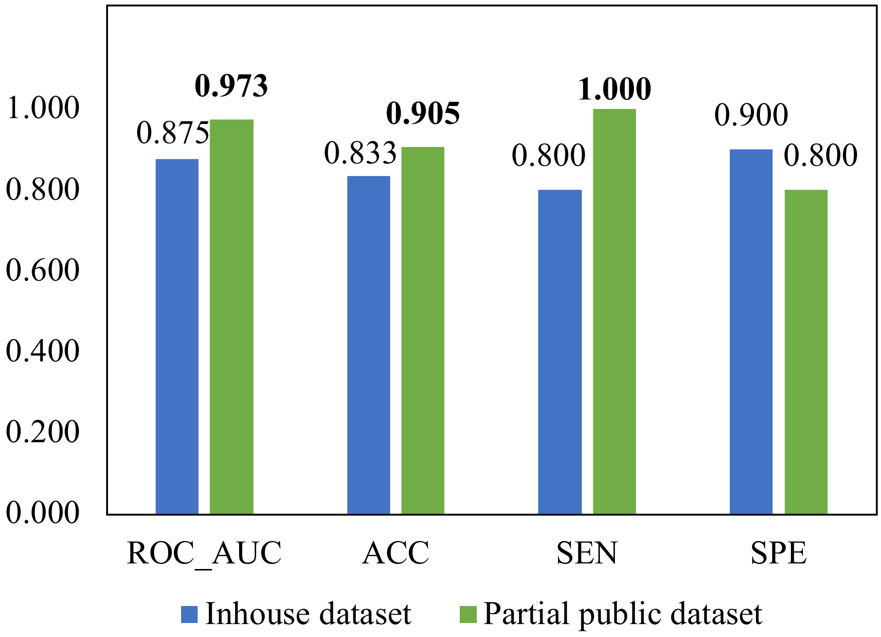


FIGURE 6.11. Experimental results on partially public datasets for differential diagnosis.

6.6.7 Generalizability Analysis with Public Datasets

It is challenging to find large public MRI data of multifocal diseases for both diagnostic and prognostic tasks, because 1) NMO data are not publicly available for the diagnostic task, 2) and progression labels (e.g., EDSS) of MS data are not publicly available for the prognostic task. As a result, we collected additional 35 MS public data from two datasets and combined with our 31 inhouse NMO data for the diagnostic validation. The public datasets include MICCAI 2008 challenge data with 20 samples [280] and MICCAI 2016 challenge data with 15 samples [281]. For MICCAI 2008 dataset, 20 training samples with T2 weighted MRI acquired from 3T Siemens scanners in two centers were used. Manual segmentation of MS lesions was provided from clinical experts. The data is publicly available at https://www.nitrc.org/frs/?group_id=745. For MICCAI 2016 dataset, 15 training samples with T2 slides were used acquired from 1.5T and 3T scanners. Manual lesion delineations were provided by independent experts. MICCAI 2016 dataset is publicly available at http://portal.fli-iam.irisa.fr/msseg-challenge/

english-msseg-data/.We implemented our DTA framework on this partial public dataset using BRF classifier. The experimental results in Figure 6.11 showed that our DTA framework retained high diagnostic performance, achieving ROC_AUC 0.973, ACC 0.905, SEN 1.00, and SPE 0.800.

6.7 Discussion

The proposed DTA framework provides the first topological modeling tool, based on persistent homology, for systematically profiling the spatial patterns of multifocal lesions, which fills the blank in current feature engineering and deep learning methods. In our DTA framework, we quantify the constructed hierarchical multi-scale network as global geometrical invariants and attributed community dynamics to capture the geometrical structure and local clusters of multifocal lesions. The experimental results demonstrated the discriminability of the quantified spatial patterns on both diagnostic and prognostic tasks, outperforming the seven state-of-the-art multi-scale topology quantification methods based on persistent homology and six reported performance of feature engineering and deep learning methods. We also illustrated the visual interpretability of our DTA framework with two case studies, showing the potential for trustworthy clinical assistance.

6.7.1 The Value of Inter-lesion Spatial Patterns Captured by DTA

Our first finding is that the inter-lesion spatial patterns, which are neglected by current feature engineering and deep learning methods, have strong predictive value towards clinical classification tasks. This finding was validated by the experimental results in Table 6.4 on both diagnostic and prognostic tasks through comparisons with feature engineering and deep learning methods. For the differential diagnosis of MS and NMO, the topological patterns that were extracted with our DTA framework achieved ROC_AUC 0.875 (Table 6.4a), which demonstrated a 9.2% increase of ROC_AUC performance compared with state-of-the-art feature engineering and deep learning methods. On the task of prognostic prediction, the topological patterns achieved ROC_AUC 0.767, a 7.4% increase of ROC_AUC compared with other methods (Table 6.4b). These results indicate that our DTA framework is able to capture not only the spatial lesion patterns of different multifocal diseases for diagnostic differentiation, but also predictive spatial patterns with the prognostic value. The discriminative spatial patterns were mined in our DTA framework by 1) integrating dynamic topological information from the multi-scale network and 2) exploiting both global-level geometrical structure and community-level spatial proximity. In terms of clinical significance, the mined topological patterns potentially reflect spatial heterogeneity of multifocal lesions, which is in accordance with the clinical findings that the spatial distribution of multifocal lesions is important for diagnostic and prognostic tasks [244], [245]. It also suggests that our method has the potential to explain the clinico-radiological paradox, to assist with patient counseling about long-term prognosis and personalized treatment plans.

Comparison with graph topology learning. Graph topology learning, which is related to our work, has achieved attractive performance on classification tasks in social network, economic network, and more recently in medical data analysis. For example, SIGN [282] is a graph neural network algorithm intended for node classification problems and its variants Inception GCN [283] and Latent Graph Learning [284] have been used for the diagnostic classification such as AD/MCI/NC to classify the patient nodes in one patient population graph. However, these node-classification methods do not fit our objective of classifying multiple lesion graphs from different patients, which should belong to graph-classification tasks. On the other hand, Graph classification methods, including graph kernels and graph neural networks, exploit latent graph embedding for classification tasks, which are related to our methods. Specifically, graph kernels (such as Propagation Kernel [277]) are mostly designed to capture only global properties of a whole graph [285]; in contrast, Graph neural networks (such as GCN [278]) is a data-driven embedding method requiring no feature engineering. However, both graph kernels and graph networks are usually based on a fixed scale of the graph, while our DTA framework exploits dynamic multi-scale community and global graphs. Furthermore, our DTA framework is 1) trainable based on the limited number of samples (as showed in Figure 6.7), 2) conveniently handles data imbalance with well-established studies in machine learning [286] and 3) well preserves the topological feature meaning and interpretability as shown in case studies (Figure 6.5-6.6).

6.7.2 The Value of Attributed Community Dynamics

The second finding of our study is that the proposed attributed community dynamics are more discriminative than global geometric invariants on both diagnostic and prognostic tasks, which provides further insights into the quantified topological features. This finding was validated by the experimental results in Table 6.2-6.3 and feature importance analysis in Figure 6.8-6.9. As demonstrated in Table 6.2-6.3, our DTA framework outperformed all global geometrical quantification methods on both diagnostic and prognostic tasks, indicating that additional information of lesion clusters captured by our Adaptive Community Profile boosted the predictive performance. Furthermore, feature importance analysis in Figure 6.8-6.9 also provided supporting evidence: 1) the cumulative feature importance of community features was much higher than global features on both tasks; 2) the univariate performance of community features was generally higher than global features. To account for the discriminability of the proposed community features, there were two underlying reasons: Firstly, we considered the heterogeneity of lesions in communities by incorporating the lesion attributes into the quantification of dynamic communities, which outperformed the community persistence indicator [266] that only based on the number of communities (Table 6.2-6.3). Secondly, the inclusion of singleton communities in our method also contributed to the classification performance, such as the pattern of singleton community with large volume captured in Figure 6.6d.

Contributions of Decomposed Community Persistence. While conventional persistence homology aims to track geometrical invariants by recording the overall persistence of homological objects, in contrast, our persistence algorithm is designed to track the detailed evolution of dynamic community characteristics by recording the decomposed persistence of dynamic communities. The decomposed persistence is essential for incorporating the domain community knowledge in the subsequent quantification. Due to the difference in the algorithm objective, technical details of our Decomposed Community Persistence differs from conventional counterpart including 1) the definition of birth and death time, 2) the merging rules, and 3) the recorded persistent pairs during the merging, with the rationale elucidated as below:

(1) The conventional persistent homology mainly tracks the birth and death of homological objects and thereby ignoring the intermediate states of dynamic community evolution. However, these intermediate states are essential to monitoring the evolution of the dynamic communities and enabling the incorporation of community domain knowledge in the subsequent quantification (Adaptive Community Profile). Thus, we give different definitions on birth and death time to capture the intermediate states during the evolution (decomposed persistence).

(2) Conventionally, merging the younger to the older only captures the overall persistence, discarding all the intermediate states. In contrast, in our method, we merge the older to the younger to technically track the newly added members, thus recording the decomposed persistence.

(3) During the merging of two components in conventional persistent homology, only one homological object is considered dead while the other continues to grow. Thus, only one persistence pair is added while the intermediate states of the other would be missing until its death. In contrast, in our new mechanism, both persistence pairs are recorded at the merging, because both communities are considered dead after the merging and the merged community is considered new-born. This enables tracking and recording of all the intermediate states of both communities.

Invariance of Adaptive Community Profile. In the design of community-based topology measures (Adaptive Community Profile), we have incorporated both intra-community invariance and inter-community invariance to fortify invariance of the community dynamics. Firstly, to enhance intra-community invariance, non-topology domain knowledge (e.g., in the clinical practice, the lesions with negligible volumes ≤ 3 voxels are considered not important and can be ignored [287]) is incorporated to minimise the effects of noise and uninformative nodes when characterising the community topology. More specifically, the incorporation of community lesion volume measures would reduce the influence of the irrelevant noise pixels with negligible volume size, which could otherwise potentially affect the community topology. Secondly, to enhance inter-community invariance, statistical measurements of different communities (e.g., mean and standard deviation) are adopted to further reduce the influence of outliers and noise.

Comparison with community quantification methods. Clique Expansion is a common approach to identify community topology in fixed-scale graphs and has been recently extended to multi-scale graphs. Specifically, [288] used the densely connected substructures called maximal cliques to study local clusters in a fixed-scale graph, and loop-shaped cavities to study parallel computations in the brain structural architecture. Different from this method using the static graph to study local clusters, we proposed a dynamic k-simplex community graph to capture the multi-scale community topological features. Community Indicator [266] is one of few multi-scale community detection methods, equipped with a quantification mechanism, for graph classification tasks. It quantifies the number of communities during the evolution as its community features; however, it neglects the heterogeneity of different lesion communities. In contrast, our DTA method 1) considers the heterogeneity of lesion communities by incorporating the lesion attributes into the quantification of dynamic communities; furthermore, 2) our method incorporates the singleton communities, a clinically meaningful pattern to represent isolated lesions [112], in our quantification, which is however neglected by Community Indicator. Thus, our methods outperform Community indicator in the performance comparison on both diagnostic and prognostic tasks (Table 6.2-6.3).

6.7.3 Interpretability of DTA for Trustworthy Clinical Decisions

Thirdly, our DTA framework demonstrated visual interpretability to facilitate clinical understanding of multifocal lesions and to assist with trustworthy clinical decisions. This finding was supported by the graphical visualization of multi-scale network construction and interpretation on the topology quantification, which were illustrated with case studies in Figure 6.5-6.6. Specifically, for the diagnostic task, Figure 6.5b shows that MS tended to have chain-shaped lesion communities with larger community volume while NMO tended to have ball-shaped community with smaller volume in multi-scale network. It was captured by Adaptive Community Profile in Figure 6.5c, which showed that communities of MS had lower clustering coefficient and larger community volume. In addition to the visualization of network construction and feature quantification, we provided further interpretation on the quantified features in Figure 6.5d by revealing the details of lesion connection and lesion attributes of the communities. As the other example, Figure 6.6b showed patients with progression (P) tend to have more densely connected communities than patients with non-progression (NP) at scale 17. It was quantified by Adaptive Community Profile in Figure 6.6c, which showed that lesion communities in P had more communities with higher clustering coefficient. Figure 6.6c also showed that patients with progression tended to have lesion communities (including singleton communities) with large volumes, which is in line with the previous clinical study that larger T2 lesions are more likely related to disease progression [289]. To sum up, our DTA framework could provide intuitive visualization of spatial lesion patterns in multi-scale network and interpretation of the pattern quantification, thus could facilitate clinical understanding of lesion spatial pattern, assist with clinical decision making, and subsequently exploit it for the therapeutic gain.

Topological feature description and its biological interpretation. Topological feature description and its biological interpretation of community and global topological features are summarised in Table 6.6. Adaptive Community Profile is community-level topological features designed to measure the statistics of community lesion volumes and community lesion density across dynamic graph sparsity by incorporating the lesion volume (lv) and clustering coefficient (cc) into community topology. In accordance with biology, for instance, larger community volume, as measured by ACP_lv, may indicate more severe damage to the structure and functionality of the brain [250], [251]. The denser lesion connectivity, as measured by ACP_cc, in the community may indicate stronger interaction among lesions and more active lesion development. In contrast, Persistence Image is global-level topological features generated to measure the invariant persistence of lesion cycles and lesion connections

TABLE 6.6. Feature description and biological meaning of topological features.

Topological features	**Feature description**	**Biological meaning**
Adaptive Community Profile (lesion volume)	The community features describe statistics about community lesion volumes across graph scales.	The larger community lesion volume may indicate more severe damage to both structure and functionality of the brain.
Adaptive Community Profile (clustering coefficient)	The community features describe statistics about community lesion density across graph scales.	The denser lesion community may indicate stronger interaction among lesions and involvement of more active development.
Persistence Image (1-dimensional homology)	The global features describe 1-dimension homology (persistence of all lesion cycles across graph scales).	The persistence of global lesion cycles (group of lesions around a hole) may indicate the incidence of normal-appearing white matter lesions inside the circle of lesions.
Persistence Image (0-dimensional homology)	The global features describe 0-dimension homology (persistence of all connected components across graph scales).	The persistence of global lesion connections may help to identify the lesion pathway.

across dynamic graph sparsity through 1-dimensional homology (h1) and 0-dimensional homology (h0). In correspondence to biology, for example, the persistence of lesion cycles (h1) may indicate the incidence of normal-appearing white matter lesions inside the circle of lesions, which are not perceptible in conventional MRI for the clinical routine [248], [249]. The persistence of lesion connections (h0) may help to identify the lesion pathway.

Future work. In our future work, in addition to inter-lesion topological relationships, we will consider further incorporating the absolute position of lesions for comprehensive spatial evaluation of multifocal lesions. Also, we will explore the relationship and interaction between topological features and other hand-craft features (e.g., radiomic features) to improve the classification performance. Furthermore, we will investigate the adaption of DTA framework to graph analysis in other domains such as brain connectome network.

6.8 Chapter Summary

In this chapter, we proposed DTA framework to characterize the global geometry and local clusters of multifocal lesions on MRI. In particular, we addressed the sparsity control challenge in graph construction with Dynamic Hierarchical Network and addressed the challenge of multi-scale community quantification via the proposed Decomposed Community Persistence algorithm and Adaptive Community Profile. Our DTA framework achieved high predictive performance on two independent clinical challenges of multifocal diseases, including differential diagnosis of MS and NMO and prognostic prediction of MS, which outperformed seven state-of-the-art persistent homology based methods and six feature engineering or deep learning methods. To summarize, the proposed DTA framework provides a platform for better understanding a wide spectrum of multifocal diseases and effectively spotting discriminative biomarkers for improving diagnosis and prognosis.

CHAPTER 7

Conclusion

7.1 Conclusions

This thesis provides Three-level Multimodal Fusion framework for quantitative analysis in the medical domain, including feature-level, information-level and knowledge-level fusion. Our framework tackles the challenges of 1) multimodal biomarker mining from high-dimensional small-sample multimodal data, 2) integration and interpretation of inter-modal and intra-modal information in multimodal deep learning, and 3) knowledge distillation from graph-based multi-focus regions incorporated with domain knowledge. The framework can be leveraged to support a wide range of medical applications including diagnostic classification, prognostic prediction, and unsupervised biomarker discovery.

In feature-level fusion, to address the challenge of biomarker discovery from high-dimensional small-sample multimodal data, we proposed an Integrative Multimodal Biomarker Mining framework to select interpretable, relevant, non-redundant and generalizable multimodal biomarkers. The framework leverages consensus clustering, Wilcoxon filter, sequential forward selection, and correlation analysis to explore the feature criterion of representativeness, robustness, discriminability, and non-redundancy. The feature selection framework was validated on two essential clinical tasks and successfully mined diagnostic and prognostic biomarkers from medical imaging data (such as CT, T1-MRI and T2-MRI) and non-imaging data.

In information-level fusion, to integrate and interpret inter-modal association and intra-modal information, we proposed a Interpretable Deep Correlational Fusion Framework based on canonical correlation analysis. Two novel fusion loss functions are proposed for supervised multimodal learning and unsupervised clustering, jointly exploiting inter-modal consensus and intra-modal discriminative information. Furthermore, an interpretation module is proposed to decode the complex non-linear cross-modal association, leveraging both deep learning interpreability and multimodal consensus interpretability. Our Deep Fusion Framework was validated on tasks in three different domains including clinical diagnosis, computer vision, and audio recognition, outperforming the state-of-the-art consensus-based multi-view learning algorithms in terms of supervised classification and unsupervised clustering.

In knowledge-level fusion, to distill knowledge from multi-focus regions and incorporated with domain knowledge, we proposed a DTA framework based on a graph-based mathematical tool named persistent homology. Different from conventional feature engineering and deep learning techniques which often focused on texture features from single-focused regions, our DTA framework is able to quantify global geometrical structure and local clustered groups from multi-focused regions. In our method, higher-order graph named simplicial complex is constructed to represent the graph structure, from which topological features are quantified and incorporated with domain knowledge. The framework was validated on diagnostic and prognostic tasks of diseases with multifocal lesion on MRI with high performance, and provided a computational tool to extract new perspective topological information for multi-focus fusion.

7.2 Future Outlook

Future research on multimodal fusion should focus on incorporating more comprehensive information for precise decision and enhancing the interpretability of extracted multimodal information. Thus, the multimodal data fusion in the medical domain could better assist clinicians with more accurate and trustworthy decisions and improve the treatment procedure and patient care.

Personalized multimodal biomarkers. In feature-level fusion, personalized medicine is one of the ultimate goals for multimodal data fusion. The major goal of personalized medicine is to tailor and optimize medical decisions (including diagnosis, prognosis, and treatment plans) for individual patients, by leveraging multimodal imaging and and non-imaging modalities. More comprehensive medical modalities should be considered, including genomics, proteomics to enhance the ability of fusion models to exploit deep pathological associations and make accurate decisions. Robustness and reproducibility should be considered for personalized multimodal biomarkers to be used in clinical practise.

Disentangled multimodal network. In information-level fusion, integration and interpretability of deep multimodal networks could be further improved by seeking disentangled consensus and complementary information. As current deep multimodal networks often non-linearly project the multi-view information in a entangled latent low-dimensional space, it is difficult to interpret the computed multimodal representations, understand the essence representations (whether it is consensus or complementary). The identified disentangled multimodal representations could facilitate the understanding of the fusion process, help explore the disease mechanism, and further improve the fusion performance based on newly acquired information.

Deep topological knowledge. In knowledge-level fusion, distilled graph knowledge from multi-focused regions could provide more value if it is equipped with the strong ability of statistical learning in deep learning or integrated other perspectives of information (such as conventional radiomics). The current dynamic topological model focuses on the construction and quantification of dynamic graph models for extracting topological features; however, the topological feature learning process could be coupled with deep learning framework to enhance the learning of more complex deep topological knowledge. In addition, the interaction between our topological perspective of knowledge and conventional perspectives of knowledge (such as texture perspective of lesions) should be further assessed for a more comprehensive knowledge-level fusion model.

Bibliography

[1] W. Raghupathi and V. Raghupathi, ‘Big Data Analytics in Healthcare: Promise and Potential,’ *Health Information Science and Systems*, vol. 2, no. 1, pp. 1–10, 2014.

[2] N. Brown, ‘Healthcare Data Growth: An Exponential Problem,’ 2015.

[3] R. Fang, S. Pouyanfar, Y. Yang *et al.*, ‘Computational Health Informatics in the Big Data Age: A Survey,’ *ACM Computing Surveys*, vol. 49, no. 1, pp. 1–36, 2016.

[4] S. E. Viswanath, P. Tiwari, G. Lee *et al.*, ‘Dimensionality Reduction-Based Fusion Approaches for Imaging and Non-Imaging Biomedical Data: Concepts, Workflow, and Use-Cases,’ *BMC Medical Imaging*, vol. 17, no. 1, pp. 1–17, 2017.

[5] S. Trip and D. Miller, ‘Imaging in Multiple Sclerosis,’ *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 76, no. suppl 3, pp. iii11–iii18, 2005.

[6] Y.-D. Zhang, Z. Dong, S.-H. Wang *et al.*, ‘Advances in Multimodal Data Fusion in Neuroimaging: Overview, Challenges, and Novel Orientation,’ *Information Fusion*, vol. 64, pp. 149–187, 2020.

[7] M. Vallieres, E. Kay-Rivest, L. J. Perrin *et al.*, ‘Radiomics Strategies for Risk Assessment of Tumour Failure in Head-and-Neck Cancer,’ *Scientific Reports*, vol. 7, no. 1, pp. 1–14, 2017.

[8] Y. Liu, D. Dong, L. Zhang *et al.*, ‘Radiomics in Multiple Sclerosis and Neuromyelitis Optica Spectrum Disorder,’ *European Radiology*, vol. 29, no. 9, pp. 4670–4677, 2019.

[9] I. Sarikaya, ‘PET Studies in Epilepsy,’ *American Journal of Nuclear Medicine and Molecular Imaging*, vol. 5, no. 5, p. 416, 2015.

[10] H. Laufs, A. Kleinschmidt, A. Beyerle *et al.*, ‘EEG-correlated fMRI of Human Alpha Activity,’ *NeuroImage*, vol. 19, no. 4, pp. 1463–1476, 2003.

[11] N. T. Issa, S. W. Byers and S. Dakshanamurthy, ‘Big Data: The Next Frontier for Innovation in Therapeutics and Healthcare,’ *Expert Review of Clinical Pharmacology*, vol. 7, no. 3, pp. 293–298, 2014.

[12] S. E. White, ‘A Review of Big Data in Health Care: Challenges and Opportunities,’ *Open Access Bioinformatics*, vol. 6, pp. 13–18, 2014.

[13] G. V. Trunk, ‘A Problem of Dimensionality: A Simple Example,’ *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 3, pp. 306–307, 1979.

[14] L. Papp, C. P. Spielvogel, I. Rausch *et al.*, ‘Personalizing Medicine Through Hybrid Imaging and Medical Big Data Analysis,’ *Frontiers in Physics*, vol. 6, p. 51, 2018.

[15] Y. Balagurunathan, Y. Gu, H. Wang *et al.*, ‘Reproducibility and Prognosis of Quantitative Features Extracted From CT Images,’ *Translational Oncology*, vol. 7, no. 1, pp. 72–87, 2014.

[16] W. K. Shams and A. W. A. Rahman, ‘Characterizing Autistic Disorder Based on Principle Component Analysis,’ in *IEEE Symposium on Industrial Electronics and Applications*, 2011, pp. 653–657.

[17] C. Morris and I. Rekik, ‘Autism Spectrum Disorder Diagnosis Using Sparse Graph Embedding of Morphological Brain Networks,’ in *Graphs in Biomedical Image Analysis, Computational Anatomy and Imaging Genetics*, Springer, 2017, pp. 12–20.

[18] I. Mhiri and I. Rekik, ‘Joint Functional Brain Network Atlas Estimation and Feature Selection for Neurological Disorder Diagnosis With Application to Autism,’ *Medical Image Analysis*, vol. 60, p. 101 596, 2020.

[19] V. Kumar, Y. Gu, S. Basu *et al.*, ‘Radiomics: The Process and the Challenges,’ *Magnetic Resonance Imaging*, vol. 30, no. 9, pp. 1234–1248, 2012.

[20] P. Seeböck, S. M. Waldstein, S. Klimscha *et al.*, ‘Unsupervised Identification of Disease Marker Candidates in Retinal OCT Imaging Data,’ *IEEE Transactions on Medical Imaging*, vol. 38, no. 4, pp. 1037–1047, 2018.

[21] K. Chaudhuri, S. M. Kakade, K. Livescu *et al.*, ‘Multi-View Clustering via Canonical Correlation Analysis,’ in *Proceedings of the International Conference on Machine Learning*, 2009, pp. 129–136.

[22] W. Härdle and L. Simar, *Applied Multivariate Statistical Analysis*. Springer, 2007, vol. 22007.

[23] M. I. Alpert and R. A. Peterson, 'On the Interpretation of Canonical Analysis,' *Journal of Marketing Research*, vol. 9, no. 2, pp. 187–192, 1972.

[24] C. J. Ter Braak, 'Interpreting Canonical Correlation Analysis Through Biplots of Structure Correlations and Weights,' *Psychometrika*, vol. 55, no. 3, pp. 519–531, 1990.

[25] A. Klami, S. Virtanen and S. Kaski, 'Bayesian Canonical Correlation Analysis.,' *Journal of Machine Learning Research*, vol. 14, no. 4, 2013.

[26] M. D. Zeiler and R. Fergus, 'Visualizing and Understanding Convolutional Networks,' in *European Conference on Computer Vision*, 2014, pp. 818–833.

[27] R. Fong and A. Vedaldi, 'Net2vec: Quantifying and Explaining How Concepts Are Encoded by Filters in Deep Neural Networks,' in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8730–8738.

[28] K. Simonyan, A. Vedaldi and A. Zisserman, 'Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps,' *arXiv Preprint arXiv:1312.6034*, 2013.

[29] M. Sundararajan, A. Taly and Q. Yan, 'Axiomatic Attribution for Deep Networks,' in *Proceedings of the International Conference on Machine Learning*, 2017, pp. 3319–3328.

[30] A. Shrikumar, P. Greenside and A. Kundaje, 'Learning Important Features Through Propagating Activation Differences,' in *Proceedings of the International Conference on Machine Learning*, 2017, pp. 3145–3153.

[31] Y. Yoo, L. Y. Tang, S.-H. Kim *et al.*, 'Hierarchical Multimodal Fusion of Deep-Learned Lesion and Tissue Integrity Features in Brain Mris for Distinguishing Neuromyelitis Optica From Multiple Sclerosis,' in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2017, pp. 480–488.

[32] N. M. Sepahvand, T. Hassner, D. L. Arnold *et al.*, 'CNN Prediction of Future Disease Activity for Multiple Sclerosis Patients From Baseline MRI and Lesion Labels,'

in *International Conference on Medical Image Computing and Computer Assisted Intervention Brainlesion Workshop*, 2018, pp. 57–69.

[33] L. Wang, T. Dong, B. Xin *et al.*, ‘Integrative Nomogram of CT Imaging, Clinical, and Hematological Features for Survival Prediction of Patients With Locally Advanced Non-Small Cell Lung Cancer,’ *European Radiology*, vol. 29, no. 6, pp. 2958–2967, 2019.

[34] K. Stepniak, A. Ursani, N. Paul *et al.*, ‘Novel 3D Printing Technology for CT Phantom Coronary Arteries With High Geometrical Accuracy for Biomedical Imaging Applications,’ *Bioprinting*, vol. 18, e00074, 2020.

[35] H. Li, C. Xu, B. Xin *et al.*, ‘18f-FDG PET/CT Radiomic Analysis With Machine Learning for Identifying Bone Marrow Involvement in the Patients With Suspected Relapsed Acute Leukemia,’ *Theranostics*, vol. 9, no. 16, p. 4730, 2019.

[36] Z. Xue, B. Xin, D. Wang *et al.*, ‘Radiomics-Enhanced Multi-Task Neural Network for Non-Invasive Glioma Subtyping and Segmentation,’ in *International Conference on Medical Image Computing and Computer Assisted Intervention Workshop*, 2019, pp. 81–90.

[37] H. Zhang, L.-l. Guo, W.-j. Tao *et al.*, ‘Comparison of the Clinical Application Value of Mo-Targeted X-Ray, Color Doppler Ultrasound and MRI in Preoperative Comprehensive Evaluation of Breast Cancer,’ *Saudi Journal of Biological Sciences*, vol. 26, no. 8, pp. 1973–1977, 2019.

[38] S. H. Jeon, C. Song, E. K. Chie *et al.*, ‘Combining Radiomics and Blood Test Biomarkers to Predict the Response of Locally Advanced Rectal Cancer to Chemoradiation,’ *In Vivo*, vol. 34, no. 5, pp. 2955–2965, 2020.

[39] A. J. Vargas and C. C. Harris, ‘Biomarker Development in the Precision Medicine Era: Lung Cancer as a Case Study,’ *Nature Reviews Cancer*, vol. 16, no. 8, pp. 525–537, 2016.

[40] C. M. Florkowski, ‘Sensitivity, Specificity, Receiver-operating Characteristic (ROC) Curves and Likelihood Ratios: Communicating the Performance of Diagnostic Tests,’ *Clinical Biochemist Reviews*, vol. 29, no. Suppl 1, p. 83, 2008.

[41] K. K. Dobbin, D. G. Beer, M. Meyerson *et al.*, 'Interlaboratory Comparability Study of Cancer Gene Expression Analysis Using Oligonucleotide Microarrays,' vol. 11, no. 2, pp. 565–572, 2005.

[42] H. Cao, S. Bernard, R. Sabourin *et al.*, 'Random Forest Dissimilarity Based Multi-View Learning for Radiomics Application,' *Pattern Recognition*, vol. 88, pp. 185–197, 2019.

[43] H. Aerts, E. Velazquez, R. Leijenaar *et al.*, 'Decoding Tumour Phenotype by Noninvasive Imaging Using a Quantitative Radiomics Approach,' *Nature Communication*, vol. 5, no. 4006, 2014.

[44] S. Parisot, S. I. Ktena, E. Ferrante *et al.*, 'Disease Prediction Using Graph Convolutional Networks: Application to Autism Spectrum Disorder and Alzheimer's Disease,' *Medical Image Analysis*, vol. 48, pp. 117–130, 2018.

[45] H. Aerts, W. Fias, K. Caeyenberghs *et al.*, 'Brain Networks Under Attack: Robustness Properties and the Impact of Lesions,' *Brain*, vol. 139, no. 12, pp. 3063–3083, 2016.

[46] S. Javed, A. Mahmood, M. M. Fraz *et al.*, 'Cellular Community Detection for Tissue Phenotyping in Colorectal Cancer Histology Images,' *Medical Image Analysis*, p. 101 696, 2020.

[47] J. S. Lewis Jr, S. Ali, J. Luo *et al.*, 'A Quantitative Histomorphometric Classifier (QuHbIC) Identifies Aggressive Versus Indolent P16-Positive Oropharyngeal Squamous Cell Carcinoma,' *The American Journal of Surgical Pathology*, vol. 38, no. 1, p. 128, 2014.

[48] I. Guyon and A. Elisseeff, 'An Introduction to Variable and Feature Selection,' *Journal of Machine Learning Research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[49] C. Lazar, J. Taminau, S. Meganck *et al.*, 'A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis,' *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 4, pp. 1106–1119, 2012.

[50] G. H. John, R. Kohavi and K. Pfleger, 'Irrelevant Features and the Subset Selection Problem,' in *Machine Learning Proceedings*, Elsevier, 1994, pp. 121–129.

[51] T. M. Cover, *Elements of Information Theory*. John Wiley & Sons, 1999.

[52] M. Dash and H. Liu, 'Consistency-Based Search in Feature Selection,' *Artificial Intelligence*, vol. 151, no. 1-2, pp. 155–176, 2003.

[53] K. Kira and L. A. Rendell, 'The Feature Selection Problem: Traditional Methods and a New Algorithm,' in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 2, 1992, pp. 129–134.

[54] I. Kononenko, E. Šimec and M. Robnik-Šikonja, 'Overcoming the Myopia of Inductive Learning Algorithms With RELIEFF,' *Applied Intelligence*, vol. 7, no. 1, pp. 39–55, 1997.

[55] H. Liu and R. Setiono, 'A Probabilistic Approach to Feature Selection-a Filter Solution,' in *Proceedings of the International Conference on Machine Learning*, vol. 96, 1996, pp. 319–327.

[56] Z. Xu, I. King, M. R.-T. Lyu *et al.*, 'Discriminative Semi-Supervised Feature Selection via Manifold Regularization,' *IEEE Transactions on Neural Networks*, vol. 21, no. 7, pp. 1033–1047, 2010.

[57] D. R. Cox, *Analysis of Survival Data. Publisher Chapman and Hall*, 2018.

[58] S. Pölsterl, S. Conjeti, N. Navab *et al.*, 'Survival Analysis for High-Dimensional, Heterogeneous Medical Data: Exploring Feature Extraction as an Alternative to Feature Selection,' *Artificial intelligence in medicine*, vol. 72, pp. 1–11, 2016.

[59] P. Pudil, F. J. Ferri, J. Novovicová *et al.*, 'Floating Search Methods for Feature Selection With Nonmonotonic Criterion Functions,' in *Proceedings of the International Conference on Pattern Recognition*, vol. 2, 1994, pp. 279–283.

[60] J. Reunanen, 'Overfitting in Making Comparisons Between Variable Selection Methods,' *Journal of Machine Learning Research*, vol. 3, no. Mar, pp. 1371–1382, 2003.

[61] P. Somol, P. Pudil, J. Novovičová *et al.*, 'Adaptive Floating Search Methods in Feature Selection,' *Pattern Recognition Letters*, vol. 20, no. 11-13, pp. 1157–1163, 1999.

[62] D. E. Goldberg, 'Genetic Algorithms in Search,' *Optimization, and MachineLearning*, 1989.

[63] Y. Sun, C. Babbs and E. Delp, 'A Comparison of Feature Selection Methods for the Detection of Breast Cancers in Mammograms: Adaptive Sequential Floating Search vs.

Genetic Algorithm,' in *IEEE engineering in medicine and biology annual conference*, 2006, pp. 6532–6535.

[64] A. H. Beg and M. Z. Islam, 'Advantages and Limitations of Genetic Algorithms for Clustering Records,' in *IEEE Conference on Industrial Electronics and Applications*, 2016, pp. 2478–2483.

[65] V. Kotsyubynsky, V. Moklyak and A. Hrubiak, 'Synthesis and Mossbauer Studies of Mesoporous γ-Fe2O3,' *Materials Science-Poland*, vol. 32, no. 3, pp. 481–486, 2014.

[66] D. Ghosh and A. M. Chinnaiyan, 'Classification and Selection of Biomarkers in Genomic Data Using LASSO,' *Journal of Biomedicine and Biotechnology*, vol. 2005, no. 2, p. 147, 2005.

[67] I. Guyon, J. Weston, S. Barnhill *et al.*, 'Gene Selection for Cancer Classification Using Support Vector Machines,' *Machine Learning*, vol. 46, no. 1, pp. 389–422, 2002.

[68] H. B. Suh, Y. S. Choi, S. Bae *et al.*, 'Primary Central Nervous System Lymphoma and Atypical Glioblastoma: Differentiation Using Radiomics Approach,' *European Radiology*, vol. 28, no. 9, pp. 3832–3839, 2018.

[69] X. Ma, L. Zhang, D. Huang *et al.*, 'Quantitative Radiomic Biomarkers for Discrimination Between Neuromyelitis Optica Spectrum Disorder and Multiple Sclerosis,' *Journal of Magnetic Resonance Imaging*, vol. 49, no. 4, pp. 1113–1121, 2019.

[70] T. Di Noto, J. von Spiczak, M. Mannil *et al.*, 'Radiomics for Distinguishing Myocardial Infarction From Myocarditis at Late Gadolinium Enhancement at MRI: Comparison With Subjective Visual Analysis,' *Radiology: Cardiothoracic Imaging*, vol. 1, no. 5, e180026, 2019.

[71] H. Li, P. Boimel, J. Janopaul-Naylor *et al.*, 'Deep Convolutional Neural Networks for Imaging Data Based Survival Analysis of Rectal Cancer,' in *International Symposium on Biomedical Imaging*, 2019, pp. 846–849.

[72] J. Shi, X. Zheng, Y. Li *et al.*, 'Multimodal Neuroimaging Feature Learning With Multimodal Stacked Deep Polynomial Networks for Diagnosis of Alzheimer's Disease,' *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 173–183, 2017.

[73] T. Xu, H. Zhang, X. Huang *et al.*, 'Multimodal Deep Learning for Cervical Dysplasia Diagnosis,' in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2016, pp. 115–123.

[74] G. Litjens, T. Kooi, B. E. Bejnordi *et al.*, 'A Survey on Deep Learning in Medical Image Analysis,' *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[75] X. Cai, F. Nie, H. Huang *et al.*, 'Heterogeneous Image Feature Integration via Multi-Modal Spectral Clustering,' in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1977–1984.

[76] Q. Yin, S. Wu, R. He *et al.*, 'Multi-View Clustering via Pairwise Sparse Subspace Representation,' *Neurocomputing*, vol. 156, pp. 12–21, 2015.

[77] J. Liu, C. Wang, J. Gao *et al.*, 'Multi-View Clustering via Joint Nonnegative Matrix Factorization,' in *Proceedings of SIAM International Conference on Data Mining*, 2013, pp. 252–260.

[78] J. Nikkilä, C. Roos, E. Savia *et al.*, 'Exploratory Modeling of Yeast Stress Response and Its Regulation With gCCA and Associative Clustering,' *International Journal of Neural Systems*, vol. 15, no. 04, pp. 237–246, 2005.

[79] M. B. Blaschko and C. H. Lampert, 'Correlational Spectral Clustering,' in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[80] Z. Huang, J. T. Zhou, X. Peng *et al.*, 'Multi-View Spectral Clustering Network.,' in *International Joint Conference of Artificial Intelligence*, 2019, pp. 2563–2569.

[81] R. Li, C. Zhang, H. Fu *et al.*, 'Reciprocal Multi-Layer Subspace Learning for Multi-View Clustering,' in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8172–8180.

[82] G. Andrew, R. Arora, J. Bilmes *et al.*, 'Deep Canonical Correlation Analysis,' in *Proceedings of the International Conference on Machine Learning*, 2013, pp. 1247–1255.

[83] W. Wang, R. Arora, K. Livescu *et al.*, 'On Deep Multi-View Representation Learning,' in *Proceedings of the International Conference on Machine Learning*, 2015, pp. 1083–1092.

[84] Q. Tang, W. Wang and K. Livescu, 'Acoustic Feature Learning Using Cross-Domain Articulatory Measurements,' in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 4849–4853.

[85] A. Benton, H. Khayrallah, B. Gujral *et al.*, 'Deep Generalized Canonical Correlation Analysis,' *arXiv Preprint arXiv:1702.02519*, 2017.

[86] J. Xie, R. Girshick and A. Farhadi, 'Unsupervised Deep Embedding for Clustering Analysis,' in *Proceedings of the International Conference on Machine Learning*, 2016, pp. 478–487.

[87] B. Yang, X. Fu, N. D. Sidiropoulos *et al.*, 'Towards K-Means-Friendly Spaces: Simultaneous Deep Learning and Clustering,' in *Proceedings of the International Conference on Machine Learning*, 2017, pp. 3861–3870.

[88] K. Ghasedi Dizaji, A. Herandi, C. Deng *et al.*, 'Deep Clustering via Joint Convolutional Autoencoder Embedding and Relative Entropy Minimization,' in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017, pp. 5736–5745.

[89] W. Hu, T. Miyato, S. Tokui *et al.*, 'Learning Discrete Representations via Information Maximizing Self-Augmented Training,' in *Proceedings of the International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 70, 2017, pp. 1558–1567.

[90] D. Das, R. Ghosh and B. Bhowmick, 'Deep Representation Learning Characterized by Inter-Class Separation for Image Clustering,' in *IEEE Winter Conference on Applications of Computer Vision*, 2019, pp. 628–637.

[91] M. Caron, P. Bojanowski, A. Joulin *et al.*, 'Deep Clustering for Unsupervised Learning of Visual Features,' in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 132–149.

[92] F. Li, H. Qiao and B. Zhang, 'Discriminatively Boosted Image Clustering With Fully Convolutional Auto-Encoders,' *Pattern Recognition*, vol. 83, pp. 161–173, 2018.

[93] X. Yang, C. Deng, F. Zheng *et al.*, 'Deep Spectral Clustering Using Dual Autoencoder Network,' in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4066–4075.

[94] Z. Jiang, Y. Zheng, H. Tan *et al.*, 'Variational Deep Embedding: An Unsupervised Generative Approach to Clustering,' in *International Joint Conference on Artificial Intelligence*, vol. 0, 2017, pp. 1965–1972.

[95] S. Chandar, M. M. Khapra, H. Larochelle *et al.*, 'Correlational Neural Networks,' *Neural Computation*, vol. 28, no. 2, pp. 257–285, 2016.

[96] J. Zhou and O. G. Troyanskaya, 'Predicting Effects of Noncoding Variants With Deep Learning–Based Sequence Model,' *Nature Methods*, vol. 12, no. 10, pp. 931–934, 2015.

[97] L. M. Zintgraf, T. S. Cohen, T. Adel *et al.*, 'Visualizing Deep Neural Network Decisions: Prediction Difference Analysis,' *International Conference of Learning Representations*, 2017.

[98] J. Zou, M. Huss, A. Abid *et al.*, 'A Primer on Deep Learning in Genomics,' *Nature Genetics*, vol. 51, no. 1, pp. 12–18, 2019.

[99] T. Zhao, Y. Xu and Y. He, 'Graph Theoretical Modeling of Baby Brain Networks,' *NeuroImage*, vol. 185, pp. 711–727, 2019.

[100] W. H. Lee, E. Bullmore and S. Frangou, 'Quantitative Evaluation of Simulated Functional Brain Networks in Graph Theoretical Analysis,' *NeuroImage*, vol. 146, pp. 724–733, 2017.

[101] A. T. Reid, J. Lewis, G. Bezgin *et al.*, 'A Cross-Modal, Cross-Species Comparison of Connectivity Measures in the Primate Brain,' *NeuroImage*, vol. 125, pp. 311–331, 2016.

[102] N. S. Schaadt, R. Schönmeyer, G. Forestier *et al.*, 'Graph-Based Description of Tertiary Lymphoid Organs at Single-Cell Level,' *PLoS Computational Biology*, vol. 16, no. 2, e1007385, 2020.

[103] A. Madabhushi, 'Digital Pathology Image Analysis: Opportunities and Challenges,' *Imaging in Medicine*, vol. 1, no. 1, p. 7, 2009.

[104] S. I. Ktena, S. Parisot, E. Ferrante *et al.*, 'Distance Metric Learning Using Graph Convolutional Networks: Application to Functional Brain Networks,' in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2017, pp. 469–477.

[105] V. Solo, J.-B. Poline, M. A. Lindquist *et al.*, 'Connectivity in fMRI: Blind Spots and Breakthroughs,' *IEEE Transactions on Medical Imaging*, vol. 37, no. 7, pp. 1537–1550, 2018.

[106] Y. Li, J. Liu, X. Gao *et al.*, 'Multimodal Hyper-Connectivity of Functional Networks Using Functionally-Weighted LASSO for MCI Classification,' *Medical Image Analysis*, vol. 52, pp. 80–96, 2019.

[107] B. C. Van Wijk, C. J. Stam and A. Daffertshofer, 'Comparing Brain Networks of Different Size and Connectivity Density Using Graph Theory,' *PloS One*, vol. 5, no. 10, e13701, 2010.

[108] Z. Wang, X. Zhu, E. Adeli *et al.*, 'Multi-Modal Classification of Neurodegenerative Disease by Progressive Graph-Based Transductive Learning,' *Medical Image Analysis*, vol. 39, pp. 218–230, 2017.

[109] X. Zhang, H. Dou, T. Ju *et al.*, 'Fusing Heterogeneous Features From Stacked Sparse Autoencoder for Histopathological Image Analysis,' *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 5, pp. 1377–1383, 2015.

[110] J. Saramäki, M. Kivelä, J.-P. Onnela *et al.*, 'Generalizations of the Clustering Coefficient to Weighted Complex Networks,' *Physical Review E*, vol. 75, no. 2, p. 027 105, 2007.

[111] J. Liu, M. Li, Y. Pan *et al.*, 'Complex Brain Network Analysis and Its Applications to Brain Disorders: A Survey,' *Complexity*, vol. 2017, 2017.

[112] Z. Chen, H. Strange, A. Oliver *et al.*, 'Topological Modeling and Classification of Mammographic Microcalcification Clusters,' *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 4, pp. 1203–1214, 2014.

[113] H. Edelsbrunner and J. Harer, *Computational Topology: An Introduction*. American Mathematical Soc., 2010.

[114] M. E. Aktas, E. Akbas and A. El Fatmaoui, 'Persistence Homology of Networks: Methods and Applications,' *Applied Network Science*, vol. 4, no. 1, p. 61, 2019.

[115] Y. Mileyko, S. Mukherjee and J. Harer, 'Probability Measures on the Space of Persistence Diagrams,' *Inverse Problems*, vol. 27, no. 12, p. 124 007, 2011.

[116] J. Reininghaus, S. Huber, U. Bauer *et al.*, 'A Stable Multi-Scale Kernel for Topological Machine Learning,' in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4741–4748.

[117] G. Kusano, Y. Hiraoka and K. Fukumizu, 'Persistence Weighted Gaussian Kernel for Topological Data Analysis,' in *Proceedings of the International Conference on Machine Learning*, 2016, pp. 2004–2013.

[118] T. Le and M. Yamada, 'Persistence Fisher Kernel: A Riemannian Manifold Kernel for Persistence Diagrams,' in *Advances in Neural Information Processing Systems*, 2018, pp. 10 007–10 018.

[119] Y. Umeda, 'Time Series Classification via Topological Data Analysis,' *Information and Media Technologies*, vol. 12, pp. 228–239, 2017.

[120] P. Bubenik, 'Statistical Topological Data Analysis Using Persistence Landscapes,' *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 77–102, 2015.

[121] H. Adams, T. Emerson, M. Kirby *et al.*, 'Persistence Images: A Stable Vector Representation of Persistent Homology,' *Journal of Machine Learning Research*, vol. 18, no. 1, pp. 218–252, 2017.

[122] O. Graa and I. Rekik, 'Multi-View Learning-Based Data Proliferator for Boosting Classification Using Highly Imbalanced Classes,' *Journal of Neuroscience Methods*, vol. 327, p. 108 344, 2019.

[123] C. Lartizien, M. Rogez, E. Niaf *et al.*, 'Computer-Aided Staging of Lymphoma Patients With FDG PET/CT Imaging Based on Textural Information,' *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 3, pp. 946–955, 2013.

[124] J. Ren, J. Tian, Y. Yuan *et al.*, 'Magnetic Resonance Imaging Based Radiomics Signature for the Preoperative Discrimination of Stage I-Ii and III-IV Head and Neck Squamous Cell Carcinoma,' *European Journal of Radiology*, vol. 106, pp. 1–6, 2018.

[125] J. Wu, T. Aguilera, D. Shultz *et al.*, 'Early-Stage Non–Small Cell Lung Cancer: Quantitative Imaging Characteristics of 18F Fluorodeoxyglucose PET/CT Allow Prediction of Distant Metastasis,' *Radiology*, vol. 281, no. 1, pp. 270–278, 2016.

[126] L. Antunovic, F. Gallivanone, M. Sollini *et al.*, '18F FDG PET/CT Features for the Molecular Characterization of Primary Breast Tumors,' *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 44, no. 12, pp. 1945–1954, 2017.

[127] S. H. Lee, P. Han, R. K. Hales *et al.*, 'Multi-View Radiomics and Dosiomics Analysis With Machine Learning for Predicting Acute-Phase Weight Loss in Lung Cancer Patients Treated With Radiotherapy,' *Physics in Medicine & Biology*, vol. 65, no. 19, p. 195 015, 2020.

[128] Q. He, X. Li, D. N. Kim *et al.*, 'Feasibility Study of a Multi-Criteria Decision-Making Based Hierarchical Model for Multi-Modality Feature and Multi-Classifier Fusion: Applications in Medical Prognosis Prediction,' *Information Fusion*, vol. 55, pp. 207–219, 2020.

[129] Y. Peng, L. Bi, M. Fulham *et al.*, 'Multi-Modality Information Fusion for Radiomics-Based Neural Architecture Search,' in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2020, pp. 763–771.

[130] M. Vallières, C. R. Freeman, S. R. Skamene *et al.*, 'A Radiomics Model From Joint FDG-PET and MRI Texture Features for the Prediction of Lung Metastases in Soft-Tissue Sarcomas of the Extremities,' *Physics in Medicine & Biology*, vol. 60, no. 14, p. 5471, 2015.

[131] W. Mu, J. Qi, H. Lu *et al.*, 'Radiomic Biomarkers From PET/CT Multi-Modality Fusion Images for the Prediction of Immunotherapy Response in Advanced Non-Small Cell Lung Cancer Patients,' in *Medical Imaging 2018: Computer-Aided Diagnosis*, vol. 10575, 2018, 105753S.

[132] G. Wang, L. He, C. Yuan *et al.*, 'Pretreatment MR Imaging Radiomics Signatures for Response Prediction to Induction Chemotherapy in Patients With Nasopharyngeal Carcinoma,' *European Journal of Radiology*, vol. 98, pp. 100–106, 2018.

[133] F. Chamming's, Y. Ueno, R. Ferré *et al.*, 'Features From Computerized Texture Analysis of Breast Cancers at Pretreatment MR Imaging Are Associated With Response to Neoadjuvant Chemotherapy,' *Radiology*, vol. 286, no. 2, pp. 412–420, 2018.

[134] J. Pérez-Beteta, D. Molina-García, J. A. Ortiz-Alhambra *et al.*, 'Tumor Surface Regularity at MR Imaging Predicts Survival and Response to Surgery in Patients With Glioblastoma,' *Radiology*, vol. 288, no. 1, pp. 218–225, 2018.

[135] Y. Wu, L. Xu, P. Yang *et al.*, 'Survival Prediction in High-Grade Osteosarcoma Using Radiomics of Diagnostic Computed Tomography,' *EBioMedicine*, vol. 34, pp. 27–34, 2018.

[136] B. Zhang, J. Tian, D. Dong *et al.*, 'Radiomics Features of Multiparametric MRI as Novel Prognostic Factors in Advanced Nasopharyngeal Carcinoma,' *Clinical Cancer Research*, vol. 23, no. 15, pp. 4259–4269, 2017.

[137] H. Park, Y. Lim, E. S. Ko *et al.*, 'Radiomics Signature on Magnetic Resonance Imaging: Association With Disease-Free Survival in Patients With Invasive Breast Cancer,' *Clinical Cancer Research*, vol. 24, no. 19, pp. 4705–4714, 2018.

[138] P. Lovinfosse, M. Polus, D. Van Daele *et al.*, 'FDG PET/CT Radiomics for Predicting the Outcome of Locally Advanced Rectal Cancer,' *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 45, no. 3, pp. 365–375, 2018.

[139] J. Wu, Y. Cui, X. Sun *et al.*, 'Unsupervised Clustering of Quantitative Image Phenotypes Reveals Breast Cancer Subtypes With Distinct Prognoses and Molecular Pathways,' *Clinical Cancer Research*, vol. 23, no. 13, pp. 3334–3342, 2017.

[140] J. Chen, L. Milot, H. M. Cheung *et al.*, 'Unsupervised Clustering of Quantitative Imaging Phenotypes Using Autoencoder and Gaussian Mixture Model,' in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 575–582.

[141] H. Li, M. Galperin-Aizenberg, D. Pryma *et al.*, 'Unsupervised Machine Learning of Radiomic Features for Predicting Treatment Response and Overall Survival of Early Stage Non-Small Cell Lung Cancer Patients Treated With Stereotactic Body Radiation Therapy,' *Radiotherapy and Oncology*, vol. 129, no. 2, pp. 218–226, 2018.

[142] S. Hussein, P. Kandel, C. W. Bolan *et al.*, 'Lung and Pancreatic Tumor Characterization in the Deep Learning Era: Novel Supervised and Unsupervised Learning Approaches,' *IEEE Transactions on Medical Imaging*, vol. 38, no. 8, pp. 1777–1787, 2019.

[143] A. J. Solomon, R. T. Naismith and A. H. Cross, 'Misdiagnosis of Multiple Sclerosis: Impact of the 2017 McDonald Criteria on Clinical Practice,' *Neurology*, vol. 92, no. 1, pp. 26–33, 2019.

[144] M. Carmosino, K. Brousseau, D. Arciniegas *et al.*, 'Initial Evaluations for Multiple Sclerosis in a University Multiple Sclerosis Center: Outcomes and Role of Magnetic Resonance Imaging in Referral,' *Archives of Neurology*, vol. 62, pp. 585–90, 2005.

[145] S. Liu, J. Kullnat, D. Bourdette *et al.*, 'Prevalence of Brain Magnetic Resonance Imaging Meeting Barkhof and McDonald Criteria for Dissemination in Space Among Headache Patients,' *Multiple Sclerosis*, vol. 19, pp. 1101–5, 2013.

[146] W. McDonald, A. Compston, G. Edan *et al.*, 'Recommended Diagnostic Criteria for Multiple Sclerosis: Guidelines From the International Panel on the Diagnosis of Multiple Sclerosis,' *Annals of Neurology*, vol. 50, pp. 121–7, 2001.

[147] A. Compston and A. Coles, 'Multiple Sclerosis,' *Lancet*, vol. 359, pp. 1221–31, 2002.

[148] A. Solomon, D. Bourdette, A. Cross *et al.*, 'The Contemporary Spectrum of Multiple Sclerosis Misdiagnosis: A Multicenter Study,' *Neurology*, vol. 87, pp. 1393–9, 2016.

[149] D. Wingerchuk, V. Lennon, C. Lucchinetti *et al.*, 'The Spectrum of Neuromyelitis Optica,' *Lancet Neurology*, vol. 6, pp. 805–15, 2007.

[150] W. Franca, G. Lorenz, H. Arsany *et al.*, 'Rebound After Fingolim,od and a Single Daclizumab Injection in a Patient Retrospectively Diagnosed With NMO Spectrum Disorder-Mri Apparent Diffusion Coefficient Maps in Differential Diagnosis of Demyelinating CNS Disorders,' *Front Neurology*, vol. 9, no. 782, 2018.

[151] R. J. Gillies, P. E. Kinahan and H. Hricak, 'Radiomics: Images Are More Than Pictures, They Are Data,' *Radiology*, vol. 278, pp. 563–577, 2016.

[152] S. Kremer, F. Renard, S. Achard *et al.*, 'Use of Advanced Magnetic Resonance Imaging Techniques in Neuromyelitis Optica Spectrum Disorder,' *JAMA Neurology*, vol. 72, pp. 815–22, 2015.

[153] Z. Qian, Y. Li, Y. Wang *et al.*, 'Differentiation of Glioblastoma From Solitary Brain Metastases Using Radiomic Machine-Learning Classifiers,' *Cancer Letter*, vol. 451, pp. 128–35, 2019.

[154] Z. Liu, S. Wang, D. Dong *et al.*, 'The Applications of Radiomics in Precision Diagnosis and Treatment of Oncology: Opportunities and Challenges,' *Theranostics*, vol. 9, pp. 1303–22, 2019.

[155] C. H. Polman, S. C. Reingold, B. Banwell *et al.*, 'Diagnostic Criteria for Multiple Sclerosis: 2010 Revisions to the McDonald Criteria,' *Annals of Neurology*, vol. 69, no. 2, pp. 292–302, 2011.

[156] D. M. Wingerchuk, B. Banwell, J. L. Bennett *et al.*, 'International Consensus Diagnostic Criteria for Neuromyelitis Optica Spectrum Disorders,' *Neurology*, vol. 85, no. 2, pp. 177–189, 2015.

[157] J. Kurtzke, 'Rating Neurologic Impairment in Multiple Sclerosis: An Expanded Disability Status Scale (EDSS),' *Neurology*, vol. 33, pp. 1444–52, 1983.

[158] V. Wylde, S. Palmer, I. Learmonth *et al.*, 'Test-Retest Reliability of Quantitative Sensory Testing in Knee Osteoarthritis and Healthy Participants,' *Osteoarthritis Cartilage*, vol. 19, pp. 655–8, 2011.

[159] Y. Saeys, I. Inza and P. Larranaga, 'A Review of Feature Selection Techniques in Bioinformatics,' *Bioinformatics*, vol. 23, pp. 2507–17, 2007.

[160] F. Ferri, P. Pudil and M. Hatef, 'Comparative Study of Techniques for Large-Scale Feature Selection,' *Machine Intelligence and Pattern Recognition*, vol. 16, pp. 403–13, 1994.

[161] C. Chen, A. Liaw and L. Breiman, 'Using Random Forest to Learn Imbalanced Data,' *University of California, Berkeley*, vol. 110, no. 1-12, p. 24, 2004.

[162] J. Y. Kim, J. E. Park, Y. Jo *et al.*, 'Incorporating Diffusion-and Perfusion-Weighted MRI Into a Radiomics Model Improves Diagnostic Performance for Pseudoprogression in Glioblastoma Patients,' *Neuro-oncology*, vol. 21, no. 3, pp. 404–414, 2019.

[163] J. Qu, C. Shen, J. Qin *et al.*, 'The Mr Radiomic Signature Can Predict Preoperative Lymph Node Metastasis in Patients With Esophageal Cancer,' *European Radiology*, vol. 29, no. 2, pp. 906–914, 2019.

[164] B. X. Ren, I. Huen, Z. J. Wu *et al.*, 'Early Postnatal Irradiation-Induced Age-Dependent Changes in Adult Mouse Brain: MRI Based Characterization,' *BMC Neuroscience*, vol. 22, no. 1, pp. 1–14, 2021.

[165] F. Orlhac, S. Boughdad, C. Philippe *et al.*, 'A Postreconstruction Harmonization Method for Multicenter Radiomic Studies in Pet,' *Journal of Nuclear Medicine*, vol. 59, no. 8, pp. 1321–1328, 2018.

[166] M. Sheikhan, M. Bejani and D. Gharavian, 'Modular Neural-Svm Scheme for Speech Emotion Recognition Using Anova Feature Selection Method,' *Neural Computing and Applications*, vol. 23, no. 1, pp. 215–227, 2013.

[167] M. Bennasar, Y. Hicks and R. Setchi, 'Feature Selection Using Joint Mutual Information Maximisation,' *Expert Systems with Applications*, vol. 42, no. 22, pp. 8520–8532, 2015.

[168] B. Zhang, X. He, F. Ouyang *et al.*, 'Radiomic Machine-Learning Classifiers for Prognostic Biomarkers of Advanced Nasopharyngeal Carcinoma,' *Cancer Letters*, vol. 403, pp. 21–27, 2017.

[169] H. Zou and T. Hastie, 'Regularization and Variable Selection via the Elastic Net,' *Journal of the royal statistical society: series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[170] X. Wang, D. Wang, Z. Yao *et al.*, 'Machine Learning Models for Multiparametric Glioma Grading With Quantitative Result Interpretations,' *Frontiers of Neuroscience*, vol. 12, no. 1046, 2018.

[171] E. Strumbelj and K. I, 'Explaining Prediction Models and Individual Predictions With Feature Contributions,' *Knowledge and Information Systems*, vol. 41, pp. 647–65, 2014.

[172] S.-Y. Huh, J.-H. Min, W. Kim *et al.*, 'The Usefulness of Brain MRI at Onset in the Differentiation of Multiple Sclerosis and Seropositive Neuromyelitis Optica Spectrum Disorders,' *Multiple Sclerosis Journal*, vol. 20, no. 6, pp. 695–704, 2014.

[173] H. Kim, Y. Lee, Y.-H. Kim *et al.*, 'Deep Learning-Based Method to Differentiate Neuromyelitis Optica Spectrum Disorder From Multiple Sclerosis,' *Frontiers in Neurology*, vol. 11, p. 1642, 2020.

[174] S. Lalan, M. Khan, B. Schlakman *et al.*, 'Differentiation of Neuromyelitis Optica From Multiple Sclerosis on Spinal Magnetic Resonance Imaging,' *International Journal of MS Care*, vol. 14, no. 4, pp. 209–214, 2012.

[175] L. Matthews, R. Marasco, M. Jenkinson *et al.*, 'Distinction of Seropositive NMO Spectrum Disorder and MS Brain Lesion Distribution,' *Neurology*, vol. 80, no. 14, pp. 1330–1337, 2013.

[176] J. M. Nielsen, T. Korteweg, F. Barkhof *et al.*, 'Overdiagnosis of Multiple Sclerosis and Magnetic Resonance Imaging Criteria,' *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, vol. 58, no. 5, pp. 781–783, 2005.

[177] M. Filippi, W. Bruck, D. Chard *et al.*, 'Association Between Pathological and MRI Findings in Multiple Sclerosis,' *Lancet Neurology*, vol. 18, pp. 198–210, 2019.

[178] Q. Wang, Q. Li, R. Mi *et al.*, 'Radiomics Nomogram Building From Multiparametric MRI to Predict Grade in Patients With Glioma: A Cohort Study,' *Journal of Magnetic Resonance Imaging*, vol. 49, pp. 825–33, 2019.

[179] G. Chandrashekar and F. Sahin, 'A Survey on Feature Selection Methods,' *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[180] S. Lundberg, B. Nair, M. Vavilala *et al.*, 'Explainable Machine-Learning Predictions for the Prevention of Hypoxaemia During Surgery,' *Nature Biomedical Engineering*, vol. 2, pp. 749–60, 2018.

[181] M. Calabrese, V. Poretto, A. Favaretto *et al.*, 'Cortical Lesion Load Associates With Progression of Disability in Multiple Sclerosis,' *Brain*, vol. 135, pp. 2952–61, 2012.

[182] D. Li, U. Held, J. Petkau *et al.*, 'MRI T2 Lesion Burden in Multiple Sclerosis: A Plateauing Relationship With Clinical Disability,' *Neurology*, vol. 66, pp. 1384–9, 2006.

[183] J. Hara, A. Wu, J. Villanueva-Meyer *et al.*, 'Clinical Applications of Quantitative 3-Dimensional MRI Analysis for Pediatric Embryonal Brain Tumors,' *International Journal of Radiation Oncology, Biology, Physics*, vol. 102, pp. 744–56, 2018.

[184] D. Karussis, 'The Diagnosis of Multiple Sclerosis and the Various Related Demyelinating Syndromes: A Critical Review,' *Journal of Autoimmunity*, vol. 48, pp. 134–142, 2014.

[185] A. Jemal, F. Bray, M. M. Center *et al.*, 'Global Cancer Statistics,' *CA: A Cancer Journal for Clinicians*, vol. 61, pp. 134–134, 2011.

[186] P. Yang, M. S. Allen, M. C. Aubry *et al.*, ‘Clinical Features of 5,628 Primary Lung Cancer Patients: Experience at Mayo Clinic From 1997 to 2003,’ *Chest*, vol. 128, pp. 452–462, 2005.

[187] A. Aupérin, C. Le Péchoux, E. Rolland *et al.*, ‘Meta-Analysis of Concomitant Versus Sequential Radiochemotherapy in Locally Advanced Non-Small-Cell Lung Cancer,’ *Journal of Clinical Oncology*, vol. 28, pp. 2181–2190, 2010.

[188] N. Hanna, M. Neubauer, C. Yiannoutsos *et al.*, ‘Phase III Study of Cisplatin, Etoposide, and Concurrent Chest Radiation With or Without Consolidation Docetaxel in Patients With Inoperable Stage III Non–Small-Cell Lung Cancer: The Hoosier Oncology Group and U.S,’ *Journal of Clinical Oncology*, vol. 26, pp. 5755–5760, 2008.

[189] E. E. Vokes, J. E. Herndon, M. J. Kelley *et al.*, ‘Induction Chemotherapy Followed by Chemoradiotherapy Compared With Chemoradiotherapy Alone for Regionally Advanced Unresectable Stage III Non–Small-Cell Lung Cancer: Cancer and Leukemia Group B,’ *Journal of Clinical Oncology*, vol. 25, no. 1698, 2007.

[190] P. Lambin, R. T. Leijenaar, T. M. Deist *et al.*, ‘Radiomics: The Bridge Between Medical Imaging and Personalized Medicine,’ *Nature Reviews Clinical Oncology*, vol. 14, no. 749, 2017.

[191] V. Agrawal, T. P. Coroller, Y. Hou *et al.*, ‘Radiologic-Pathologic Correlation of Response to Chemoradiation in Resectable Locally Advanced NSCLC,’ *Lung Cancer*, vol. 102, no. 1, 2016.

[192] H. Mathieu, M. Mohamed and M. V, ‘Hatt, Mathieu and Majdoub, Mohamed and Vallières, Martin and Tixier, Florent and Le Rest, Catherine Cheze and Groheux, David and Hindié, Elif and Martineau, Antoine and Pradier, Olivier and Hustinx, Roland and others,’ *Journal of Nuclear Medicine*, vol. 56, no. 38, 2015.

[193] M. Kirienko, F. Gallivanone, M. Sollini *et al.*, ‘FDG PET/CT as Theranostic Imaging in Diagnosis of Non-Small Cell Lung Cancer,’ *Frontiers of Bioscience*, vol. 22, no. 1713, 2017.

[194] M.-C. Desseroit, D. Visvikis, F. Tixier *et al.*, ‘Development of a Nomogram Combining Clinical Staging With 18f-FDG PET/CT Image Features in Non-Small-Cell Lung

Cancer Stage I-Iii,' *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 43, pp. 1477–1485, 2016.

[195] T. Pyka, R. A. Bundschuh, N. Andratschke *et al.*, 'Textural Features in Pre-Treatment [F18]-FDG-Pet/Ct Are Correlated With Risk of Local Recurrence and Disease-Specific Survival in Early Stage NSCLC Patients Receiving Primary Stereotactic Radiation Therapy,' *Radiation Oncology*, vol. 10, no. 100, 2015.

[196] A. Schernberg, S. Reuze, F. Orlhac *et al.*, 'A Score Combining Baseline Neutrophilia and Primary Tumor SUV Peak Measured From FDG PET Is Associated With Outcome in Locally Advanced Cervical Cancer,' *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 45, pp. 187–195, 2018.

[197] K. A. Scilla, S. M. Bentzen, V. K. Lam *et al.*, 'Neutrophil-Lymphocyte Ratio Is a Prognostic Marker in Patients With Locally Advanced (Stage IIIA and IIIB) Non-Small Cell Lung Cancer Treated With Combined Modality Therapy,' *Oncologist*, vol. 22, pp. 737–742, 2017.

[198] R. M. Bremnes, K. Al-Shibli, T. Donnem *et al.*, 'The Role of Tumor-Infiltrating Immune Cells and Chronic Inflammation at the Tumor Site on Cancer Development, Progression, and Prognosis: Emphasis on Non-Small Cell Lung Cancer,' *Journal of Thoracic Oncology*, vol. 6, no. 4, pp. 824–833, 2011.

[199] R. D. Schreiber, L. J. Old and M. J. Smyth, 'Cancer Immunoediting: Integrating Immunity's Roles in Cancer Suppression and Promotion,' *Science*, vol. 331, pp. 1565–1570, 2011.

[200] B. Xin, C. Xu, L. Wang *et al.*, 'Integrative Clustering and Supervised Feature Selection for Clinical Applications,' in *International Conference on Control, Automation, Robotics and Vision*, 2018, pp. 1316–1320.

[201] H. Cui, X. Wang, J. Zhou *et al.*, 'Topology Polymorphism Graph for Lung Tumor Segmentation in PET-CT Images,' *Physics in Medicine and Biology*, vol. 60, pp. 4893–4914, 2015.

[202] H. Cui, X. Wang, J. Zhou *et al.*, 'A Topo-Graph Model for Indistinct Target Boundary Definition From Anatomical Images,' *Computer Methods and Programs in Biomedicine*, vol. 159, 2018.

[203] S. Leger, A. Zwanenburg, K. Pilz *et al.*, ‘A Comparative Study of Machine Learning Methods for Time-to-Event Survival Data for Radiomics Risk Modelling,’ *Scientific Report*, vol. 7, no. 13206, 2017.

[204] S. Monti, P. Tamayo, J. Mesirov *et al.*, ‘Consensus Clustering: a Resampling-based Method for Class Discovery and Visualization of Gene Expression Microarray Data,’ *Machine learning*, vol. 52, no. 1, pp. 91–118, 2003.

[205] T. A. Gerds, M. W. Kattan, M. Schumacher *et al.*, ‘Estimating a Time-Dependent Concordance Index for Survival Prediction Models With Covariate Dependent Censoring,’ *Statistics in Medicine*, vol. 32, pp. 2173–2184, 2013.

[206] C. Oberije, D. De Ruysscher, R. Houben *et al.*, ‘A Validated Prediction Model for Overall Survival From Stage III Non-Small Cell Lung Cancer: Toward Survival Prediction for Individual Patients,’ *International Journal of Radiation Oncology Biology Physics*, vol. 92, no. 935, 2015.

[207] S. Tanadini-Lang, J. Rieber, A. R. Filippi *et al.*, ‘Nomogram Based Overall Survival Prediction in Stereotactic Body Radiotherapy for Oligo-Metastatic Lung Disease,’ *Radiotherapy and Oncology*, vol. 123, no. 2, pp. 182–188, 2017.

[208] X.-R. Tang, Y.-Q. Li, S.-B. Liang *et al.*, ‘Development and Validation of a Gene Expression-Based Signature to Predict Distant Metastasis in Locoregionally Advanced Nasopharyngeal Carcinoma: A Retrospective, Multicentre, Cohort Study,’ *Lancet Oncology*, 2018.

[209] X. ave, L. Zhang, J. Yang *et al.*, ‘Delta-Radiomics Features for the Prediction of Patient Outcomes in Non–small Cell Lung Cancer,’ *Scientific Report*, vol. 7, no. 588, 2017.

[210] J. Lee, B. Li, Y. Cui *et al.*, ‘A Quantitative CT Imaging Signature Predicts Survival and Complements Established Prognosticators in Stage I Non-Small Cell Lung Cancer,’ *International Journal of Radiation Oncology Biology Physics*, 2018.

[211] Y. Huang, Z. Liu, L. He *et al.*, ‘Radiomics Signature: A Potential Biomarker for the Prediction of Disease-Free Survival in Early-Stage (I or II,’ *Radiology*, vol. 281, no. 947, 2016.

[212] N. Ohri, F. Duan, B. S. Snyder *et al.*, ‘Pretreatment 18fdg-Pet Textural Features in Locally Advanced Non-Small Cell Lung Cancer: Secondary Analysis of ACRIN 6668/Rtog 0235,’ *Journal of Nuclear Medicine*, vol. 57, pp. 228–233, 2016.

[213] A. Salavati, F. Duan, B. S. Snyder *et al.*, ‘Optimal FDG PET/CT Volumetric Parameters for Risk Stratification in Patients With Locally Advanced Non-Small Cell Lung Cancer: Results From the ACRIN 6668/Rtog 0235 Trial,’ *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 44, pp. 1969–1983, 2017.

[214] S.-X. Rao, D. M. Lambregts, R. S. Schnerr *et al.*, ‘CT Texture Analysis in Colorectal Liver Metastases: A Better Way Than Size and Volume Measurements to Assess Response to Chemotherapy?’ *United European Gastroenterology Journal*, vol. 4, pp. 257–263, 2016.

[215] X. Dong, X. Sun, L. Sun *et al.*, ‘Early Change in Metabolic Tumor Heterogeneity During Chemoradiotherapy and Its Prognostic Value for Patients With Locally Advanced Non-Small Cell Lung Cancer,’ *PLoS One*, vol. 11, no. 6, e0157836, 2016.

[216] A. Cunliffe, S. Iii, R. Castillo *et al.*, ‘Lung Texture in Serial Thoracic Computed Tomography Scans: Correlation of Radiomics-Based Features With Radiation Therapy Dose and Radiation Pneumonitis Development,’ *International Journal of Radiation Oncology Biology Physics*, vol. 91, pp. 1048–1056, 2015.

[217] G. Cox, R. Walker, A. Andi *et al.*, ‘Prognostic Significance of Platelet and Microvessel Counts in Operable Non-Small Cell Lung Cancer,’ *Lung Cancer*, vol. 29, pp. 169–177, 2000.

[218] F. Balkwill and A. Mantovani, ‘Inflammation and Cancer: Back to Virchow?’ *Lancet*, vol. 357, pp. 539–545, 2001.

[219] M. R. Galdiero, C. Garlanda, S. Jaillon *et al.*, ‘Tumor Associated Macrophages and Neutrophils in Tumor Progression,’ *Journal of Cellular Physiology*, vol. 228, pp. 1404–1412, 2013.

[220] H. A. Smith and Y. Kang, ‘The Metastasis-Promoting Roles of Tumor-Associated Immune Cells,’ *Journal of Molecular Medicine*, vol. 91, pp. 411–429, 2013.

[221] H. Liu, X. Gu, X. Ma *et al.*, 'Preoperative Platelet Count in Predicting Lymph Node Metastasis and Prognosis in Patients With Non-Small Cell Lung Cancer,' *Neoplasma*, vol. 60, pp. 203–208, 2013.

[222] Y. Li, H. Jia, W. Yu *et al.*, 'Nomograms for Predicting Prognostic Value of Inflammatory Biomarkers in Colorectal Cancer Patients After Radical Resection,' *International Journal of Cancer*, vol. 139, pp. 220–231, 2016.

[223] N. A. Cannon, J. Meyer, P. Iyengar *et al.*, 'Neutrophil–lymphocyte and Platelet–lymphocyte Ratios as Prognostic Factors After Stereotactic Radiation Therapy for Early-Stage Non–small-Cell Lung Cancer,' *Journal of Thoracic Oncology*, vol. 10, pp. 280–285, 2015.

[224] H. Gittleman, D. Lim, M. W. Kattan *et al.*, 'An Independently Validated Nomogram for Individualized Estimation of Survival Among Patients With Newly Diagnosed Glioblastoma: NRG Oncology RTOG 0525 and 0825,' *Neuro-Oncology*, vol. 19, pp. 669–677, 2017.

[225] Y.-H. Yuan, Q.-S. Sun, Q. Zhou *et al.*, 'A Novel Multiset Integrated Canonical Correlation Analysis Framework and Its Application in Feature Fusion,' *Pattern Recognition*, vol. 44, no. 5, pp. 1031–1040, 2011.

[226] H. Zhao, Z. Ding and Y. Fu, 'Multi-View Clustering via Deep Matrix Factorization,' in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.

[227] L. v. d. Maaten and G. Hinton, 'Visualizing Data Using T-Sne,' *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[228] Q.-S. Sun, S.-G. Zeng, Y. Liu *et al.*, 'A New Method of Feature Fusion and Its Application in Image Recognition,' *Pattern Recognition*, vol. 38, no. 12, pp. 2437–2448, 2005.

[229] J. Yang, J.-y. Yang, D. Zhang *et al.*, 'Feature Fusion: Parallel Strategy vs. Serial Strategy,' *Pattern Recognition*, vol. 36, no. 6, pp. 1369–1381, 2003.

[230] X. Liu, Y. Dou, J. Yin *et al.*, 'Multiple Kernel K-Means Clustering With Matrix-Induced Regularization,' in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016.

[231] H. He, Y. Bai, E. A. Garcia *et al.*, 'ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning,' in *International Joint Conference on Neural Networks*, 2008, pp. 1322–1328.

[232] Y. LeCun, L. Bottou, Y. Bengio *et al.*, 'Gradient-Based Learning Applied to Document Recognition,' *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[233] J. R. Westbury, G. Turner and J. Dembowski, 'X-Ray Microbeam Speech Production Database User's Handbook,' *University of Wisconsin*, 1994.

[234] J. Ngiam, A. Khosla, M. Kim *et al.*, 'Multimodal Deep Learning,' in *Proceedings of the International Conference on Machine Learning*, 2011, pp. 689–696.

[235] H. W. Kuhn, 'The Hungarian Method for the Assignment Problem,' *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[236] W. Wang, X. Yan, H. Lee *et al.*, 'Deep Variational Canonical Correlation Analysis,' *arXiv Preprint arXiv:1610.03454*, 2017.

[237] V. L. Feigin, A. A. Abajobir, K. H. Abate *et al.*, 'Global, Regional, and National Burden of Neurological Disorders During 1990-2015: A Systematic Analysis for the Global Burden of Disease Study 2015,' *Lancet Neurology*, vol. 16, no. 11, pp. 877–97, 2017.

[238] D. S. Reich, C. F. Lucchinetti and P. A. Calabresi, 'Multiple Sclerosis,' *New England Journal of Medicine*, vol. 378, no. 2, pp. 169–180, 2018.

[239] C. Barillot, G. Edan and O. Commowick, 'Imaging Biomarkers in Multiple Sclerosis: From Image Analysis to Population Imaging,' *Medical Image Analysis*, vol. 33, pp. 134–139, 2016.

[240] F. Barkhof, 'The Clinico-Radiological Paradox in Multiple Sclerosis Revisited,' *Current Opinion in Neurology*, vol. 15, no. 3, pp. 239–245, 2002.

[241] M. Habes, A. Sotiras, G. Erus *et al.*, 'White Matter Lesions: Spatial Heterogeneity, Links to Risk Factors, Cognition, Genetics, and Atrophy,' *Neurology*, vol. 91, no. 10, e964–e975, 2018.

[242] C. Lucchinetti, W. Brück, J. Parisi *et al.*, 'Heterogeneity of Multiple Sclerosis Lesions: Implications for the Pathogenesis of Demyelination,' *Annals of Neurology: Official*

Journal of the American Neurological Association and the Child Neurology Society, vol. 47, no. 6, pp. 707–717, 2000.

[243] V. G. Young, G. M. Halliday and J. J. Kril, 'Neuropathologic Correlates of White Matter Hyperintensities,' *Neurology*, vol. 71, no. 11, pp. 804–811, 2008.

[244] M. Juryńczyk, G. Tackley, Y. Kong *et al.*, 'Brain Lesion Distribution Criteria Distinguish MS From AQP4-antibody NMOSD and MOG-antibody Disease,' *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 88, no. 2, pp. 132–136, 2017.

[245] M. Muthuraman, J. Kroth, D. Ciolac *et al.*, 'Lesion Patterns Topology Is Associated With Regional Cortical Atrophy and Predicts Disease-Related Disability,' *Multiple Sclerosis Journal*, vol. 24, pp. 205–205, 2018.

[246] A. Altermatt, L. Gaetano, S. Magon *et al.*, 'Clinical Correlations of Brain Lesion Location in Multiple Sclerosis: Voxel-Based Analysis of a Large Clinical Trial Dataset,' *Brain Topography*, vol. 31, no. 5, pp. 886–894, 2018.

[247] M. Vellinga, J. Geurts, E. Rostrup *et al.*, 'Clinical Correlations of Brain Lesion Distribution in Multiple Sclerosis,' *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 29, no. 4, pp. 768–773, 2009.

[248] C. P. Loizou, M. Pantzaris and C. S. Pattichis, 'Normal Appearing Brain White Matter Changes in Relapsing Multiple Sclerosis: Texture Image and Classification Analysis in Serial MRI Scans,' *Magnetic Resonance Imaging*, vol. 73, pp. 192–202, 2020.

[249] T. Frisch, M. L. Elkjaer, R. Reynolds *et al.*, 'Multiple Sclerosis Atlas: A Molecular Map of Brain Lesion Stages in Progressive Multiple Sclerosis,' *Network and Systems Medicine*, vol. 3, no. 1, pp. 122–129, 2020.

[250] M. Lombardo, R. Barresi, E. Bilotta *et al.*, 'Demyelination Patterns in a Mathematical Model of Multiple Sclerosis,' *Journal of Mathematical Biology*, vol. 75, no. 2, pp. 373–417, 2017.

[251] K. Bendfeldt, J. O. Blumhagen, H. Egger *et al.*, 'Spatiotemporal Distribution Pattern of White Matter Lesion Volumes and Their Association With Regional Grey Matter Volume Reductions in Relapsing-Remitting Multiple Sclerosis,' *Human Brain Mapping*, vol. 31, no. 10, pp. 1542–1555, 2010.

[252] A. Doyle, D. Precup, D. L. Arnold *et al.*, 'Predicting Future Disease Activity and Treatment Responders for Multiple Sclerosis Patients Using a Bag-of-Lesions Brain Representation,' in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2017, pp. 186–194.

[253] A. Eshaghi, S. Riyahi-Alam, R. Saeedi *et al.*, 'Classification Algorithms With Multi-Modal Data Fusion Could Accurately Distinguish Neuromyelitis Optica From Multiple Sclerosis,' *NeuroImage: Clinical*, vol. 7, pp. 306–314, 2015.

[254] C. Louapre, B. Bodini, C. Lubetzki *et al.*, 'Imaging Markers of Multiple Sclerosis Prognosis,' *Current Opinion in Neurology*, vol. 30, no. 3, pp. 231–236, 2017.

[255] B. Taschler, T. Ge, K. Bendfeldt *et al.*, 'Spatial Modeling of Multiple Sclerosis for Disease Subtype Prediction,' in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2014, pp. 797–804.

[256] V. Wottschel, D. Alexander, P. Kwok *et al.*, 'Predicting Outcome in Clinically Isolated Syndrome Using Machine Learning,' *NeuroImage: Clinical*, vol. 7, pp. 281–287, 2015.

[257] C. Grana, D. Borghesani and R. Cucchiara, 'Optimized Block-Based Connected Components Labeling With Decision Trees,' *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1596–1609, 2010.

[258] F. Battiston, G. Cencetti, I. Iacopini *et al.*, 'Networks Beyond Pairwise Interactions: Structure and Dynamics,' *Physics Reports*, 2020.

[259] A. Zomorodian, 'Fast Construction of the Vietoris-Rips Complex,' *Computers & Graphics*, vol. 34, no. 3, pp. 263–271, 2010.

[260] S. Fortunato and D. Hric, 'Community Detection in Networks: A User Guide,' *Physics Reports*, vol. 659, pp. 1–44, 2016.

[261] S. Fortunato, 'Community Detection in Graphs,' *Physics Reports*, vol. 486, no. 3-5, pp. 75–174, 2010.

[262] G. Palla, I. Derényi, I. Farkas *et al.*, 'Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society,' *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[263] A. Amelio and C. Pizzuti, ‘Overlapping Community Discovery Methods: A Survey,’ in *Social networks: Analysis and case studies*, Springer, 2014, pp. 105–125.

[264] H. Shen, X. Cheng, K. Cai *et al.*, ‘Detect Overlapping and Hierarchical Community Structure in Networks,’ *Physica A: Statistical Mechanics and its Applications*, vol. 388, no. 8, pp. 1706–1712, 2009.

[265] S. Jabbour, N. Mhadhbi, B. Radaoui *et al.*, ‘Detecting Highly Overlapping Community Structure by Model-Based Maximal Clique Expansion,’ in *2018 IEEE International Conference on Big Data (Big Data)*, IEEE, 2018, pp. 1031–1036.

[266] B. Rieck, U. Fugacci, J. Lukasczyk *et al.*, ‘Clique Community Persistence: A Topological Visual Analysis Approach for Complex Networks,’ *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 822–831, 2017.

[267] R. Seidel and M. Sharir, ‘Top-Down Analysis of Path Compression,’ *SIAM Journal on Computing*, vol. 34, no. 3, pp. 515–525, 2005.

[268] X. Yang, Q. Song and Y. Wang, ‘A Weighted Support Vector Machine for Data Classification,’ *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 21, no. 05, pp. 961–976, 2007.

[269] K. H. Brodersen, C. S. Ong, K. E. Stephan *et al.*, ‘The Balanced Accuracy and Its Posterior Distribution,’ in *International Conference on Pattern Recognition*, IEEE, 2010, pp. 3121–3124.

[270] A. Eshaghi, V. Wottschel, R. Cortese *et al.*, ‘Gray Matter MRI Differentiates Neuromyelitis Optica From Multiple Sclerosis Using Random Forest,’ *Neurology*, vol. 87, no. 23, pp. 2463–2470, 2016.

[271] Y. Zhao, B. C. Healy, D. Rotstein *et al.*, ‘Exploration of Machine Learning Techniques in Predicting Multiple Sclerosis Disease Course,’ *PLoS One*, vol. 12, no. 4, e0174866, 2017.

[272] M. T. Law, A. L. Traboulsee, D. K. Li *et al.*, ‘Machine Learning in Secondary Progressive Multiple Sclerosis: An Improved Predictive Model for Short-Term Disability Progression,’ *Multiple Sclerosis Journal–Experimental, Translational and Clinical*, vol. 5, no. 4, p. 2 055 217 319 885 983, 2019.

[273] A. Tousignant, P. Lematre, D. Precup *et al.*, 'Prediction of Disease Progression in Multiple Sclerosis Patients Using Deep Learning Analysis of MRI Data,' in *International Conference on Medical Imaging With Deep Learning*, 2019, pp. 483–492.

[274] A. Eshaghi, F. Prados, W. J. Brownlee *et al.*, 'Deep Gray Matter Volume Loss Drives Disability Worsening in Multiple Sclerosis,' *Annals of Neurology*, vol. 83, no. 2, pp. 210–222, 2018.

[275] P. Schmidt, 'Bayesian Inference for Structured Additive Regression Models for Large-Scale Problems With Applications to Medical Imaging,' Ph.D. dissertation, lmu, 2017.

[276] L. Zhang and W. Zhou, 'On the Sparseness of 1-Norm Support Vector Machines,' *Neural Networks*, vol. 23, no. 3, pp. 373–385, 2010.

[277] M. Neumann, R. Garnett, C. Bauckhage *et al.*, 'Propagation Kernels: Efficient Graph Kernels From Propagated Information,' *Machine Learning*, vol. 102, no. 2, pp. 209–245, 2016.

[278] T. N. Kipf and M. Welling, 'Semi-Supervised Classification With Graph Convolutional Networks,' *International Conference on Representation Rearning*, 2016.

[279] T. G. Dietterich, 'Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms,' *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[280] M. Styner, J. Lee, B. Chin *et al.*, '3d Segmentation in the Clinic: A Grand Challenge II: MS Lesion Segmentation,' *Midas Journal*, vol. 2008, pp. 1–6, 2008.

[281] O. Commowick, F. Cervenansky and R. Ameli, 'Msseg Challenge Proceedings: Multiple Sclerosis Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure,' in *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2016.

[282] F. Frasca, E. Rossi, D. Eynard *et al.*, 'Sign: Scalable Inception Graph Neural Networks,' *arXiv preprint arXiv:2004.11198*, 2020.

[283] A. Kazi, S. Shekarforoush, S. A. Krishna *et al.*, 'Inceptiongcn: Receptive Field Aware Graph Convolutional Network for Disease Prediction,' in *International Conference on Information Processing in Medical Imaging*, Springer, 2019, pp. 73–85.

[284] L. Cosmo, A. Kazi, S.-A. Ahmadi *et al.*, ‘Latent-Graph Learning for Disease Prediction,’ in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2020, pp. 643–653.

[285] H. Cai, V. W. Zheng and K. C.-C. Chang, ‘A Comprehensive Survey of Graph Embedding: Problems, Techniques, and Applications,’ *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 9, pp. 1616–1637, 2018.

[286] J. M. Johnson and T. M. Khoshgoftaar, ‘Survey on Deep Learning With Class Imbalance,’ *Journal of Big Data*, vol. 6, no. 1, pp. 1–54, 2019.

[287] C. Köhler, H. Wahl, T. Ziemssen *et al.*, ‘Exploring Individual Multiple Sclerosis Lesion Volume Change Over Time: Development of an Algorithm for the Analyses of Longitudinal Quantitative MRI Measures,’ *NeuroImage: Clinical*, vol. 21, p. 101 623, 2019.

[288] A. E. Sizemore, C. Giusti, A. Kahn *et al.*, ‘Cliques and Cavities in the Human Connectome,’ *Journal of Computational Neuroscience*, vol. 44, no. 1, pp. 115–145, 2018.

[289] M. Filippi, P. Preziosa, B. L. Banwell *et al.*, ‘Assessment of Lesions on Magnetic Resonance Imaging in Multiple Sclerosis: Practical Guidelines,’ *Brain*, vol. 142, no. 7, pp. 1858–1875, 2019.