UNCERTAINTY MITIGATION IN IMAGE-BASED MACHINE LEARNING MODELS FOR PRECISION MEDICINE

A Dissertation
Presented to
The Academic Faculty

by

Lujia Wang

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
August, 2022

UNCERTAINTY MITIGATION IN IMAGE-BASED MACHINE LEARNING MODELS FOR PRECISION MEDICINE

Approved by:

Dr. Jing Li, Advisor
School of Industrial & Systems Engineering
Georgia Institute of Technology

Dr. Yajun Mei
School of Industrial & Systems Engineering
Georgia Institute of Technology

Dr. Jianjun Shi
School of Industrial & Systems Engineering
Georgia Institute of Technology

Dr. Nathan Gaw
Department of Operational Sciences
Air Force Institute of Technology

Dr. Kamran Paynabar
School of Industrial & Systems Engineering
Georgia Institute of Technology

Date Approved: July 21, 2022

To all of the people that light up my life.

ACKNOWLEDGEMENTS

I would like to express my heartfelt thanks to many people who guided me, encouraged me, supported and helped me over the years through my doctoral studies. First and foremost, I would like to express my deepest gratitude to my advisor, Dr. Jing Li, for her continuous support, invaluable wisdom, tremendous guidance, immense help and encouragement over the past years. Her continual guidance and support, as well as her genuine passion for research, is inspirational. It is an honor and great pleasure to work under her supervision and insightful guidance.

I would like to express my special thanks to Dr. Jianjun Shi, who have provided very valuable suggestions and invaluable supports to my research and future life. I am very grateful to Dr. Nathan Gaw, who have given me a lot of encouragements to overcome the obstacles in academic research and daily life. I also want to appreciate my thesis committee members, Dr. Kamran Paynabar and Dr. Yajun Mei, for spending their precious time on the examination of my thesis research. Many thanks for their support, guidance and suggestions.

Special thanks go to my co-workers/collaborators in Mayo Clinic, Drs. Leland S. Hu, Kristin R. Swanson, Andrea Hawkins-Daarud, Nhan L. Tran, Kyle W. Singleton, and Lee Curtin. I appreciate their patience, constructive criticism, and valuable instruction. The experience we work together is so valuable to me. Their technical and academic depth broadened my horizons.

Finally, I would like to thank my family members for their unconditional love and unending support during this journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Machine learning (ML) algorithms have been developed to build predictive models in medicine and healthcare. In most cases, the performance of ML models/algorithms is measured by predictive accuracy or accuracy-related measures only. In medicine, the model results are intended to guide physicians to make critical decisions regarding patient care. This means that quantifying and mitigating the uncertainty of the output is also very important as it will allow decision makers to know how much they can rely on the model output.

My dissertation focuses on studying model uncertainty of image-based ML in the context of precision medicine of brain cancer. Specifically, I focus on developing ML models to predict intra-tumor heterogeneity of genomic and molecular markers based on multi-contrast magnetic resonance imaging (MRI) data for glioblastoma (GBM) – the most aggressive type of brain cancer. Intra-tumor heterogeneity has been found to be a leading cause of treatment failure of GBM. Devising a non-invasive approach to map out the molecular/genomic distribution using MRI helps develop treatment with high precision. My dissertation research addresses the model uncertainties due to high-dimensional and noisy features, sparsity of labeled data, and utility of domain knowledge.

In the first study, we developed a Semi-supervised Gaussian Process with Uncertainty-minimizing Feature-selection (SGP-UF), which can incorporate selected unlabeled samples (i.e. unbiopsied regions of a tumor) in the model training, and integrate feature selection with a new criterion of seeking features that minimize the prediction uncertainty.

In the second study, we developed a Knowledge-infused Global-Local data fusion (KGL) framework, which optimally fuses three sources of data/information including biopsy samples (labeled data, local/sparse), images (unlabeled data, global), and knowledge-driven mechanistic models.

In the third study, we developed a Weakly Supervised Ordinal Support Vector Machine (WSO-SVM), which aims to leverage a combination of data sources including biopsy/labeled samples and unlabeled samples from the tumor and image data from the normal brain, as well as their intrinsic ordinal relationship.

We demonstrate that these novel methods significantly reduce prediction uncertainty while at the same time achieving higher accuracy in precision medicine, which can inform personalized targeted treatment decisions that potentially improve clinical outcome.

CHAPTER 1. INTRODUCTION

1.1 Background

1.1.1 Machine Learning and Model Uncertainties

Thanks to advances in computing power and the increasing availability of digital data, machine learning (ML) is gaining great popularity in recent years. It is widely applied to various fields, like weather forecasting, financial services, social media services and healthcare. In most cases, the performance of ML models/algorithms is measured by predictive accuracy or accuracy-related measures only. Recently, more and more research has started to address the importance of understanding the uncertainties of ML models. Since the output of ML models is typically used to facilitate subsequent decision-making, it is critical for decision makers to know how much they can rely on the model output.

Uncertainty is usually classified into two distinct categories: aleatoric uncertainty and epistemic uncertainty. The decomposition of different uncertainty is shown in Figure 1.1. Aleatoric uncertainty is also known as numerical uncertainty, discrete uncertainty, physical variability, irreducible uncertainty, inherent uncertainty, stochastic uncertainty and uncertainty due to chance [1][2]. This type of uncertainty comes from the variability due to numerical measurement errors or numerical approximations from the experimental systems. It is the variation inherent to the system. Epistemic uncertainty is also known as systematic uncertainty, reducible uncertainty, subjective uncertainty, and uncertainty due to lack of knowledge [3]. It stems from incomplete knowledge (lack of knowledge from the insufficient or biased data) about some characteristic of a system or phenomenon in

any phase or activity of the modeling process [4]. The epistemic uncertainty of a model is reducible with more collected data or increased information [5]. According to [6], epistemic uncertainty can be classified into three categories: parametric uncertainty, structural or model uncertainty, and subjective judgment. Parametric uncertainty comes from the variability of model parameters under the condition that the model is almost always well given [7]. Structural or model uncertainty is associated with the accuracy a mathematical model has to describe the real-world problem [7]. Subjective judgment comes from uncertainties in experts' opinions on the interpretations of data [6]. My dissertation focuses on studying model uncertainty of image-based ML in the application of precision medicine of brain cancer.

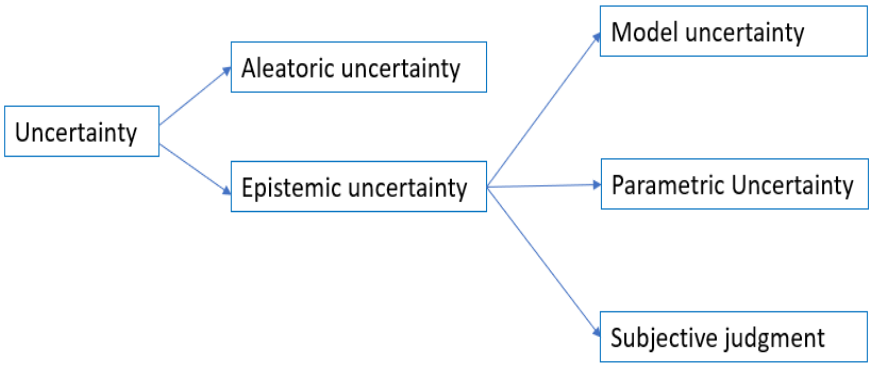

Figure 1.1 – Summary of different types of uncertainty

1.1.2 Application Background

The study of images becomes popular in the field of ML with the rapid development of computer vision and image recognition. The ability for ML techniques with strong subtle imaging patterns discrimination can support better decision making, especially in health

care with medical imaging (e.g., CT or MRI) for diagnostic accuracy improvement and clinical efficiency.

My dissertation research focuses on developing imaging-based ML models to characterize intra-tumor heterogeneity of brain cancer, in particular glioblastoma (GBM). GBM is the most aggressive type of brain tumor with median survival of 15 months [8]. Intra-tumor molecular heterogeneity has been found to be one of the leading causes of treatment failure. The invasive nature of biopsy makes it impossible to sample every sub-region to understand the regionally-specific molecular characteristics. Neuroimaging such as magnetic resonance imaging (MRI) portrays the entire brain non-invasively, providing the opportunity to estimate/predict the spatial molecular distributions across each individual tumor, such as the spatial distributions of tumor cell density and aberrations of driver genes. Such capability will enable unprecedented precision of treatment: radiation therapy can be spatially optimized to avoid over- and under-treating certain regions of the brain, gene therapy can be adapted to regional aberration patterns, etc.

On the other hand, since critical therapeutic decisions would heavily rely on the spatial molecular maps predicted by ML based on imaging, it is critically important to understand and reduce the uncertainty of the ML models. This will enable greater confidence of the clinicians in using the ML output to assist their decision making and lead to greater transferability of the models into clinical practice.

1.1.3 Subtypes of Model Uncertainty Driven by the Application Need

1) High-dimensional and noisy features: To build an ML model that uses MRI to predict spatial molecular distribution, the first step is to extract features from MRI images.

Many features can be extracted by applying various kinds of texture analysis algorithms. However, not all the features will contribute to the prediction of regional molecular status, i.e., they are noise. While feature selection has been studied by many ML researchers, the majority of the existing work focuses on selecting features to enhance accuracy, but not to reduce the uncertainty of the model prediction.

2) Sparsity of labeled data: In contrast with the large number of features, the labeled data is sparse in our application. Here, labeled data refers to biopsy samples acquired from the tumoral area of each patient's brain, which provide direct measurement for the molecular marker of interest. Due to the invasive nature of biopsy, only a few samples can be obtained from each patient. The small sample size leads to greater uncertainty of the ML model.

3) Uncertainty in domain knowledge: Domain knowledge is the background knowledge of a specific, specialized discipline or field or environment. However, domain knowledge usually has multiple sources of uncertainty, such as the approximate or incomplete estimation from mechanistic models, useful knowledge unexplored, and intuitive uncertainty in assessment of physical uncertainty by experts. It is a significant challenge for ML to take advantage of the domain knowledge.

1.2 Expected Original Contribution

My dissertation research develops ML models to address the aforementioned subtypes of model uncertainty within the context of using imaging to characterize intra-tumor heterogeneity of GBM to facilitate precision medicine. The following original contributions are expected:

- **Development of a Semi-supervised Gaussian Process with Uncertainty-minimizing Feature-selection (SGP-UF) (Chapter 2).** The methodological contributions are: SGP-UF is an extension of GP with two innovations: 1) While GP is a supervised learning model (i.e., the model is the trained using labeled samples only), SGP-UP can additionally incorporate selected unlabeled samples (i.e. unbiopsied regions of a tumor) in the model training. Since unbiopsied regions are also what the trained model will predict for, including the image features from these regions help mitigate the risk of extrapolation and therefore reduce the prediction uncertainty. 2) SGP-UP integrates feature selection with a new criterion of seeking features that minimize the prediction uncertainty. The practical impact of this work is: we apply both SGP-UP to the prediction of regional EGFR amplification status for each GBM tumor using MRI. EGFR is a well-known GBM driver gene and serves as a common therapeutic target for many clinically available drugs. We demonstrate that, by incorporating unlabeled samples and a uncertainty-driven criterion of features selection in model training, SGP-UP significantly reduces prediction uncertainty while at the same time achieving higher accuracy (i.e., lower prediction error between the true EGFR status and the predictive mean of the EGFR status in the predictive distribution).

- **Development of a Knowledge-infused Global-Local data fusion (KGL) framework (Chapter 3)**: The methodological contributions are: KGL optimally fuses three sources of data/information including biopsy samples (labeled data, local/sparse), images (unlabeled data, global), and knowledge-driven mechanistic models. The key idea of KGL is to build a predictive model that uses global imagery to predict the regional distribution for molecular markers, where the model parameters are optimized to simultaneously serve three purposes: 1) maximizing the accuracy on labeled samples (i.e., regions with direct measurement); 2) reducing the prediction uncertainty on unlabeled samples (i.e., regions without direct measurement but only imagery); 3) being consistent with the trend or patterns conveyed by domain knowledge. The practical impact of this work is: We present a real-data application of predicting the spatial distribution of Tumor Cell Density (TCD)—an important molecular marker for brain cancer. A total of 82 biopsy samples were acquired from 18 patients with glioblastoma, together with 6 MRI contrast images from each patient and biological knowledge encoded by a PDE simulator-based mechanistic model called Proliferation-Invasion (PI). KGL achieved 12-14% reduction of prediction error and over 60% reduction of prediction uncertainty compared with competing methods. The result has important implications for providing individualized, spatially-optimized treatment for each patient.
- **Development of a Weakly Supervised Ordinal Support Vector Machine (WSO-SVM) model (Chapter 4):** The methodological contributions are: WSO-SVM aims to leverage a combination of data sources including

biopsy/labeled samples and unlabeled samples from the tumor and image data from the normal brain, as well as their intrinsic ordinal relationship. The key idea of WSO-SVM is to leverage unlabeled samples in tumor ROI and impose a mathematical constraint to "teach" the model that these samples should not be classified to normal brain even though their true membership is unknown. The practical impact of this work is: WSO-SVM was applied to a unique dataset of 318 image-localized biopsies with spatially matched multiparametric MRI from 74 GBM patients. The model was trained to predict the regional genetic alteration of three GBM driver genes (EGFR, PDGFRA and PTEN) based on features extracted from the corresponding region of five MRI contrast images. WSO-SVM achieved higher accuracies, outperforming the existing ML algorithms. The accuracies improved with higher certain prediction. The prediction maps revealed a great amount of variability between patients in terms of the genetic alteration patterns. Within each individual's tumor, there was also region-to-region variation for the genetic alteration patterns.

CHAPTER 2. SEMI-SUPERVISED GAUSSIAN PROCESS WITH UNCERTAINTY-MINIMIZING FEATURE SELECTION FOR INTRA-TUMOR GENOMIC PREDICTION USING MRI

*This chapter is based on my published paper "*Hu, L.S., *Wang, L., Hawkins-Daarud, A., ..., *Swanson, K.R., *Li, J. (2021) Uncertainty Quantification in the Radiogenomics Modeling of EGFR Amplification in Glioblastoma. Scientific Reports, 11(1): 1-14. (*Contributed equally)".*

2.1 Background

The field of machine-learning (ML) has exploded in recent years, thanks to advances in computing power and the increasing availability of digital data. Some of the most exciting developments in ML have centered on computer vision and image recognition, with broad applications ranging from e-commerce to self-driving cars. But these advances have also naturally dovetailed with applications in healthcare, and in particular, the study of medical images. The ability for computer algorithms to discriminate subtle imaging patterns has led to myriad ML tools aimed at improving diagnostic accuracy and clinical efficiency [9].

One of the most transformative applications of imaging-based ML is to use images such as MRI to predict genetic aberrations of the tumor of each patient with cancer. This field is known as radiogenomics [10]. In the context of individualized oncology, radiogenomics non-invasively diagnoses the unique genetic drug sensitivities for each

patient's tumor, which can inform personalized targeted treatment decisions that potentially improve clinical outcome.

My dissertation research focuses on the most aggressive type of brain tumor called GBM. One of the leading factors of treatment failure of GBM is intra-tumoral genomic heterogeneity, meaning that different sub-regions of each tumor may have different genetic status. As a result, the different sub-regions may respond to the treatment differently. This means that the treatment needs to be adapted to not only patient difference but also regional difference within each individual tumor. Although radiogenomics has been a popular research area, there is less research that uses images to predict the genetic status across different sub-regions of each tumor, while the majority of the existing research has focused on predicting an average/overall genetic status of the tumor [11][12][13].

My research aims to develop an ML model to predict regional genetic status using images (e.g., MRI). Given the training model, the ultimate goal is to produce a predicted map for the regional genetic distribution of each tumor. This map can assist physicians to make appropriate treatment decisions for each patient. However, there are several challenging issues for accomplishing this goal:

1) The sample size of labeled data is inherently small. Labeled data in this context refers to biopsy samples from each patient. Due to the invasive nature of biopsy, only a few samples can be obtained per patient. If a model is trained on the limited biopsy samples and used to generate predictions on the large number of unbiopsied regions of a tumor, the predictions may be highly uncertainty due to the risk of extrapolation. As with any data-driven approach,

the scope of training data establishes the upper and lower bounds of the model domain, which guides the predictions for all new unseen test cases. In the ideal scenario, the new test data will fall within the distribution of the training domain, which allows for interpolation of model predictions, and the lowest degree of predictive uncertainty. If the test data fall outside of the training domain, then the model must extrapolate predictions, at the cost of greater model uncertainty. In our application, since the predictions on test samples (i.e., the large number of unbiopsied samples from each tumor) will be used to guide treatment decisions, it is critically important to reduce the risk of extrapolation and the prediction uncertainty.

2) Many features can be extracted from an image and some (or even majority of them) may be noise. Selecting the right features is important for building the ML model to predict regional genetic status. As uncertainty reduction of the model prediction is critical in our application, feature selection shall be geared toward this objective. However, although feature selection is a popular research area in ML [14][15], most existing algorithms were designed to maximize the accuracy of prediction (i.e., minimizing the prediction error between the true and predicted response variables), but not to reduce the uncertainty of the prediction.

To address these challenges, we propose a new model called Semi-supervised Gaussian Process with Uncertainty-minimizing Feature selection (SGP-UF). GP is a well-known model for its capability of generating a predictive distribution to quantify the uncertainty of the prediction [16][17]. The proposed SGP-UF is an extension of GP with

two innovations: 1) While GP is a supervised learning model (i.e., the model is the trained using labeled samples only), SGP-UP can additionally incorporate selected unlabeled samples (i.e. unbiopsied regions of a tumor) in the model training. Since unbiopsied regions are also what the trained model will predict for, including the image features from these regions help mitigate the risk of extrapolation and therefore reduce the prediction uncertainty. 2) SGP-UP integrates feature selection with a new criterion of seeking features that minimize the prediction uncertainty.

As a case study, we apply both SGP-UP and GP (as a competing method) to the prediction of regional EGFR amplification status for each GBM tumor using MRI. EGFR is a well-known GBM driver gene and serves as a common therapeutic target for many clinically available drugs. We demonstrate that, by incorporating unlabeled samples and a uncertainty-driven criterion of features selection in model training, SGP-UP significantly reduces prediction uncertainty while at the same time achieving higher accuracy (i.e., lower prediction error between the true EGFR status and the predictive mean of the EGFR status in the predictive distribution). Our overarching goal of this work is to provide a pathway to clinically integrating reliable regional genomic predictions and intra-tumor heterogeneity characterization as part of decision support within the paradigm of individualized oncology.

2.2 Introduction of Gaussian Process (GP)

A GP model offers the flexibility of identifying nonlinear relationships between input and output variables [18][19]. More importantly, it generates a probability

distribution for each prediction, which quantifies the uncertainty of the prediction. This capability is important for using the prediction result to guide clinical decision making.

Next we illustrate how GP works. GP is a supervised learning mode, which required a labeled dataset for training. Assume the training dataset includes N samples, which are to N biopsy samples in our application. Let $\{y_1, \dots, y_N\}$ be the response variables (e.g., measurements of the EGFR gene) of the N samples. Let $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ be the input variables, where $\mathbf{x}_i$ contains the image features extracted from the region of the i-th biopsy sample, $i = 1, \dots, N$. GP assumes a set of random functions corresponding to the samples, $\{f(\mathbf{x}_1), \dots, f(\mathbf{x}_N)\}$. This is called a Gaussian Process because any subset of $\{f(\mathbf{x}_1), \dots, f(\mathbf{x}_N)\}$ follows a joint Gaussian distribution with zero mean and the covariance between two samples $\mathbf{x}_i$ and $\mathbf{x}_j$ computed by a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$. The input variables are linked with the response by $y_i = f(\mathbf{x}_i) + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$ is a Gaussian noise.

Let $\{\mathbf{X}_L, \mathbf{y}_L\}$ denote a training dataset, where $\mathbf{X}_L$ is a matrix containing the input variables of all training samples and $\mathbf{y}_L$ is a vector containing the response variables of these samples. Here, we use a subscript "L" to highlight that GP is trained using labeled data only. This is to distinguish with our proposed model in the next section that can additionally incorporate unlabeled data. Furthermore, let $\boldsymbol{\theta}$ contain the parameters to be estimated for a GP model. $\widehat{\boldsymbol{\theta}}$ can be estimated by maximizing the marginal likelihood of the labeled samples,

$$\min_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} -\log p(\mathbf{y}_L | \mathbf{X}_L, \boldsymbol{\theta})) \tag{2.1}$$

After the parameters are estimated, the trained GP model can be used to generate a predictive distribution for a new test sample $\mathbf{x}^*$ (e.g., $\mathbf{x}^*$ is the image feature vector corresponding to an unbiopsied region), which takes the following form:

$$f(\mathbf{x}^*) \sim N\left(\mu^*, {\sigma^*}^2\right), \tag{2.2}$$

where

$$\mu^* = K(\mathbf{x}^*, \mathbf{X})\left(K(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I}\right)^{-1}\mathbf{Y}, \tag{2.3}$$

$${\sigma^*}^2 = K(\mathbf{x}^*, \mathbf{x}^*) - K(\mathbf{x}^*, \mathbf{X})\left(K(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I}\right)^{-1} K(\mathbf{x}^*, \mathbf{X})^T. \tag{2.4}$$

The predictive mean in (2.2) can be used as a point estimator for the EGFR of the new sample, while the predictive variance in (2.3) reflects the uncertainty of the prediction.

2.3 Development of the proposed SGP-UF

We first discuss the SGP for a given set of features in Section 2.3.1, and then discuss how to integrate feature selection for uncertainty minimization with SGP in Section 2.3.2.

2.3.1 Semi-supervised Gaussian Process (SGP)

Expanding upon the notations of Section 2.2, we use $\{\mathbf{X}_L, \mathbf{Y}_L\}$ and $\{\mathbf{X}_U\}$ to denote the labeled and unlabeled samples used in training, respectively. For a new test sample $\mathbf{x}^*$, denote the predictive distribution by

$$f_{SGP}(\mathbf{x}^*) \sim N\left(\mu^*_{SGP}, {\sigma^*_{SGP}}^2\right). \tag{2.5}$$

The predictive variance of a GP model does not utilize the response variables in the training data but only the input variables, as can be seen from (2.4). Because of this, we can combine the input variables of both labeled and unlabeled samples and derive ${\sigma^*_{SGP}}^2$ using a similar formula as (2.4), i.e.,

$${\sigma^*_{SGP}}^2 = K(\mathbf{x}^*, \mathbf{x}^*) - \begin{pmatrix} K(\mathbf{x}^*, \mathbf{X}_L) \\ K(\mathbf{x}^*, \mathbf{X}_U) \end{pmatrix}^T \begin{pmatrix} K(\mathbf{X}_L, \mathbf{X}_L) + \sigma^2 \mathbf{I} & K(\mathbf{X}_L, \mathbf{X}_U) \\ K(\mathbf{X}_L, \mathbf{X}_U)^T & K(\mathbf{X}_U, \mathbf{X}_U) + \sigma^2 \mathbf{I} \end{pmatrix}^{-1} \begin{pmatrix} K(\mathbf{x}^*, \mathbf{X}_L) \\ K(\mathbf{x}^*, \mathbf{X}_U) \end{pmatrix}. \quad (2.6)$$

We consider μ^*_{SGP} to be weighted sum of the labeled and unlabeled samples, i.e., $\mu^*_{SGP} = \mathbf{w}_L^T k(x^*, \mathbf{X}_L) + \mathbf{w}_U^T k(x^*, \mathbf{X}_U)$ with the weights being $\mathbf{w}_L = (K(\mathbf{X}_L, \mathbf{X}_L) + \sigma^2 I)^{-1}\mathbf{Y}_L$ and $\mathbf{w}_U = (K(\mathbf{X}_U, \mathbf{X}_U) + \sigma^2 I)^{-1}\mathbf{Y}_U$. Since $\mathbf{Y}_U$ is unknown, we use the predictive means of the unlabeled samples based on a GP trained on the labeled samples as an estimate, i.e., $\widehat{\mathbf{Y}}_U = K(\mathbf{X}_L, \mathbf{X}_U)(K(\mathbf{X}_L, \mathbf{X}_L) + \sigma^2 I)^{-1}\mathbf{Y}_L$. Using this estimate and through some algebra, we get the final formula for μ^*_{SGP}, i.e.,

$$\mu^*_{SGP} = K_{SGP}(K(\mathbf{X}_L, \mathbf{X}_L) + \sigma^2 I)^{-1}\mathbf{Y}_L,$$

with

$$K_{SGP} = K(\mathbf{x}^*, \mathbf{X}_L) + K(\mathbf{x}^*, \mathbf{X}_U)(K(\mathbf{X}_U, \mathbf{X}_U) + \sigma^2 I)^{-1}K(\mathbf{X}_U, \mathbf{X}_L).$$

The SGP model was originally proposed by [20] as an empirical procedure, but no theoretical justification was provided as to why and in which aspects SGP outperforms GP. We provide some theoretical analysis in Theorem 2.1 and 2.2. The proofs are shown in Appendices A and B.

Theorem 2.1: When applying both GP and SGP to predict a test sample $\mathbf{x}^*$, the predictive variance of SGP is no greater than GP, i.e., $\sigma_{SGP}^{*\,2} < \sigma^{*2}$.

Theorem 2.2: Consider a test sample $\mathbf{x}^*$. Let $\{\mathbf{X}_L, \mathbf{Y}_L\}$ be the set of labeled samples used in training by GP. Let $\{\mathbf{X}_U\}$ be the set of unlabeled samples used in training by SGP in addition to the labeled set. If $K(\mathbf{X}_U, \mathbf{X}_L) \to \mathbf{0}$ and $K(\mathbf{X}_U, \mathbf{x}^*) \to \mathbf{0}$, i.e., the distances of the unlabeled samples with respect to the labeled samples and the test sample go to infinity, then the predictive distribution for $\mathbf{x}^*$ by SGP, $f_{SGP}(\mathbf{x}^*)$, converges to that by GP, $f(\mathbf{x}^*)$, with respect to Kullback–Leibler divergence , i.e., $f^{SGP} \xrightarrow{D} f$.

A final note in this section is that although in theory SGP-UF can include any samples from unbiopsied regions of a tumor as unlabeled data, we found a reasonable subset of unbiopsied regions to include are the 8 closest neighboring regions of each biopsy region. This approach is used in our application case study.

2.3.2 Integration of SGP and Uncertainty-Minimizing Feature Selection (SGP-UF)

Note that the SGP was proposed under a given set of features. When there is a high-dimensional feature set and many of them may be noise, a feature selection algorithm is needed. There are two key components in a feature selection algorithm: the optimal criterion and the search path. For the former, most existing feature selection algorithms use prediction accuracy. We propose to use prediction uncertainty, i.e., an optimal feature subset is one that minimizes the prediction uncertainty. Specific to SGP, the predictive variance $\sigma_{SGP}^{*\,2}$ reflects the prediction uncertainty. Alternatively, the uncertainty can be computed as a p value from a hypothesis test on the predictive mean. For example, if the

interest is to know whether the predictive mean is smaller or greater than a pre-specified value (considered as H_0 and H_1 of a hypothesis test, respectively), the p value reflects the uncertainty of the test result. The smaller the p value, the less uncertainty for rejecting H_0 (i.e., acknowledging that the predictive mean is greater than the pre-specified value). Although both the predictive variance and the p value reflect uncertainty of the prediction, we found that the p value criterion has a better performance in our application, which will be focused on in the remainder of our discussion in this section.

The second key component of a feature selection algorithm is to design a search path, such that different subsets of the features can be efficiently evaluated using the optimal criterion to find the optimal subset. There are many existing algorithms such as forward selection, backward selection, evolutionary algorithms (e.g., GA), and swarm intelligence algorithms (e.g., PSO). We adopt forward selection due to its capability of generating parsimonious features with fast computation and reasonably good results. Other algorithms are also possible.

A final note is that, to avoid overfitting in feature selection, a commonly used strategy is to evaluate each feature subset along the search path under a cross-validation (CV) scheme. We adopt a specific CV schemed called leave-one-patient-out cross validation (LopoCV), due to the natural grouping of samples within patients. Specifically, for each feature subset on the search path, the samples from $N - 1$ patients are used to train an SGP model using the feature subset. Then, the trained model is used to predict the samples of the remaining patient and the p value of the prediction for each sample can be computed. This training-prediction process is iterated for N times, so that the samples of

each patient can be predicted using other patients' samples for training. In the end, there is a p value associated with the prediction of every sample for each of the N patients. These p values are summed together as a metric to reflect the uncertainty/certainty of the given feature subset. Different feature subsets on the search path can be compared to find the optimal one with the smallest sum of p values.

Once the optimal feature subset is selected, we re-estimate the SGP model parameters using the selected features to maximize the LopoCV prediction accuracy for those samples with low uncertainty (p value<0.05, 0.1, etc.).

2.4 Application

2.4.1 Acquisition and Processing of Clinical MRI and Histologic Data

Patient recruitment and Surgical biopsies: We recruited 25 patients with clinically suspected GBM undergoing preoperative stereotactic MRI for surgical resection as previously described [21]. We confirmed histologic diagnosis of GBM in all cases. We obtained institutional review board approval and informed consent from each subject prior to enrollment. Neurosurgeons used pre-operative conventional MRI, including T1-Weighted contrast-enhanced (T1+C) and T2-Weighted sequences (T2W), to guide multiple stereotactic biopsies as previously described [21][22]. In short, each neurosurgeon collected an average of 3-4 tissue specimens from each tumor using stereotactic surgical localization, following the smallest possible diameter craniotomies to minimize brain shift. Neurosurgeons selected targets separated by at least 1 cm from both enhancing core (ENH) and non-enhancing T2/FLAIR abnormality in pseudorandom fashion, and recorded biopsy locations via screen capture to allow subsequent coregistration with multiparametric MRI

datasets. In this study, a dataset of 95 image-localized biopsies from these 25 GBM patients was collected.

Histologic analysis and tissue treatment: Tissue specimens (target volume of 125mg) were flash frozen in liquid nitrogen within 1-2 min from collection in the operating suite and stored in -80oC freezer until subsequent processing. Tissue was retrieved from the freezer and embedded frozen in optimal cutting temperature (OCT) compound. Tissue was cut at 4 um sections in a -20 degree C cryostat (Microm-HM-550) utilizing microtome blade [23][24]. Tissue sections were stained with hematoxylin and eosin (H&E) for neuropathology review to ensure adequate tumor content (≥50%).

Genetic Profiling and Analysis: We performed DNA isolation and determined copy number variant (CNV) for all tissue samples using array comparative genomic hybridization (aCGH) and exome sequencing as previously published [21][25][26][27]. This included application of previously described CNV detection to whole genome long insert sequencing data and exome sequencing [21][25][26][27].

Copy Number Variant Aberrations for EGFR: TCGA has previously identified a set of highly recurrent and biologically significant DNA gains and losses through copy number analysis, which comprise known therapeutic targets and/or core GBM pathways [28][29]. For this study, we focused on amplification of the receptor tyrosine kinase epidermal growth factor receptor (EGFR), given the high number of clinically tested and available drug inhibitors against this target [30].

MRI protocol, parametric maps, and image coregistration: **Conventional MRI and general acquisition conditions**: We performed all imaging at 3 Tesla field strength

(Sigma HDx; GE-Healthcare Waukesha Milwaukee; Ingenia, Philips Healthcare, Best, Netherlands; Magnetome Skyra; Siemens Healthcare, Erlangen Germany) within 1 day prior to stereotactic surgery. Conventional MRI included standard pre- and post-contrast T1-Weighted (T1-C, T1+C, respectively) and pre-contrast T2-Weighted (T2W) sequences. T1W images were acquired using spoiled gradient recalled-echo inversion-recovery prepped (SPGR-IR prepped) (TI/TR/TE=300/6.8/2.8ms; matrix=320×224; FOV=26cm; thickness=2mm). T2W images were acquired using fast-spin-echo (FSE) (TR/TE=5133/78ms; matrix=320x192; FOV=26cm; thickness=2mm). T1+C images were acquired after completion of Dynamic Susceptibility-weighted Contrast-enhanced (DSC) Perfusion MRI (pMRI) following total Gd-DTPA (gadobenate dimeglumine) dosage of 0.15 mmol/kg as described below [21][22][31]. **Diffusion Tensor (DTI)**: DTI imaging was performed using Spin-Echo Echo-planar imaging (EPI) [TR/TE 10000/85.2ms, matrix 256x256; FOV 30cm, 3mm slice, 30 directions, ASSET, B=0,1000]. The original DTI image DICOM files were converted to a FSL recognized NIfTI file format, using MRIConvert (http://lcni.uoregon.edu/downloads/mriconvert), before processing in FSL from semi-automated script. DTI parametric maps were calculated using FSL (http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/), to generate whole-brain maps of mean diffusivity (MD) and fractional anisotrophy (FA) based on previously published methods [32]. **DSC-pMRI:** Prior to DSC acquisition, preload dose (PLD) of 0.1 mmol/kg was administered to minimize T1W leakage errors. After PLD, we employed Gradient-echo (GE) EPI [TR/TE/flip angle=1500ms/20ms/60o, matrix 128x128, thickness 5mm] for 3 minutes. At 45 sec after the start of the DSC sequence, we administered another 0.05 mmol/kg i.v. bolus Gd-DTPA [21][22][31]. The initial source volume of images from the GE-EPI scan

contained negative contrast enhancement (i.e., susceptibility effects from the PLD administration) and provided the MRI contrast labeled EPI+C. At approximately 6 minutes after the time of contrast injection, the T2*W signal loss on EPI+C provides information about tissue cell density from contrast distribution within the extravascular, extracellular space [22][33]. We performed leakage correction and calculated relative cerebral blood (rCBV) based on the entire DSC acquisition using IB Neuro (Imaging Biometrics, LLC) as referenced [34][35]. We also normalized rCBV values to contralateral normal appearing white matter as previously described [21][22][31]. **Image coregistration:** For image coregistration, we employed tools from ITK (www.itk.org) and IB Suite (Imaging Biometrics, LLC) as previously described [21][22][31]. All datasets were coregistered to the relatively high quality DTI B0 anatomical image volume. This offered the additional advantage of minimizing potential distortion errors (from data resampling) that could preferentially impact the mathematically sensitive DTI metrics. Ultimately, the coregistered data exhibited in plane voxel resolution of ~1.17 mm (256x256 matrix) and slice thickness of 3mm.

ROI segmentation, Image feature extraction and Texture Analysis Pipeline: We generated regions of interest (ROIs) measuring 8×8×1 voxels (9.6×9.6×3mm) for each corresponding biopsy location. A board-certified neuroradiologist (L.S.H.) visually inspected all ROIs to ensure accuracy [21][22]. From each ROI, we employed our in-house texture analysis pipeline to extract a total of 336 texture features from each sliding window. This pipeline, based on previous iterations [21][22], included measurements of first-order statistics from raw image signals (18 features): mean (M) and standard deviation (SD) of gray-level intensities, Energy, Total Energy, Entropy, Minimum, 10th percentile, 90th

percentile, Maximum, Median, Interquartile Range, Range, Mean Absolute Deviation (MAD), Robust Mean Absolute Deviation (rMAD), Root Mean Squared (RMS), Skewness, Kurtosis, Uniformity [36]. We mapped intensity values within each window onto the range of 0–255. This step helped standardize intensities and reduced effects of intensity nonuniformity on features extracted during subsequent texture analysis. Texture analysis consisted of two separate but complementary texture algorithms: gray level co-occurrence matrix (GLCM) [37][38], and Gabor Filters (GF) [39], based on previous work showing high relevance to regional molecular and histologic characteristics [21][22]. The output from the pipeline comprised a feature vector from each sliding window, composed of 56 features across each of the 6 MRI contrasts, for a total of 336 (6*56) features.

EGFR CNV data transformation: As shown in prior work, the log-scale CNV data for EGFR status can also exhibit heavily skewed distributions across a population of biopsy samples, which can manifest as a long tail with extremely large values (up to 22-fold log scale increase) in a relative minority of EGFR amplified samples [40]. Such skewed distributions can present challenges for model training. We adopt a data transformation driven by domain knowledge. It is a reasonable belief that CNV>3.5 corresponds to EGFR amplification [41], which is an abnormal genetic status that is highly valuable to be detected [42][43]. Driven by this, we transformed the original CNV data that maintained identical biopsy sample ordering between transformed and original scales, but condensed the spacing between samples with extreme values on the transformed scale, such that the distribution width of samples with CNV>3.5 approximated that of the samples with CNV< 3.5.

Table 2.1 – Differences in Image features and model complexity when comparing GP and SGP-UF

	SGP-UF	**GP**
Selected image texture features	1.T2.Information.Measure.of.Correlation.2_Avg_1 2.T2.Angular.Second.Moment_Avg_1 3.T2.Kurtosis 4.rCBV.Contrast_Avg_1	1.T2.Difference.Entropy_Avg_3 2.T2.Contrast_Avg_1 3.T2.Entropy_Avg_1 4.SPGRC.Sum.Variance_Avg_3 5.SPGRC.Gabor_Std_0.4_0.1 6.rCBV.Angular.Second.Moment_Avg_1 7.rCBV.Difference.Variance_Avg_1 8.rCBV.Kurtosis 9.EPI.Gabor_Mean_0.4_0.1 10.EPI.Angular.Second.Moment_Avg_1 11.FA.Angular.Second.Moment_Avg_1 12.FA.Skewness 13.FA.Difference.Variance_Avg_3 14.FA.Entropy_Avg_1 15.FA.Sum.Variance_Avg_1 16.FA.Information.Measure.of.Correlation.1_Avg_1 17.FA.Range"

2.4.2 Application of GP and SGP-UF

We applied both GP and SGP-UF to the dataset, aiming at training a model that predicts regional EGFR using MRI features and then using the model to predict unbiopsied regions of each tumor to produce a predicted EGFR map to facilitate clinical decision making. Since GP does not have an inherent mechanism for feature selection like SGP-UP, we used a commonly adopted feature selection algorithm called “Boruta” [44] for GP. Boruta is a powerful feature selection algorithm that uses random forest to warrant the selection of a robust feature subset. It is a wrapper algorithm so it can be integrated with any supervised learning model. We also tried other wrapper algorithms to integrate with GP, but Boruta turned out to have the best performance.

Once the GP and SGP-UF models are trained, the models are used to generate predictions for all the unbiopsied regions of each tumor. The predictions are in forms of predictive distributions. Because our special interest here is to detect CNV>3.5 (i.e., EGFR amplification) versus CNV< 3.5, the predictive distribution of each unbiopsied region is used to test the hypothesis of CNV>3.5 versus CNV< 3.5. Using a cutoff on the p value of this test, e.g., 0.05, the prediction can be converted to a binary classification result (EGFR amplification versus non-amplification/wildtype).

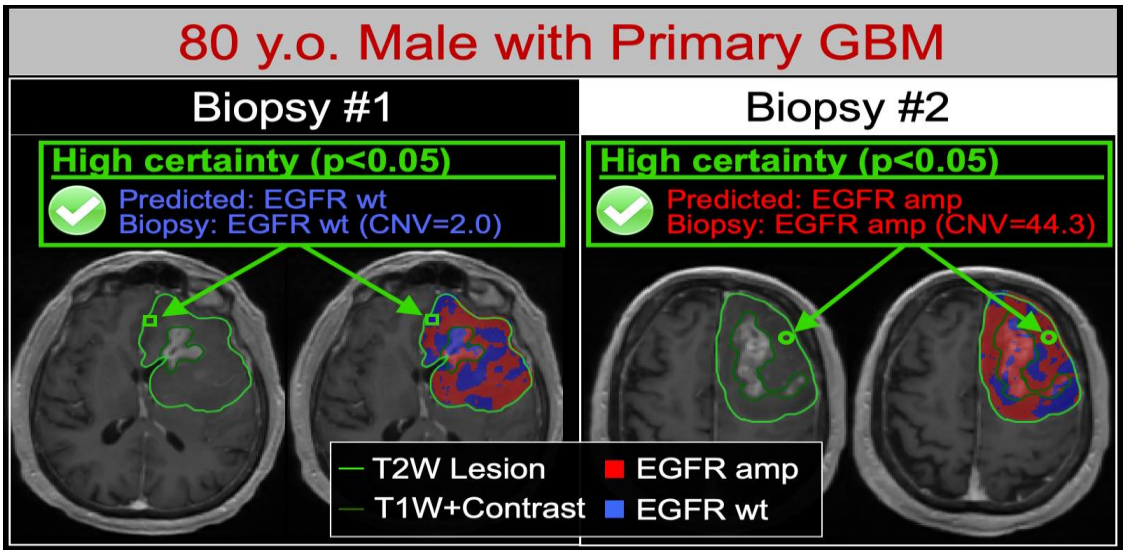

Figure 2.1 – Predicted EGFR amplification vs. wildtype maps using SGP-UF resolve the regional intratumoral heterogeneity of EGFR amplification status in GBM

SGP-UF achieved an overall LopoCV accuracy of 75% (77% sensitivity, 74% specificity) across the entire cohort (n=95). Figure 2.1 illustrates how the spatially resolved predictive maps correspond with stereotactic biopsies from the regionally distinct genetic subpopulations that can co-exist within a single patient's GBM tumor. According to Figure 2.1, it shows two different image-localized biopsies (Biopsy #1, Biopsy #2) from the same

GBM tumor in a single patient. For each biopsy, T1+C images (left) demonstrate the enhancing tumor segment (dark green outline, T1W+Contrast) and the peripheral non-enhancing tumor segment (light green outline, T2W lesion). Color maps for each biopsy (right) also show regions of predicted EGFR amplification (amp, red) and EGFR wildtype (wt, blue) status overlaid on the T1+C images. For biopsy #1 (green square), the map correctly predicted low EGFR copy number variant (CNV) and wildtype status with high predictive certainty ($p<0.05$). Conversely for biopsy #2 (green circle), the maps correctly predicted high EGFR CNV and amplification status, also with high predictive certainty ($p<0.05$). Note that both biopsies originated from the non-enhancing tumor segment, suggesting the feasibility for quantifying EGFR drug target status for residual subpopulations that are typically left unresected followed gross total resection.

We observed substantially different sizes in feature sets and overall model complexity when comparing the GP and SGP-UF models, as summarized in Table 2.1. While the GP model selected 17 image features (across 5 different MRI contrasts), the SGP-UF model selected only 4 features (across 2 MRI contrasts). The lower complexity of the SGP-UF model likely stemmed from a key difference in model training: the SGP-UF model first prioritized feature selection that minimized average predictive uncertainty (i.e., lowest sum of p-values), which helped to narrow candidate features to a relevant and focused subset. Only then did the model prioritize predictive accuracy, within this smaller feature subset. Meanwhile, the GP model selected from the entire original feature set to maximize predictive accuracy, without constraints on predictive uncertainty. Although the accuracy was optimized on training data, the GP model could not achieve the same level

of cross validated model performance (60% accuracy, 31% sensitivity, 71% specificity) compared to the S GP model, due largely to lack of control of extrapolation risks.

Existing published studies have used predictive accuracy to report model performance, but have not yet addressed model uncertainty. Our data suggest that leveraging both accuracy and uncertainty can further optimize model performance and applicability. When stratifying SGP sample predictions based on predictive uncertainty, we observed a striking difference in model performance. The subgroup of sample predictions with the lowest uncertainty (i.e., the most certain predictions) ($p<0.05$) (n=72) achieved the highest predictive performance (83% accuracy, 83% sensitivity, 83% specificity) compared to the entire cohort as a whole (75% accuracy, n=95). This could be explained by the substantially lower performance (48% accuracy, 63% sensitivity, 40% specificity) observed amongst the subgroup of highly uncertain sample predictions ($p>0.05$) (n=23). These discrepancies in model performance persisted even when stratifying with less stringent uncertainty thresholds (e.g., $p<0.10$, $p<0.15$), which we summarize in Table 2.2. Together, these results suggest that predictive uncertainty can inform the likelihood of achieving an accurate sample prediction, which can help discriminate model outputs, not only across patients, but at the regional level within the same patient tumor (Figure 2.2). We obtained two separate biopsies (#1 and #2) from the same tumor in a 44 year-old male patient with primary GBM. The (A) T2W and (B) T1+C images demonstrate the enhancing (dark green outline, T1W+Contrast) and peripheral non-enhancing tumor segments (light green outline, T2W lesion). The (C) color map shows regions of predicted EGFR amplification (amp, red) and EGFR wildtype (wt, blue) status overlaid on the T1+C images. For biopsy #1 (green circle), the SGP-UF model predicted

EGFR wildtype status (blue) with high certainty, which was concordant with biopsy results (green box). For biopsy #2 (yellow circle), the SGP-UF model showed poor certainty (i.e., high uncertainty), with resulting discordance between predicted (red) and actual EGFR status (yellow box).

We collected separate dataset of 24 image-localized biopsies from a cohort of 7 primary high-grade glioma patients. The same MRI techniques and processing pipeline as the training set were adopted. We applied the transductive learning GP model developed on the training set to validate model performance. The transductive learning model achieved an overall accuracy of 67% across the entire pooled cohort (n=24), without stratifying based on the predictive certainty. When stratifying transductive learning GP sample predictions based on predictive uncertainty, we observed an increase in model performance, which was the same trend observed in the training cohort. Specifically, the subgroup of sample predictions with the lowest uncertainty (i.e., the most certain predictions) ($p<0.05$) (n=18) achieved the highest predictive performance (78% accuracy, 75% sensitivity, 79% specificity) compared to samples with high uncertainty ($p>0.05$) (33% accuracy, n=6) and the entire cohort as a whole (67% accuracy, n=24) (Table 2.3).

2.5 Conclusion and Discussion

In this study, we highlight the challenges of predictive uncertainty in radiogenomics and present a novel approach, SGP-UF, that not only quantifies model uncertainty, but also leverages it to enhance model performance and interpretability. This work offers a pathway to clinically integrating reliable radiogenomics predictions as part of decision support within the paradigm of individualized oncology.

Table 2.2 – Differences in predictive accuracy related to certain versus uncertain sample predictions by SGP-UF

Uncertainty threshold	Number of samples (n)	Overall accuracy	Sensitivity/Specificity (%/%)
Entire pooled cohort	95	75%	77/74
p<0.05 (certain)	72	83%	83/83
p>0.05 (uncertain)	23	48%	63/40
p<0.10 (certain)	78	79%	83/78
p>0.10 (uncertain)	17	53%	63/44
p<0.15 (certain)	81	79%	84/77
p>0.15 (uncertain)	14	50%	57/43

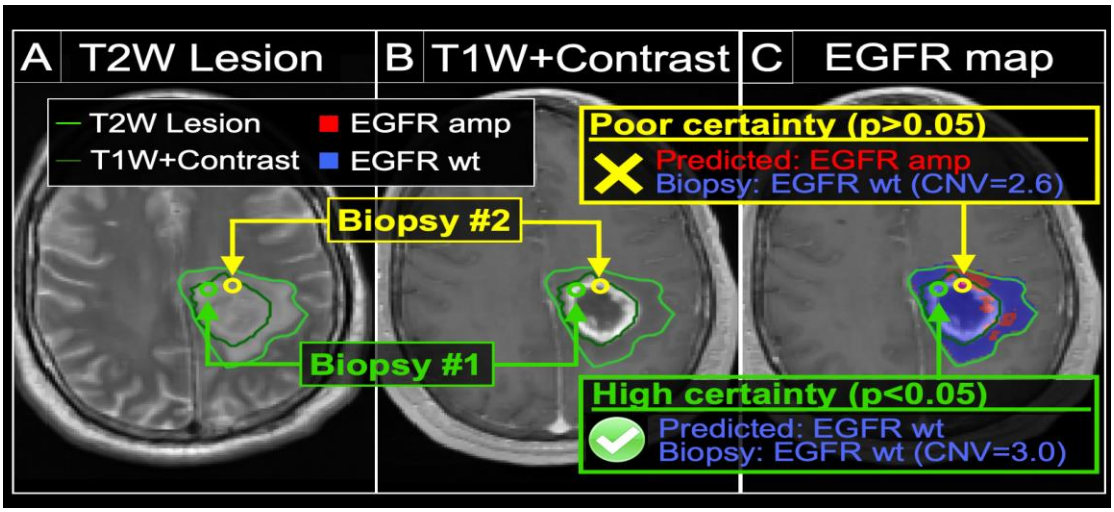

Figure 2.2 – Model uncertainty informing the likelihood of achieving an accurate prediction for EGFR amplification status by SGP-UF

Table 2.3 – Differences in predictive accuracy related to certain versus uncertain sample predictions in the validation set

Uncertainty threshold	Number of samples (n)	Overall accuracy	Sensitivity/Specificity (%/%)
Entire pooled cohort	24	67%	43/76
$p<0.05$ (certain)	18	78%	75/79
$p>0.05$ (uncertain)	6	33%	0/67

CHAPTER 3. KNOWLEDGE-INFUSED GLOBAL-LOCAL DATA FUSION FOR SPATIAL PREDICTION OF TUMOR CELL DENSITY USING MRI

This chapter is based on my published paper "Wang, L., Hawkins-Daarud, A., Swanson, K.R., Hu, L.S., Li, J. (2021) Knowledge-infused Global-Local Data Fusion for Spatial Predictive Modeling in Precision Medicine. IEEE Transactions on Automation Science and Engineering, 19 (3): 2203-2215".

3.1 Background

In many science and engineering domains, the automated capability for generating a spatial prediction map of a variable of interest is critical for decision making. Here we give three examples:

- In Precision Medicine of cancer, one leading cause of treatment failure is intra-tumor heterogeneity [45] [22]. This means that molecular markers, which are typically used to guide treatment decisions, do not uniformly distribute across a tumor. Existing treatments do not adapt well to this regional heterogeneity, leading to sub-optimal treatment outcomes. If the spatial molecular distribution could be precisely mapped out for each tumor, cancer treatments could be greatly improved.
- In forest fire management, the ability for predicting regional fire risk across the forest is important for early detection and prevention [46].
- In poverty management and reduction, one important first step is to map out regional poverty status across a developing world. This information can help

optimally allocate resources [47].

The challenge is that direct measurement for the variable of interest at every spatial location is impossible due to feasibility and cost constraints. Related to the above examples, direct measurement for molecular markers must be done through biopsy. Due to its invasive nature, only a few biopsy samples from a patient can be obtained. Similarly, direct measurement for fire risk must be done through aerial or ground survey, which can only sample a few locations of the forest. For the same reason, survey data that directly reflects poverty levels may only be available for some regions across a developing world. As a result of these constraints, many spatial locations do not have direct measurement data for the variable of interest, i.e., these locations are "blank". This creates a tremendous difficulty for decision making.

On the other hand, indirect or proxy measurement data may be available global-wide. One typical form of such data is imagery. In medicine, clinical imaging such as CT and MRI has been widely used to support diagnosis and treatment. Imaging can be taken non-invasively and portrays the entire host organ of the tumor. Also, imaging of different kinds is designed to measure microscopic tissue structure, morphology, microvasculature, and metabolism, which provide insight into the phenotypic presentation of the molecular characteristics of the tumor. In the other two examples, global proxy data is provided by satellite imagery: spectroradiometer satellite images can help detect fire risk across a forest; regional poverty levels can be reflected in satellite nightlight images portraying power density and daytime images portraying infrastructure, housing, etc.

In addition to sparsely-sampled local data and global imagery, another important source of information is domain knowledge. For example, in cancer biology, mechanistic

models exist for some molecular markers based on biological knowledge and principles [48][49]. These models take the form of algebraic equations, PDEs, or ODEs, and can produce a prediction map for the spatial distribution of some molecular markers across a tumor. However, these models are typically based on simplified assumptions. As a result, the prediction map may only capture some general trend of the molecular distribution but lacks localized precision. In forest fire management, similar forms of domain knowledge exist from forest fire simulators and bio-ecological models [50]. Furthermore, domain knowledge may exist in a looser form. For example, it may be known that some molecular characteristics are more likely to be present at certain regions of a tumor. In the poverty example, there may be historical knowledge that certain regions are less or more wealthy than others.

In summary, with the final objective of generating a spatial prediction map for a variable of interest, there are three sources of pertinent data and information. Please see Table 3.1 for what these data/information sources are in different science and engineering applications. Using a single data/information source by itself does not lead to an optimal solution. This paper proposes a novel computational machine learning framework to optimally fuse the multiple sources of data/information, which is called the methodology of knowledge-infused global-local data fusion (KGL). Please see Figure 3.1 for a schematic overview of the KGL framework. The key idea of KGL is to build a predictive model that uses global imagery to predict the regional distribution for the variable of interest, where the model parameters are optimized to simultaneously serve three purposes: 1) maximizing the accuracy on labeled samples (i.e., regions with direct measurement); 2) reducing the prediction uncertainty on unlabeled samples (i.e., regions without direct measurement but

only imagery); 3) being consistent with the trend or patterns conveyed by domain knowledge.

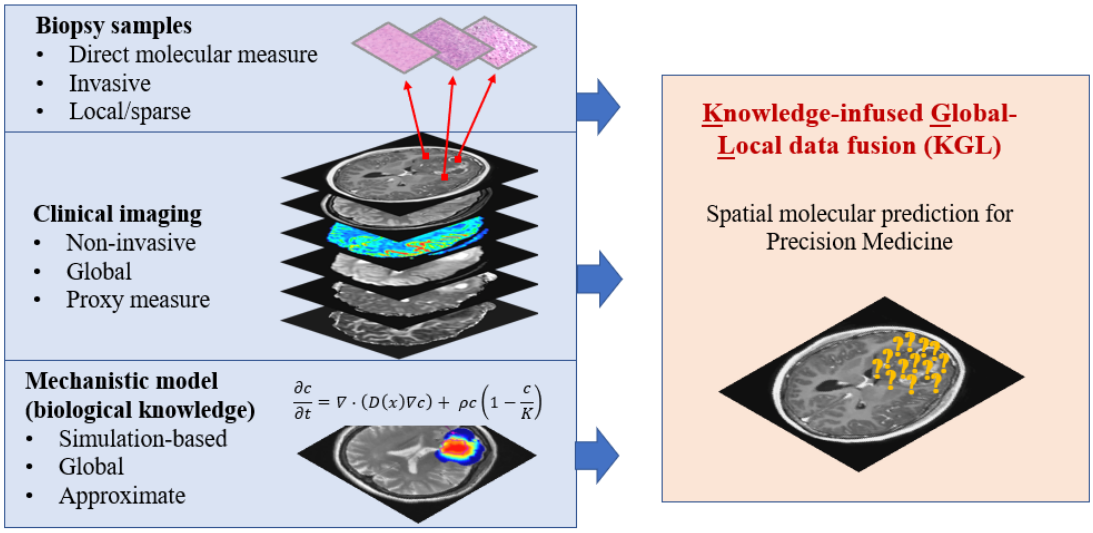

Figure 3.1 – A schematic overview of the multi-data/information fusion framework by the proposed KGL methodology

The contributions of this paper are summarized as follows:

- **New fusion framework**: To our best knowledge, KGL is the first methodology that optimally fuses local and global data together with domain knowledge. There is no existing machine learning framework that immediately targets to achieve this goal.
- **Novel machine learning development**: KGL primarily intersects with two sub-fields in statistical modeling and machine learning: semi-supervised learning (SSL) and Gaussian Process (GP) model. The intersection with SSL is that KGL uses both labeled and unlabeled samples to train the predictive model. Leveraging unlabeled samples to alleviate the sample size limitation is the core idea of SSL. The intersection with GP is that KGL uses a GP to relate regional image features with

the regional variable of interest. While in theory this relationship may be built by some other models, GP is chosen due to its advantages of being non-parametric, non-linear, and most importantly the capacity for generating a predictive probability distribution instead of just a point estimator. This allows for uncertainty quantification and reduction. However, as shown in the next section of Related Works, the existing models in SSL and GP do not provide the capability of KGL.

- **Theoretical insight:** We demonstrate that the formulation of KGL belongs to the machine learning paradigm called Posterior Regularization (PostReg) [47] [46] . PostReg was motivated by the need of integrating domain knowledge with data-driven machine learning algorithms. In probabilistic models, a typical way to incorporate domain knowledge is via Bayesian inference, in which the knowledge is imposed through specification of the prior. However, in many applications such as the examples mentioned in Table 3.1 of this paper, it is difficult to encode the knowledge in a Bayesian Prior. PostReg provides a flexible mechanism to incorporate the knowledge by constraining the posterior distribution. Although PostReg has been existing as a theoretical framework, our paper is the first effort that demonstrates its practical utility in integrating local data, global data, and domain knowledge for spatial prediction.

Contribution to Precision Medicine of cancer treatment: We apply KGL to a real-data application for predicting the spatial distribution of an important molecular marker called tumor cell density (TCD) for each patient with glioblastoma (GBM) – the most aggressive type of brain cancer. KGL generates predictions with higher accuracy and lower uncertainty than a variety of competing methods. The results have important

implication for improving the spatial treatment precision of each GBM tumor.

Table 3.1 – Examples in science and engineering domains that demand the proposed KGL methodology to support critical decision making

	Variable of interest	**Available sources of data/information**		
		Local data (direct measure)	Global data (proxy)	Domain knowledge
Precision Medicine of cancer	Regional molecular status	Biopsy samples	Clinical imaging	Mechanistic models
Early detection of forest fire	Regional fire potential	Ground or aerial survey	Spectro-radiometer satellite images	Forest fire simulators; ecological model
Resource allocation for poverty reduction	Regional poverty level	Household survey	Daytime and nightlight satellite images	Macro-level statistics (country-level GDP)

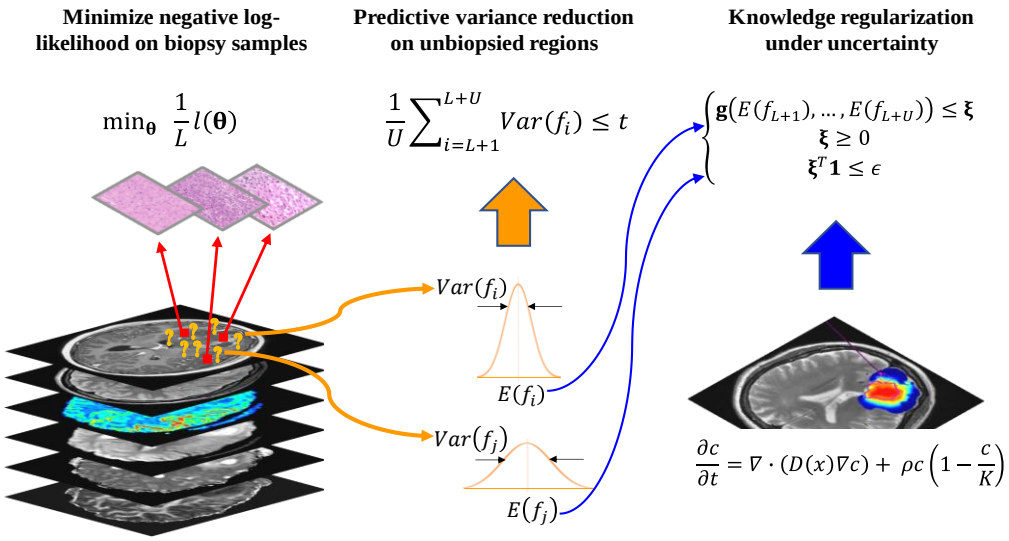

Figure 3.2 – Mathematical formulation of KGL as a constrained optimization

3.2 Knowledge-infused Global-Local Data Fusion (KGL) Model

3.2.1 Mathematical Formulation

Adopt the notation in the Preliminaries section 2.2 and let $\{\mathbf{x}_i, y_i\}_{i=1}^{L}$ be L labeled samples. Let $\{\mathbf{x}_i\}_{i=L+1}^{L+U}$ be U unlabeled samples, e.g. image features extracted from U locations of an area of interest (e.g., a tumor, a forest, a developing world). $y \in \mathbb{R}$ is the measurement of a variable of interest (e.g., a molecular marker, fire risk, poverty level). Our objective is to build a model using $\{\mathbf{x}_i, y_i\}_{i=1}^{L}$ and $\{\mathbf{x}_i\}_{i=L+1}^{L+U}$ together with domain knowledge in order to predict $\{\hat{y}_i\}_{i=L+1}^{L+U}$.

Recall that the advantage of a GP model is that it can produce a predictive distribution, in which the predictive variance σ^{*2} reflects the certainty/uncertainty of the prediction. Also note that σ^{*2} can be computed using only the image features of an unlabeled sample. This leads us to an SSL extension of the GP:

$$\min_{\boldsymbol{\theta}} \ \frac{1}{L} l(\boldsymbol{\theta}) \tag{3.1}$$

$$\text{s.t.} \quad \frac{1}{U}\sum_{i=L+1}^{L+U} Var(f_i) \leq t, \tag{3.2}$$

which minimizes the average negative marginal likelihood under a constraint that upper-bounds the sum of predictive variances on unlabeled samples. Compared with the supervised learning model in (2.1), the SSL considers uncertainty reduction in predicting the unlabeled samples, not just maximizing the likelihood of labeled samples.

Furthermore, considering that domain knowledge may exist, we add additional constraints to (3.2) on the predictive means of unlabeled samples, i.e., (3.3) - (3.7) below:

$$\min_{\boldsymbol{\theta}} \ \frac{1}{L} l(\boldsymbol{\theta}) \tag{3.3}$$

$$\text{s.t.} \quad \frac{1}{U}\sum_{i=L+1}^{L+U} Var(f_i) \leq t, \tag{3.4}$$

$$\mathbf{g}\big(E(f_{L+1}), \dots, E(f_{L+U})\big) \leq \boldsymbol{\xi} \tag{3.5}$$

$$\boldsymbol{\xi} \geq 0 \tag{3.6}$$

$$\boldsymbol{\xi}^T \mathbf{1} \leq \epsilon \tag{3.7}$$

where $\mathbf{1}$ is a vector of m ones. $\mathbf{g}(\cdot)$ contains m different functions, $g_1(\cdot), \dots, g_m(\cdot)$. Each $g_j(\cdot)$ is a function of the predictive means of unlabeled samples, $j = 1, \dots, m$. $\boldsymbol{\xi} = (\xi_1, \dots, \xi_m)$ contains the upper bounds of these functions. A special case is when $m = 1$. Then, (3.5) reduces to a single function of $g\big(E(f_{L+1}), \dots, E(f_{L+U})\big) \leq \xi$. Sometimes, a single function is not enough to represent different kinds of domain knowledge. Thus, we use a general notation in (3.5) to allow for m functions of different forms. Also note that when the domain knowledge is in the form of an equation but not an inequality, i.e., $g\big(E(f_{L+1}), \dots, E(f_{L+U})\big) = \xi$, the equation can always been represented by two inequalities of $g_1\big(E(f_{L+1}), \dots, E(f_{L+U})\big) \leq \xi$ and $-g_2\big(E(f_{L+1}), \dots, E(f_{L+U})\big) \leq -\xi$, which can be added to the constraint set in (3.5). Additionally, we consider that domain knowledge may not always be completely accurate. To accommodate this uncertainty, we use slack variables in specifying the constraints corresponding to domain knowledge, as shown in (3.5) - (3.7). ϵ controls the extent to which the domain k knowledge constraints can be violated. This adds the flexibility of allowing some small violations of these

constraints. To summarize, please see Figure 3.2 for a graphical illustration of the afore-described constrained optimization framework for KGL.

3.2.2 Optimization Algorithm for KGL Model Estimation

To solve the optimization problem in (3.3) - (3.7), we first write the corresponding Lagrangian function, i.e.,

$$\mathcal{L} = \frac{1}{L} l(\boldsymbol{\theta}) + \alpha_1 \left(\frac{1}{U} \sum_{i=L+1}^{L+U} Var(f_i) - t \right) + \sum_{j=1}^{m} \mu_j (g_j(\cdot) - \xi_j) - \sum_{j=1}^{m} v_j \xi_j + \alpha_2 \left(\sum_{j=1}^{m} \xi_j - \epsilon \right), \tag{3.8}$$

with Lagrange multipliers $\boldsymbol{\mu} = (\mu_1, \dots, \mu_m)$, $\boldsymbol{v} = (v_1, \dots, v_m)$, $\alpha_1 \in \mathbb{R}$ and $\alpha_2 \in \mathbb{R}$, and $g_j(\cdot)$ used to represent $g_j\big(E(f_{L+1}), \dots, E(f_{L+U})\big)$ for notation simplicity. Then, the optimal solution of the primal problem in (3.3) - (3.7) is equivalent to the solution of the following optimization:

$$\inf_{\boldsymbol{\theta}, \boldsymbol{\xi}} \sup_{\boldsymbol{\mu} \geq 0, \boldsymbol{v} \geq 0, \alpha_1 \geq 0, \alpha_2 \geq 0} \mathcal{L} . \tag{3.9}$$

Theorem 3.1: Let $\mathcal{L}' = \frac{1}{L} l(\boldsymbol{\theta}) + \lambda_1 \left(\frac{1}{U} \sum_{i=L+1}^{L+U} Var(f_i) \right) + \sum_{j=1}^{m} \mu_j (g_j(\cdot) - \xi_j) - \sum_{j=1}^{m} v_j \xi_j + \lambda_2 \left(\sum_{j=1}^{m} \xi_j \right)$, where λ_1 and λ_2 are tuning parameters. Then, for any $\lambda_1 > 0$ and $\lambda_2 > 0$, there exist $t > 0$ and $\epsilon > 0$ such that the optimal solution of $\inf_{\boldsymbol{\theta}, \boldsymbol{\xi}} \sup_{\boldsymbol{\mu} \geq 0, \boldsymbol{v} \geq 0, \alpha_1 \geq 0, \alpha_2 \geq 0} \mathcal{L}$ is equal to that of $\inf_{\boldsymbol{\theta}, \boldsymbol{\xi}} \sup_{\boldsymbol{\mu} \geq 0, \boldsymbol{v} \geq 0} \mathcal{L}'$ and vice versa. (Proof in Appendix C.)

According to Theorem 3.1, (3.9) can be further simplified as:

$$\inf_{\boldsymbol{\theta}, \boldsymbol{\xi}} \sup_{\boldsymbol{\mu} \geq 0, \boldsymbol{v} \geq 0} \mathcal{L}' . \tag{3.10}$$

Since $\mathcal{L}'$ is a convex function of $\xi_j, \boldsymbol{\mu}, \boldsymbol{v}$ (non-convex of $\boldsymbol{\theta}$), (3.10) is equivalent to

$$\inf_{\boldsymbol{\theta}} \sup_{\boldsymbol{\mu}\geq 0, \boldsymbol{v}\geq 0} \inf_{\boldsymbol{\xi}} \mathcal{L}' . \tag{3.11}$$

Focus on the inner minimization in (3.11). The minimizer of ξ_j must satisfy

$$\frac{\partial \mathcal{L}\prime}{\partial \xi_j} = \lambda_2 - u_j - v_j = 0, j = 1, \dots, m. \tag{3.12}$$

From (3.12), we can write $v_j = \lambda_2 - u_j$. Inserting this into (3.11), we get

$$\inf_{\boldsymbol{\theta}} \sup_{\boldsymbol{\mu}\geq 0} \mathcal{J}\left(u_j; j = 1, \dots., m\right) \tag{3.13}$$

$$\text{s.t. } 0 \leq \mu_j \leq \lambda_2, j = 1, \dots., m. \tag{3.14}$$

where $\mathcal{J}\left(u_j; j = 1, \dots., m\right) = \frac{1}{L} l(\boldsymbol{\theta}) + \lambda_1\left(\sum_{i=L+1}^{L+U} Var(f_i)\right) + \sum_{j=1}^{m} u_j g_j(\cdot)$.

It is clear that the solution of the inner maximization of (3.13) with (3.14) is $\mu_j = \begin{cases} \lambda_2, & if\ g_j(\cdot) > 0 \\ any\ value\ in\ [0, \lambda_2], & if\ g_j(\cdot) = 0 \\ 0, & if\ g_j(\cdot) < 0 \end{cases}$. Then, the final objective function becomes

$$\inf_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \inf_{\boldsymbol{\theta}} \frac{1}{L} l(\boldsymbol{\theta}) + \lambda_1 \left(\frac{1}{U}\sum_{i=L+1}^{L+U} Var(f_i)\right) + \lambda_2 \sum_{j=1}^{m} g_j(\cdot) I(g_j(\cdot) > 0), \tag{3.15}$$

The gradient of objective function in (3.15) can be written as

$$\nabla L_{\boldsymbol{\theta}} = \frac{1}{L} \nabla\, l(\boldsymbol{\theta}) + \lambda_1 \left(\frac{1}{U}\sum_{i=L+1}^{L+U} \nabla Var(f_i)\right) + \lambda_2 \sum_{j=1}^{m} \nabla g_j(\cdot) I(g_j(\cdot) > 0).$$

In this paper, this optimization is solved by a gradient descent algorithm implemented in R.

Discussion on the insight of the optimization: Note that the optimization in (3.15) simultaneously balances three aspects: maximizing the average marginal likelihood on labeled samples (recall that $l(\boldsymbol{\theta})$ is the negative marginal likelihood as defined in (2.1)); minimizing the predictive variances/uncertainty on unlabeled samples; optimizing the consistency with domain knowledge. The last term in (3.15) is particularly interesting: $I(g_j(\cdot) > 0)$ is an indicator function that takes the value of one if $g_j(\cdot) > 0$ and zero otherwise. Recall that in the KGL formulation in (3.3)- (3.7), the consistency with domain knowledge is imposed by having the constraints of $g_j(\cdot) \leq \xi_j, \xi_j \geq 0, j = 1, \ldots, m$, where we consider m different types of domain knowledge. The utility of the indicator functions is to find which subset of these constraints must be satisfied. This is the subset corresponding to $g_j(\cdot) \leq 0$ or equivalently $I\big(g_j(\cdot) > 0\big) = 0$. For the remaining constraints corresponding to $g_j(\cdot) > 0$ or equivalently $I\big(g_j(\cdot) > 0\big) = 1$, the model will try to satisfy these constraints as much as possible, but this needs to be traded off with the first two terms in the optimization, i.e., some degree of violations for these constraints is allowed. The appealing part of the model is that it does not require pre-specifying which subset of constraints must be satisfied and which not, and how much violation is allowed. All these will be automatically resolved through solving the optimization problem.

A final note is that since the optimization problem in (3.15) is non-convex, the converged solution may not be the global optimal. This is a common problem for non-convex optimization problems. A typical strategy is to use different initial values. More sophisticated non-convex optimization algorithms may be used but are left for future investigation.

3.3 Another View: KGL as Posterior Regularization (PostReg)

To incorporate domain knowledge in probabilistic models, a common approach is to specify a prior of the model M that reflects the domain knowledge, i.e., $\pi(M)$. This prior is then integrated with the data likelihood $p(D|M)$ using the Bayes' rule to obtain the posterior $p(M|D)$. In this approach, domain knowledge does not directly impact or regularize the final model estimate, but only indirectly through prior specification. Due to the indirect nature, the final model estimate may not fully comply with the knowledge. In some applications, it may be preferred that domain knowledge can be used to directly regularize the posterior. This has led to the development of the PostReg framework [46]. The basic idea of PostReg is to use a variational distribution $q(M|D)$ to approximate the posterior $p(M|D)$, while at the same time regularizing $q(M|D)$ according to domain knowledge. That is, PostReg aims to find the solution $q^*(M|D)$ for the following optimization

$$\inf_{q\in\mathcal{P}_{prob}} \; KL\big(q(M|D)\|p(M|D)\big) + \Omega\big(q(M|D)\big). \tag{3.16}$$

The first term is the Kullback–Leibler (KL)-divergence, defined as the expected log-difference between the posterior and approximate distributions. $\Omega(\cdot)$ is a function of the approximate distribution, which regularizes this distribution to comply with domain knowledge. Because of the regularization effect, $q(M|D)$ cannot be exactly equal to the posterior $p(M|D)$, but is made close to $p(M|D)$ while at the same time being consistent with the domain knowledge. $\mathcal{P}_{prob}$ denotes a proper variational family of distributions. The PostReg optimization in (3.16) is a general formulation. It has been realized for specific

models such as latent variable models under the EM framework [47], multi-view learning [47], and infinite Support Vector Machines [51].

We demonstrate that solving the optimization in (3.3)- (3.7) is equivalent to solving a specific form of the PostReg optimization. In this specific form, the choice of the regularizer $\Omega\big(q(M|D)\big)$ corresponds to variance minimization and consistency with domain knowledge in expectation. This theoretical result is summarized in Theorem 3.2. (Proof in Appendix D.)

Theorem 3.2: The optimization in (3.3)- (3.7) is equivalent to a PostReg optimization taking the form of, i.e.,

$$\inf_{q\in\mathcal{P}_{prob}} \; KL\big(q(M|D)\|p(M|D)\big) + \Omega\big(q(M|D)\big), \tag{3.17}$$

with the following specific definitions for the notations: $M = (f, \boldsymbol{\theta})$ is the model; $D = \big(\{\mathbf{x}_i, y_i\}_{i=1}^{L}, \{\mathbf{x}_i\}_{i=L+1}^{L+U}\big)$ is the data; $\mathcal{P}_{prob} = \{q \mid q(f, \boldsymbol{\theta}|D) = p(f|\boldsymbol{\theta}, D)\delta_{\overline{\boldsymbol{\theta}}}(\boldsymbol{\theta}|D), \overline{\boldsymbol{\theta}} \in \Theta\}$ is a variational family of distributions where $q(f|\boldsymbol{\theta}, D) = p(f|\boldsymbol{\theta}, D)$ and $q(\boldsymbol{\theta}|\mathcal{D}) = \delta_{\overline{\boldsymbol{\theta}}}(\boldsymbol{\theta}|D)$ which is a Dirac delta function centered on $\overline{\boldsymbol{\theta}}$ in the parameter space Θ; $\Omega\big(q(f, \boldsymbol{\theta}|D)\big)$, denoted by a simple form of $\Omega(q)$ hereafter, is given by

$$\Omega(q) = \inf_{t,\boldsymbol{\xi}} \left\{ (\lambda_1 t + \lambda_2 \textstyle\sum_{j=1}^{m} \xi_j) \left| \begin{array}{c} \frac{1}{U}\sum_{i=L+1}^{L+U}\left(\int_{f,\boldsymbol{\theta}} q \times \big(f(\mathbf{x}_i) - E_q[f(\mathbf{x}_i)]\big)^2 \, d\eta(f, \boldsymbol{\theta})\right) \le t; \\ \mathbf{g}\left(\int_{f,\boldsymbol{\theta}} q \times f(\mathbf{x}_{L+1}) \, d\eta(f, \boldsymbol{\theta}), \ldots, \int_{f,\boldsymbol{\theta}} q \times f(\mathbf{x}_{L+U}) \, d\eta(f, \boldsymbol{\theta})\right) \le \boldsymbol{\xi}; \\ \boldsymbol{\xi} \ge 0 \end{array} \right. \right\}.$$

(3.18)

By demonstrating that KGL is a specific instance within the general PostReg framework, we can gain two insights: First, we obtain another angle to explain how domain knowledge is integrated with global and local data in KGL, i.e., domain knowledge is imposed to regularize the posterior of the model (not the prior nor by any other means). Second, KGL provides a realization of the general PostReg framework and enriches the problem set PostReg can potentially address. Although PostReg has been existing as a theoretical framework, KGL is the first effort that demonstrates practical utility of using the concept of PostReg to integrate local data, global data, and domain knowledge for spatial estimation.

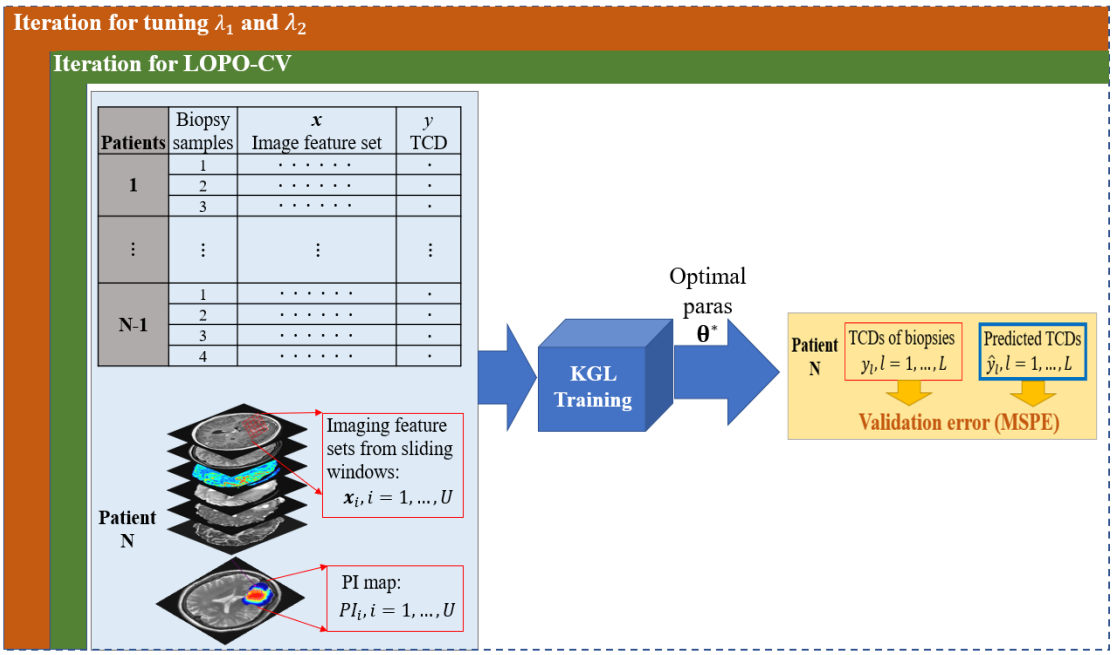

Figure 3.3 – Model training procedure for KGL

3.4 Application

3.4.1 Data Collection and Pre-processing

Glioblastoma (GBM) is the most aggressive type of brain tumor with median survival of 15 months [52]. Intra-tumor molecular heterogeneity has been found to be one of the leading causes of treatment failure. Tumor cell density (TCD) is an important molecular marker to inform surgical intervention and radiation therapy. TCD is the percentage of tumor cells within a spatial unit of the tumor. It is well-known that TCD is spatially heterogeneous, meaning that TCD varies significantly across different sub-regions of each tumor [53] [54]. Mapping out the spatial distribution of TCD across each tumor is important for a neurosurgeon to determine where to resect. The mapping will also help radiation treatment planning by informing a radiation oncologist on how to optimize the spatial dose distribution according to the regional TCD. Such optimal decision is critical to avoid overtreating some areas of the brain – causing functional impairment, and undertreating other areas – leading to tumor recurrence. To know the TCD at each sub-region of a tumor, biopsy is the gold-standard approach. However, due to its invasive nature, only a few biopsy samples can be taken. MRI portrays the entire brain non-invasively. But MRI does not directly measure TCD while only providing proxy data. In this experiment, we apply KGL to predict regional TCD of each tumor by integrating MRI, biopsy samples, and mechanistic model/domain knowledge.

Patients and biopsy samples: This study includes the data of 18 GBM patients provided by our collaborators at Mayo Clinic with IRB approval. Each patient has 2-14 biopsy samples, making a total of 82 samples. Pre-operative MRI including T1-weighted

contrast-enhanced (T1+C) and T2-weighted sequences (T2) was used to guide biopsy selection. The neurosurgeons recorded biopsy locations via screen capture to allow subsequent co-registration with multiparametric MRI. The TCD of each biopsy specimen was assessed by a neuropathologist.

MRI pre-processing and feature extraction: Each patient went through an MRI exam prior to treatment. The MRI exam produced multiple contrast images such as T1+C, T2, dynamic contrast enhancement (EPI+C), mean diffusivity (MD), fractional anisotropy (FA), and relative cerebral blood volume (rCBV). Detailed MRI protocols and image co-registration can be found in our prior publications [55], [56]. To extract features, an 8x8 $pixel^2$ window was placed at each pixel as the center within a pre-segmented tumoral Region of Interest (ROI), which is the abnormality visible on T2. The window was slid throughout the entire ROI, and at each pixel, the average gray-level intensity was computed within the 8x8 $pixel^2$ window from each of the six contrast images and used as features. Therefore, six image features were included in model training.

Labeled and unlabeled samples: Biopsy samples are labeled samples as they have TCD. Samples corresponding to the sliding windows, except the windows at biopsy locations, are unlabeled as they only have image features not TCD.

Mechanistic model: We integrate a well-known mechanistic model called Proliferation-Invasion (PI) [55] [57]. PI is a PDE-based simulator driven by biological knowledge of how GBM tumor cells proliferate and invade to sounding brain tissues. The PDE for the PI model is:

$$\overbrace{\frac{\partial c}{\partial t}}^{\text{Rate of Change of Cell Density}} = \overbrace{\nabla \cdot (D(x)\nabla c)}^{\text{Invasion of Cells into Nearby Tissue}} + \overbrace{\rho c\left(1-\frac{c}{K}\right)}^{\text{Proliferation of cells}},$$

where $c(x,t)$ is the TCD at location x of the brain and time t, $D(x)$ is the net rate of diffusion, ρ is the net rate of proliferation and K is the cell carrying capacity. Solutions to this model are known to asymptotically set up a traveling wave in spherical symmetry. This wave has two key properties 1) the radial wave speed, known to be $2\sqrt{D\rho}$, and 2) the gradient of the wave front, which is known to be related to the ratio D/ρ. By assuming different imaging sequences of T1+C and T2 correlate with different thresholds of density on the traveling wave, one can estimate the D/ρ and generate estimations of the current gradient/shape of the tumor cell density profile [58], [59]. In line with previous papers, the T1+C and T2 images of a patient are used to calibrate the model parameters assuming the abnormality on the T1+C image corresponds to the 80% tumor cell density threshold and the T2 image to the 16% tumor cell density. By estimating D/ρ, we can generate the current TCD estimate at each pixel. The PI map can capture some general trend of the spatial TCD distribution but may lack localized precision due to simplified assumptions and with only D/ρ estimated cannot be used to predict future growth. We run the PI simulator for each patient and generate a PI map to be integrated with KGL for this single time point of interest (see the next section).

3.4.2 Application of KGL

3.4.2.1 Integration of domain knowledge encoded by PI map

In KGL, domain knowledge is incorporated through imposing constraints on the predictive means of unlabeled samples, i.e., $\mathbf{g}\big(E(f_{L+1}),\dots,E(f_{L+U})\big) \le \boldsymbol{\xi}$. Due to the aforementioned properties of the PI map, we propose to use it to regularize the general spatial trend of the TCD predictions. Specifically, based on the pixel-wise estimates of TCD generated by PI, we compute the average estimate over 64 pixels within each 8x8 pixel^2 window corresponding to an unlabeled sample. Denote this average estimate for each unlabeled sample i by $PI_i, i = L+1,\dots,L+U$. The proposed constraints are:

$$\begin{cases} g_1\big(E(f_{L+1}),\dots,E(f_{L+U})\big) \triangleq |E(f_{L+1}) - PI_{L+1}| \le \xi_1 \\ \vdots \\ g_U\big(E(f_{L+1}),\dots,E(f_{L+U})\big) \triangleq |E(f_{L+U}) - PI_{L+U}| \le \xi_U \end{cases}, \tag{3.19}$$

$$g_{U+1}\big(E(f_{L+1}),\dots,E(f_{L+U})\big) \triangleq \sum_{i=L+1,\dots,L+U;\, j>i} w_{ij}\left(E(f_i) - E\big(f_j\big)\right)^2 \le \xi_{U+1}, \tag{3.20}$$

$$\xi_1,\dots,\xi_{U+1} \ge 0, \sum_{i=1}^{U+1} \xi_i \le \epsilon, \tag{3.21}$$

where $w_{ij} = e^{-(PI_i - PI_j)^2}$. The constraints in (3.19) encourage similarity between the predictive mean and the PI estimate at the same location (unbiopsied sample). Additionally, the constraint in (3.20) encourages the predictive means of two samples to be similar if their PI estimates are similar, where the PI similarity is reflected by w_{ij}. Furthermore, considering that the PI map only provides approximates of the TCDs, a slack variable approach is used in (3.21) to make these constraints soft instead of hard constraints.

3.4.2.2 Model training and competing methods

Model training needs to determine the optimal parameter estimates $\boldsymbol{\theta}^*$ of KGL and select the tuning parameters, λ_1 and λ_2. The training procedure is depicted in Figure 3.3. The

search for the optimal turning parameters is used as the outermost iteration. At fixed λ_1 and λ_2, the KGL optimization is solved for each patient. The input to the patient-specific optimization includes labeled samples from other patients, unlabeled samples from this patient, and the PI map of this patient. To improve efficiency and robustness, a subset of the first 100 unlabeled samples with the smallest average distances from the labeled samples is included. The output is optimal parameters, $\boldsymbol{\theta}^*(\lambda_1, \lambda_2)$. Then, the model under the optimal parameters is used to generate a predictive distribution of the TCD for each biopsy sample of this patient. The predictive means of all the biopsy samples are compared with the true TCDs to compute the Mean Absolute Prediction Error (MAPE). This process is iterated with every patient in the dataset treated as "this patient", known as leave-one-patient-out cross validation (LOPO-CV). While other types of CV schemes may be adopted, LOPO-CV aligns well with the natural grouping of samples in our dataset. Finally, the best tuning parameters λ_1^* and λ_2^* are selected as the ones minimizing the average MAPE over all the patients. Under the λ_1^* and λ_2^*, the KGL optimization is solved for each patient to generate the final optimal parameters $\boldsymbol{\theta}^*$for the patient.

For comparison, we applied a range of competing algorithms to the same dataset, including:

1) The mechanistic model, i.e., PI;
2) The standard GP [53], i.e., a GP model trained using only biopsy samples;
3) Semi-GP: A semi-supervised GP model based on a data-dependent covariance function for unlabeled data [60];
4) Co-training SVR-KNN: an SSL algorithm based on co-training with support vector regression (SVR) and k-nearest neighbors (KNN) [61];

5) SSRR-AGLP: semi-supervised ridge regression with adaptive graph-based label [62];
6) SS-RT: semi-supervised regression trees [63];
7) SAFER: SAFE semi-supervised Regression [64];
8) KGL with no variance reduction: this is a special case of KGL without the constraint on predictive variances;
9) KGL with random unlabeled sample selection: this is a special case of KG by randomly selecting 100 unlabeled samples to include in model training.

The two GP models in 2) and 3) were chosen to form the baseline to compare with KGL. 4)-7) are existing SSL algorithms, each representing a major category of SSL: co-training, graph-based, and low-density separation for 4)-6), respectively, and an integrated framework to combine multiple SSL algorithms for 7). These algorithms were developed in recent years. 8) and 9) are two special cases of KGL: 8) intends to show the benefit of bias-variance tradeoff of KGL. 9) adopts an alternative strategy by randomly selecting 100 unlabeled samples to include in training, as opposed to selecting the top 100 unlabeled samples with the smallest average distances from the labeled samples. The parameters of each algorithm were optimized based on the same LOPO-CV criterion as KGL.

B.3 Generation of predicted TCD maps and uncertainty quantification

For the three GP-based methods, the trained model of each method can be used to generate a predictive distribution of the TCD for each sample (i.e., each sliding window) within the ROI. The predictive means of all the samples can be visualized by a color map

overlaid on the ROI. Also, we can use the predictive variances to quantify prediction uncertainty.

3.4.2.3 Results

Table 3.2 compares all methods for MAPE. Only GP-based methods can produce predictive variance, so they are additionally compared in terms of average predictive variance for biopsy samples. The last three KGL methods have the smallest MAPE. Their average predictive variances are also much smaller than the two existing GP-based methods. Among the three KGL methods, the last one performs the best, implying the benefit of including the variance constraint and adopting a more robust unlabeled sample selection strategy.

Figure 3.4 compares standard GP, semi-GP, and KGL in terms of the average predictive variance for all samples (i.e., sliding windows) within the ROI for each patient. KGL has a smaller MAPE. The predictive variances by KGL are much reduced for all samples and across all patients, implying greater certainty in the prediction (Averages across all patients: Standard GP=0.032; Semi-GP=0.032; KGL=0.014 (56% variance reduction compared with the other two methods)).

Furthermore, Figure 3.5 shows the predictive TCD maps from two patients as examples. Colors represent predictive means of the TCD from 0 (darkest blue) to 100% (darkest red). Below each map, we also show the distribution of the predictive variances for samples within the ROI. Patient A has one biopsy sample shown on this slice of the MRI. Both standard GP and semi-GP underestimate the TCD of this sample by a large margin, whereas KGL has a higher accuracy. Patient B has two biopsy samples for which

KGL estimates with higher accuracy. Also, the color maps produced by KGL show better spatial smoothness and aligns better with the expected tumor cell distributions from known biology, especially for the color map of patient B. This is a benefit due to incorporation of the PI map/domain knowledge in model training. Furthermore, the predictive variance distribution by KGL is much more concentrated at the low variance range, whereas standard GP and semi-GP produce predictions with large variances (large uncertainty). In all, KGL outperforms the other two methods in both prediction accuracy, prediction certainty, and compliance to biological knowledge.

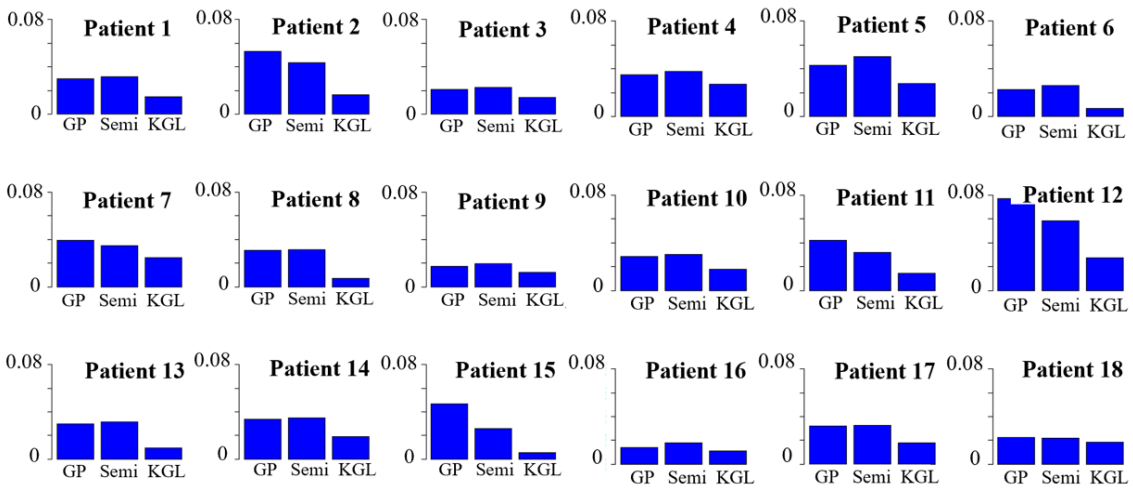

Figure 3.4 – Comparison of methods on average predictive variance of unlabeled samples for each patient

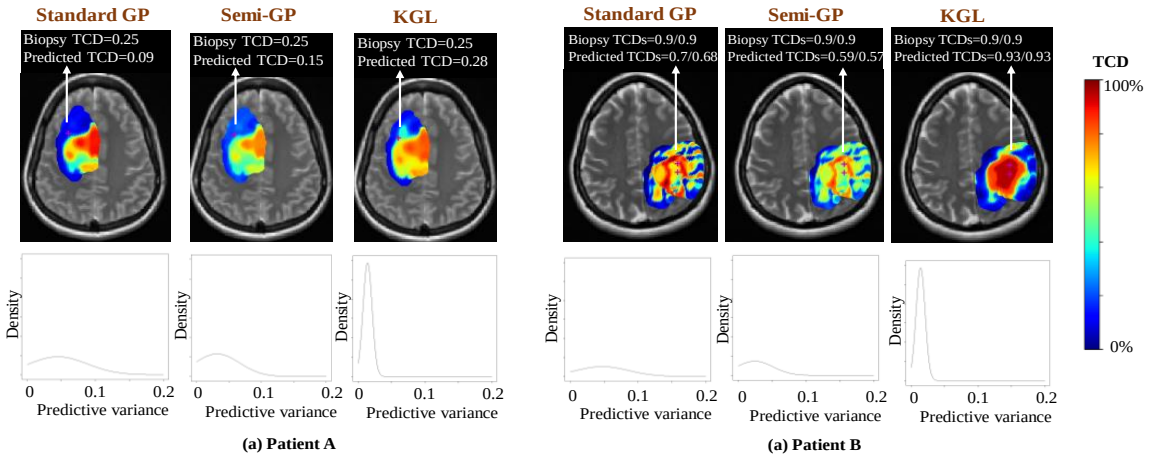

Figure 3.5 – Predicted means of TCD within a ROI shown as a color map overlaid on the patient's T2 MRI; predicted variances shown in distribution

3.4.2.4 Discussion on utilities of the results to decision making in Precision Medicine

With the predicted TCD map for each patient, the neurosurgeon can have a better reference to decide where of the brain to take out more (or less) cancerous tissues. Areas with high TCD should be maximally resected. Areas with little TCD should be preserved so as to protect the integrity of brain functions. This level of spatial precision is highly valuable for optimizing the surgical outcomes of GBM. Furthermore, the predicted TCD maps can also help radiation oncologists decide how to optimize the spatial radiation dose in radiation therapy. Areas with higher TCD should be irradiated more to kill the cancer cells, whereas areas with lower TCD should receive less dose to minimize radiation-induced complications. This level of spatial precision is much desirable for radiation treatment planning optimization. Finally, we like to point out that since KGL also generates a predictive variance in addition to the mean for each sample, the variance can be used to

quantify the uncertainty of the prediction to guide more informed and risk-conscious clinical decision making.

Table 3.2 – Comparison of methods on prediction of biopsy samples

Methods	MAPE	Average predictive variance
PI	0.252	-
Standard GP	0.191	0.038
Semi-GP	0.189	0.039
Co-training SVR-KNN	0.243	-
SSRR-AGLP	0.201	-
SS-RT	0.231	-
SAFER	0.223	-
KGL (no variance reduction)	0.174	0.023
KGL (random unlabeled sample selection)	0.171	0.018
KGL	**0.165**	**0.015**

3.5 Conclusion and Discussion

We proposed a novel machine learning framework, KGL, to optimally fuse multiple sources of data/information to predict the spatial distribution for a variable of interest. KGL was demonstrated in an application of predicting the spatial TCD distribution for GBM, and showed superior performance over competing methods. Future research includes methodological extension to non-numerical response variables, optimal selection of unlabeled samples, and development of more efficient optimization solvers.

CHAPTER 4. WEAKLY SUPERVISED ORDINAL LEARNING FOR INTRA-TUMOR MULTI-GENE PREDICTION USING MRI

This chapter is based on paper "Quantifying Intra-tumoral Genetic Heterogeneity of Glioblastoma toward Precision Medicine using MRI and a Data-inclusive Machine Learning Algorithm", submitted to Computer Methods and Programs in Biomedicine, 2022.

4.1 Background

Glioblastoma (GBM) remains one of the most aggressive and lethal of all human cancers. The current treatment regimen only increases the median overall survival to about 15 months [65]. This lack of substantial survival benefit, despite best available standard therapy, has motivated efforts to identify the underlying factors contributing to poor clinical outcomes.

One of the important factors contributing to GBM complexity is the intra-tumoral genetic heterogeneity [56] [66], which been cited as a clinical challenge for treatment [67]. Each tumor is comprised of genetically distinct subpopulations with different sensitivities to treatment. As a result, genetic targets from one biopsy location may not accurately reflect those from other parts of the same tumor [66]. Worse yet, given the invasive nature of the disease, diffusely invaded GBM cells are always left behind in the brain even after resection, and these remaining regions may be genetically distinct to the biopsy samples collected during surgery [68]–[70]. The region-to-region genetic variability provides potential mechanisms for therapeutic escape and makes single targeted therapies less

effective [71]. For instance, EGFR represents the one of the most common gene driver alterations in GBM and has been implicated in several pathogenic mechanisms. Targeted drug therapies directed at EGFR as well as other receptor tyrosine kinases (RTKs) such as PDGFRA have been developed [72] [73]. However, co-expression of other common genetic driver alterations such as PTEN loss/deletion/mutation have been implicated as a mechanism of resistance against RTK inhibitors by upregulation of the PI3K signaling pathway [74] [75]. This resistance mechanism with enhanced PI3K signaling could explain the poor clinical outcomes despite EGFR targeted drug therapies in previous clinical trials [75] [76]. This underscores the importance of identifying how the key genetic alteration combinations vary from region to region across each individual tumor to inform future treatment regimens.

There are substantial challenges for quantifying intra-tumoral genetic heterogeneity of GBM. Ideally, one would need to have biopsy samples taken from many different regions of a tumor and perform genetic analysis of each sample. This, however, is infeasible due to the invasive nature of biopsy. Although the central tumor mass can often be surgically removed, the invasive portions of the tumor are often left unresected and unbiopsied given the risk to adjacent neurologic structures. Thus, biopsy alone is insufficient to characterize the full landscape of the intra-tumoral heterogeneity [10][14].

Neuroimaging such as MRI provides a non-invasive means to evaluate the whole tumor (indeed the entire brain), which may be used to help quantify intra-tumoral genetic heterogeneity. Genetic variations of GBM result in alterations in biological characteristics of the tumor that may include changes in apoptosis, cellular proliferation, cellular invasion, and angiogenesis [77]. These biological changes, in turn, manifest physiological changes

that are detectable by a combination of MRI sequences [66]. Thus, developing a machine learning (ML) model that utilizes MRI features to predict genetic characteristics could provide a non-invasive means to quantify intra-tumoral genetic heterogeneity [78] [66] [79].

The emerging field of radiogenomics has shown the feasibility of using MRI features to predict genetic status in individualized oncology. Previous studies have investigated the use MRI features to predict the EGFR status of GBM patients. Akbari *et al.* [80] extracted multiparametric MRI features from tumor regions and trained a Support Vector Machine (SVM) to predict EGFRvIII Mutation. Tykocinski *et al.* [81] used multivariable logistic regression to predict EGFRvIII mutation based on features extracted from perfusion-weighted MRI. Kickingereder et al. [82] used common classifiers such as stochastic gradient boosting machine, random forest, and logistic regression to predict the copy number variant (CNV) status of several GBM driver genes such as EGFR, PDGFRA, and PTEN based on multiparametric MRI. Chen et al. [83] built a convolutional neural network to predict PTEN mutation using multiparametric MRI. There are also studies to classify GBM tumors into subtypes in terms of IDH mutation status [84] [85] or MGMT methylation status [86] as these subtypes have been reported to have different prognoses [87].

However, most of these published studies are non-localized, which extracted MRI features from large tumoral regions to predict the overall genetic status of the tumor. They remain incapable of resolving the intra-tumoral/regional genetic heterogeneity of a GBM tumor. The study by Hu et al. [78] is among the first ones to train ML models to predict

regional CNV status of several driver genes within each tumor based on localized multiparametric MRI features. In a more recent work, Hu et al. [79] developed a transductive learning algorithm to reduce the uncertainty of ML in predicting regional CNV of EGFR. Two recent review papers pointed out the importance of having more studies like these to resolve intra-tumoral genetic heterogeneity [23] [24].

The ML algorithms used in these existing studies are mainly supervised learning models, in which the model training uses biopsy samples (a.k.a. labeled samples) only. The limitation is that only a small number of biopsy samples can be acquired from each patient. On the other hand, since MRI images are readily available for the whole tumoral area (indeed the entire brain), there is an opportunity to include MRI features outside the biopsied regions of the tumor as unlabeled samples. Integrating unlabeled and labeled samples is known as semi-supervised learning (SSL). In our application, including unlabeled samples is not only possible but also necessary. This is because our final goal is to generate predictions for the large amount of unbiopsied regions, which will help understand the full landscape of regional genetic heterogeneity to drive treatment decision for each patient. If the predictive model was trained using only biopsy samples, the generalization performance of the model could be unsatisfactory. SSL is known to improve generalization performance [90]. Existing SSL algorithms fall into several categories such as generative methods [91], low-density separation [92], disagreement-based [93] and graph-based algorithms [94]. If directly using these algorithms, one could include biopsy/labeled samples and unlabeled samples from the parts of the tumor not being biopsied. However, this strategy is still not optimal as it overlooks the MRI features outside the tumoral area, i.e., the normal brain. Training an ML model to differentiate the normal

brain from tumor may help the model further learn how to differentiate the regions with and without a particular genetic alteration within the tumor. This motivates us to develop a data-inclusive ML algorithm called Weakly-Supervised Ordinal SVM (WSO-SVM), which integrates all sources of data to train a robust model. The contributions of this paper are summarized as follows:

- We target an emerging biomedical field that aims to quantify the intra-tumoral genetic heterogeneity of GBM. There is limited existing work. The biomedical impact of this work is to allow non-invasive prediction of regional genetic alteration using MRI for each patient, which can inform future development of adaptive therapies for individualized oncology.
- We propose a new data-inclusive ML model, WSO-SVM, that trains robust classifiers to predict regional genetic alteration status within each GBM tumor using MRI. The novelty of this model is to leverage a combination of data sources including biopsy/labeled samples and unlabeled samples from the tumor and image data from the normal brain. This capability differentiates WSO-SVM from existing supervised learning and SSL algorithms.
- Based on a real-world dataset of 74 GBM patients, we demonstrate that WSO-SVM works significantly better than a variety of existing ML algorithms for predicting the regional genetic alteration of several GBM driver genes.

4.2 Weakly-Supervised Ordinal SVM (WSO-SVM)

We first discuss the WSO-SVM specially for GBM data in Section 4.2.1, and then interpret uncertainty qualification in Section 4.2.2. Finally we will discuss the extension of WSO-SVM to general case on weakly supervised learning in Section 4.2.3.

4.2.1 Mathematical Formulation

WSO-SVM integrates three sources of data: (a) Labeled/biopsy samples, (x_i, y_i), where $y_i = 1\ or\ 2$ denotes the altered or non-altered class of a gene for the i-th biopsy sample and x_i contains image features extracted from the sliding window at the biopsy location. For notation convenience, we denote the collection of biopsy samples by $D^{(1)} \cup D^{(2)} = \left(x_i^{(1)}\right)_{i=1}^{n_1} \cup \left(x_i^{(2)}\right)_{i=1}^{n_2}$, where $x_i^{(1)}$ and $x_i^{(2)}$ denote the image features from class 1 and 2 with n_1 and n_2 samples, respectively. (b) Unlabeled samples from t-ROI, $D^{(12)} = \left(x_j^{(12)}\right)_{j=1}^{m'_{12}}$; we do not know if these samples belong to class 1 or 2. (c) Normal brain samples from c-ROI, $D^{(3)} = \left(x_k^{(3)}\right)_{k=1}^{m_3}$, which are assigned to class 3.

Our ultimate goal is to build a robust classifier to differentiate class 1 and 2. A conventional supervised learning algorithm would use the data in (a) alone. A conventional SSL algorithm would use (a) and (b) alone. Our proposed WSO-SVM integrates all the data in (a), (b), (c) by exploiting the intrinsic order of class 1, 2, 3, which correspond to decreasing abnormality, to build an ordinal classifier. Different from conventional ordinal classification algorithms trained using labeled samples in each class, WSO-SVM additionally leverages unlabeled samples in (b) and imposes a mathematical constraint to "teach" the model that these samples should not be classified to class 3 even though their true membership to class 1 or 2 is unknown.

Next, we introduce the details of the WSO-SVM design. WSO-SVM aims to build two discriminant functions, f_1 and f_2, to differentiate altered samples (class 1) vs. non-altered samples (class 2) and tumoral samples (class 1&2) vs. normal brain samples (class 3), respectively, with a constraint of $f_1 \leq f_2$ to retain the intrinsic order of the three classes. Let $f_1 = h + b_1,\ f_2 = h + b_2$, with $b_1 \leq b_2$ and h being a shared function. We adopted the support vector formulation for h due to the success of SVM in various applications, and let $h(x) = w^T\phi(x)$, which ϕ contains non-linear transformations of the features. Then, we construct WSO-SVM as the following optimization problem:

$$\min \frac{1}{2} w^T w$$

subject to:

$$\left.\begin{array}{l} w^T\phi\left(x_i^{(1)}\right) + b_1 \geq 1 - \xi_i^{(1)};\ i \in 1, \dots, n_1\ \left(x_i^{(1)} \in D^{(1)}\right) \\ w^T\phi\left(x_i^{(2)}\right) + b_1 \leq -1 + \xi_i^{(2)};\ i \in 1, \dots, n_2\ \left(x_i^{(2)} \in D^{(2)}\right) \\ \sum_{i=1}^{n_1} \xi_i^{(1)} + \sum_{i=1}^{n_2} \xi_i^{(2)} \leq \epsilon \\ \xi_i^{(1)} \geq 0,\ i = 1, \dots, n_1; \quad \xi_i^{(2)} \geq 0,\ i \in 1, \dots, n_2 \end{array}\right\} \begin{array}{c} max - margin\ separation \\ between\ altered\ (class\ 1) \\ vs. non - altered\ (class\ 2) \\ samples \end{array}$$

$$\left.\begin{array}{l} w^T\phi\left(x_j^{(12)}\right) + b_2 \geq 1 - \zeta_j^{(12)};\ j = 1, \dots, m_{12}\left(x_j^{(12)} \in D^{(1)} \cup D^{(2)} \cup D^{(12)}\right) \\ w^T\phi\left(x_k^{(3)}\right) + b_2 \leq -1 + \zeta_k^{(3)};\ k \in 1, \dots, m_3\left(x_k^{(3)} \in D^{(3)}\right) \\ \sum_{j=1}^{m_{12}} \zeta_j^{(12)} + \sum_{k=1}^{m_3} \zeta_k^{(3)} \leq e \\ \zeta_j^{(12)} \geq 0, j = 1, \dots, m_{12};\ \zeta_k^{(3)} \geq 0, k \in 1, \dots, m_3 \end{array}\right\} \begin{array}{c} max - margin\ separation \\ between\ tumoral\ (class \\ 1\&2)\ vs. normal\ brain \\ (class\ 3)\ samples \end{array}$$

$$\left.\begin{array}{l} b_1 \leq b_2 \end{array}\right\} \begin{array}{c} obey\ intrinsic\ order\ of \\ ordinal\ classes \end{array}$$

(4.1)

All notations in (4.1) have been previously defined except that $m_{12} = n_1 + n_2 + m'_{12}$ is the total sample size of labeled and unlabeled samples from t-ROI. WSO-SVM minimizes the model complexity under a set of constraints. The first block of constraints aims to achieve max-margin separation between class 1 and class 2, where $\xi_i^{(1)}$ and $\xi_i^{(2)}$ are slack variables to achieve soft-margin similar to SVM. The second block aims to achieve max-margin separation between class 1&2 and class 3 with $\zeta_j^{(12)}$ and $\zeta_k^{(3)}$ being slack variables. ϵ and e are tuning parameters. The last constraint is to enforce the discriminant functions to obey the intrinsic order of the three ordinal classes.

It is easier to solve the WSO-SVM optimization in its dual form which is given in Proposition 4.1 (See the proof in Appendix E).

Proposition 4.1: The dual form of the primal WSO-SVM optimization problem in (4.1) is:

$$\min_{\alpha,\beta} \ \frac{1}{2}\gamma^T Y K Y \gamma - \sum_{i=1}^{n_1} \alpha_i^{(1)} - \sum_{i=1}^{n_2} \alpha_i^{(2)} - \sum_{j=1}^{m_{12}} \beta_j^{(12)} - \sum_{k=1}^{m_3} \beta_k^{(3)},$$

subject to:

$$\sum_{i=1}^{n_1} \alpha_i^{(1)} - \sum_{i=1}^{n_2} \alpha_i^{(2)} + \sum_{j=1}^{m_{12}} \beta_j^{(12)} - \sum_{k=1}^{m_3} \beta_k^{(3)} = 0,$$

$$\sum_{i=1}^{n_1} \alpha_i^{(1)} - \sum_{i=1}^{n_2} \alpha_i^{(2)} \geq 0,$$

$$0 \leq \alpha_i^{(1)} \leq C_1, i = 1, \ldots, n_1; 0 \leq \alpha_i^{(2)} \leq C_1, i = 1, \ldots, n_2,$$

$$0 \leq \beta_j^{(12)} \leq C_2, j = 1, \ldots, m_{12};\ 0 \leq \beta_k^{(3)} \leq C_2, k = 1, \ldots, m_3,$$

where $\gamma = (\alpha_1^{(1)}, \ldots, \alpha_{n_1}^{(1)}, \alpha_1^{(2)}, \ldots, \alpha_{n_2}^{(2)}, \beta_1^{(12)}, \ldots, \beta_{m_{12}}^{(12)}, \beta_1^{(3)}, \ldots, \beta_{m_3}^{(3)})$,

$Y = diag\left(\overbrace{1,\ldots,1}^{n_1}, \overbrace{-1,\ldots,-1}^{n_2}, \overbrace{1,\ldots,1}^{m_{12}}, \overbrace{-1,\ldots,-1}^{m_3}\right)$, and K is a covariance matrix with $K_{ij} = \phi(x_i)^T\phi(x_j) = k(x_i, x_j)$ that can be computed by a kernel function defined on the feature space. C_1 and C_2 are tuning parameters.

The dual problem is a convex quadratic programming problem, which can be solved by a standard quadratic optimization solver such as CPLEX.

Once the optimal solutions of α and β in the dual problem are obtained, we can obtain the optimal coefficients in the primal problem, w, and further get $h(x) = \sum_{i=1}^{n_1}\alpha_i^{(1)}k\left(x, x_i^{(1)}\right) - \sum_{i=1}^{n_2}\alpha_i^{(2)}k\left(x, x_i^{(2)}\right) + \sum_{j=1}^{m_{12}}\beta_j^{(12)}k\left(x, x_j^{(12)}\right) - \sum_{k=1}^{m_3}\beta_k^{(3)}k\left(x, x_k^{(3)}\right)$. Also, b_1 and b_2 can be estimated as: $b_1 = y - h(x)$ for any $(x, y) \in D^{(1)}$ (or $D^{(2)}$) whose corresponding $\alpha^{(1)}$ (or $\alpha^{(2)}$) satisfies $0 < \alpha^{(1)}$(or $\alpha^{(2)} < C_1$); $b_2 = y - h(x)$ for any $(x, y) \in D^{(1)} \cup D^{(2)} \cup D^{(12)}$(or $D^{(3)}$) whose corresponding $\beta^{(12)}$ (or $\beta^{(3)}$) satisfies $0 < \beta^{(12)}(or\ \beta^{(3)} < C_2)$. Then, we can obtain the discriminant functions for any new sample x^*, i.e., $f_1(x^*) = sign(h(x^*) + b_1)$ and $f_2(x^*) = sign(h(x^*) + b_2)$. The decision rule for classifying the new sample x^* is: it belongs to class 1 if $f_1(x^*) \geq 0$, to class 2 if $f_1(x^*) < 0$ & $f_2(x^*) \geq 0$, and to class 3 if $f_2(x^*) < 0$.

4.2.2 Uncertainty Quantification

Calibration in statistics aims to calibrate the predictive uncertainty of individual samples. It can be applied to both regression and classification problems. In regression, it aims to estimate other values of the independent variable from the new observations of

dependent variable, which is also well known as inverse regression [96]. Calibration in classification aims to infer the classification probability with the classification score. There are various methods for calibrating, including assignment value approach [97], Bayes approach [98], Isotonic regression [99], Beta calibration [100]. In this chapter, Beta calibration via logistic regression will be adopted, which is a most popular, well-founded and easily implemented method for calibrating compared to other methods [100]. Beta calibration via logistic regression takes advantage of training data, model output score vectors on training instances, as well as output score vector on test instances, with the detailed algorithm shown in [100].

4.2.3 Extension and Discussion

The work can be extended to the general framework of weakly supervised ordinal learning with K ordinal classes. Let $\mathcal{X} = \mathbb{R}^d$ be the d-dimensional feature space and $\mathcal{Y} = \{1, \dots, K\}$ be the label space. Different from multi-class classification, the consecutive integers in $\{1, \dots, K\}$ follow an order. In this dissertation, we use 'labels' and 'classes' interchangeable to refer to the integers contained in $\mathcal{Y}$.

A typical ordinal model consists of a set of ranking functions, $f_k, k = 1, \dots, K-1$, which satisfy the constraint of $f_1 \leq \cdots \leq f_{K-1}$. To predict the label for a sample $\mathbf{x}$, one can compute the outputs from the ranking functions for this sample, $f_1(\mathbf{x}), \dots, f_{K-1}(\mathbf{x})$. Then, the predicted label can be obtained by

$$J(\mathbf{x}) = arg \min_{k=1,\dots,K-1} \{k: f_k(\mathbf{x}) \geq 0\} = 1 + \sum_{k=1}^{K-1} I(f_k(\mathbf{x}) < 0). \quad (4.2)$$

where $I(\cdot)$ is an indicator function. That is, the predicted label is the number of negative ranking functions in the sequence plus one. For example, if all ranking functions are non-negative, the predicted label is 1; if all ranking functions are negative, the predicted label is K; if the first k ($1 \leq k < K-1$) ranking functions are negative and the remaining ones are non-negative, the predicted label is $k+1$.

Furthermore, the ranking functions can be decomposed into a common function and class-specific intercepts, $f_k(\mathbf{x}) = h(\mathbf{x}) + b_k$ with $b_1 \leq \cdots \leq b_{K-1}$. $h(\mathbf{x}) = \boldsymbol{\eta}^T \phi(\mathbf{x})$, where ϕ includes transformations of the feature vector $\mathbf{x}$ and $\boldsymbol{\eta}$ contains the combination coefficients. Depending on the form of ϕ, the WSO model can be linear or non-linear. A training dataset is needed to learn the parameters such as $\boldsymbol{\eta}, b_1, \ldots, b_{K-1}$.

Consider a training dataset of n samples, $\{(\mathbf{x}_i, \mathcal{F}_i), i = 1, \ldots., n\}$. In the conventional ordinal learning setting, every training sample must have one and only one label. WSO allows the training set to include samples with interval labels, i.e., $\mathcal{F}_i = [Y_i^l, Y_i^r] \subseteq \mathcal{Y}$. For example, a sample may have an interval label of [2,4], meaning that the sample can be from class 2, 3, or 4, but we do not know which precise class it is from. HOL can incorporate samples with both interval and precise labels. When $Y_i^l < Y_i^r$, $\mathcal{F}_i$ denotes the interval label. When $Y_i^l = Y_i^r$, $\mathcal{F}_i$ denotes the precise label.

The goal of WSO is to learn an ordinal learning model based on a training set with the aforementioned characteristics. This can be formulated as the following optimization problem:

$$\min_{\mathbf{f}=(f_1,\ldots,f_{K-1})} \sum_{i=1}^{n} L(J(\mathbf{x}_i), \mathcal{F}_i) + \mu \left|\left|\mathbf{f}\right|\right|_{\mathcal{H}},$$

$$\text{s.t., } f_1 \leq \cdots \leq f_{K-1}, \tag{4.3}$$

where L is a loss function defined on the training set, $\left|\left|\cdot\right|\right|_{\mathcal{H}}$ is a norm in a metric space $\mathcal{H}$ to regularize the complexity of the ranking functions, and μ controls the trade-off between the loss and model complexity. Because $\mathcal{F}_i$ can be an interval, commonly used loss functions for supervised learning models are not applicable.

In this dissertation, we focus on one specific form of the loss function for computational ease, i.e., the 0/1 loss: A loss of '1' is incurred if the predicted label J_i falls outside the true interval $\mathcal{F}_i$, and the loss is '0' otherwise, i.e.,

$$L_{0/1}(J_i, \mathcal{F}_i) = I(J_i \notin \mathcal{F}_i). \tag{4.4}$$

Here we provide an example to illustrate these loss functions. Consider a sample with true label interval [2,4]. Under the 0/1 loss, if the predicted label is 6, the loss is 1.

It is difficult to solve the WSO optimization in (4.3) especially when the ranking functions $f_1, \ldots, f_{K-1}$ are non-linear. This is because the ranking functions are embedded in the loss functions in a complicated form, which makes the optimization intractable. To tackle this challenge, we propose a conversion method that converts the original WSO optimization into an equivalent formulation of learning $K-1$ binary classifiers with coupled parameters. Because binary classification has been much better studied in the literature, this conversion allows us to borrow ideas from binary classification to effectively and efficiently solve the WSO optimization.

Specifically, consider each ranking function f_k to be a binary classifier: if $f_k(\mathbf{x}) > 0$, $\mathbf{x}$ is classified to the interval $[1, k]$; otherwise, $\mathbf{x}$ is classified to the interval $[k+1, K]$. To train f_k, we use a subset of training samples whose label $\mathcal{F}_i$ is included in $[1, k]$ or $[k+1, K]$. This is a subset of the whole training set because we cannot include samples whose label interval includes k. Denote this subset by $D_k = \{(\mathbf{x}_i, \mathcal{F}_i) | \, Y_i^r = k \text{ or } Y_i^l = k+1; \; i = 1, \ldots, n\}$. Next, we can define the loss function of training f_k as:

$$\sum_{i \in D_k} I\big(Z_{k,i} f_k(\mathbf{x}_i) < 0\big),$$

where $Z_{k,i} = 1 \; or \; -1$ corresponds to $\mathcal{F}_i \subseteq [1, k]$ or $\mathcal{F}_i \subseteq [k+1, K]$, respectively. A loss of '1' is incurred for a sample if the predicted and the true classes of the sample do not agree. Finally, we can sum up the loss of each f_k and get the total loss for training the $K-1$ binary classifiers simultaneously, i.e.,

$$\mathcal{B}(f, Z) \triangleq \sum_{k=1}^{K-1} \sum_{i \in D_k} I\big(Z_{k,i} f_k(\mathbf{x}_i) < 0\big). \tag{4.5}$$

Theorem 4.1 proves that $\mathcal{B}(f, Z)$ is equivalent to the WSO loss in (4.3).

Theorem 4.1: Let $L(J, \mathcal{F}) \triangleq \sum_{i=1}^{n} L(J(\mathbf{x}_i), \mathcal{F}_i)$ denote the WSO loss in (4.3). $\mathcal{B}(f, Z)$ is the total loss of $K-1$ binary classifiers based on $f_1, \ldots, f_{K-1}$, as defined in (4.5). Then, $L(J, \mathcal{F}) = \mathcal{B}(f, Z)$.

Based on Theorem 1, we can convert the WSO optimization in (4.3) into an equivalent form as:

$$\min_{\mathbf{f}=(f_1, \ldots, f_{K-1})} \sum_{k=1}^{K-1} \sum_{i \in D_k} I\big(Z_{k,i} f_k(\mathbf{x}_i) < 0\big) + \mu \big|\big|\mathbf{f}\big|\big|_{\mathcal{H}},$$

$$\text{s.t., } f_1 \leq \cdots \leq f_{K-1}. \tag{4.6}$$

To solve this optimization problem is to train $K-1$ binary classifiers with coupled parameters in $f_1, \ldots, f_{K-1}$, which is more tractable than solving the original optimization. To solve the optimization in (4.6), we first propose to use a hinge loss as a surrogate for the indicator function in (4.6) to make the optimization more tractable and efficient to solve. The hinge loss is a convex upper bound of the indicator function. Using the hinge loss and spelling out the ranking functions as $f_k(\mathbf{x}_i) = \boldsymbol{\eta}^T \phi(\mathbf{x}_i) + b_k$, (4.6) becomes:

$$\min_{\boldsymbol{\eta}, b_1, \ldots, b_{K-1}} \sum_{k=1}^{K-1} \sum_{i \in D_k} \max(0, 1 - Z_{k,i}(\boldsymbol{\eta}^T \phi(\mathbf{x}_i) + b_k)) + \mu \big|\big|\boldsymbol{\eta}\big|\big|_{\mathcal{H}},$$

$$\text{s.t., } b_1 \leq \cdots \leq b_{K-1}. \tag{4.7}$$

The constraint in (4.7) enforces the set of discriminative ranking functions. It is clear to see that WSO-SVM for GBM data is a special case of the formulation of (4.7) with three ordinal classes, i.e., $K = 3$.

Next, we will present the algorithms for solving (4.7) under the 0/1 loss. It is easier to solve the WSO optimization in its dual form which is given in Proposition 4.2.

Proposition 4.2: The dual form of the primal WSO-SVM optimization problem in (4.7) is:

$$\min_{\boldsymbol{\eta}, \boldsymbol{\xi}} \frac{1}{2} \big|\big|\boldsymbol{\eta}\big|\big|^2 + C \sum_{k=1}^{K-1} \sum_{i \in D_k} \zeta_i^k$$

$$\text{s.t., } Z_{k,i}(\boldsymbol{\eta}^T \phi(\mathbf{x}_i) + b_k) \geq 1 - \zeta_i^k;$$

$$\zeta_i^k \geq 0;$$

$$i \in D_k\ , k = 1, \dots, K-1;$$

$$b_1 \leq \cdots \leq b_{K-1}.$$

The dual problem is a convex quadratic programming problem. When the sample size is small, it can be solved directly by a standard quadratic optimization solver such as CPLEX. When the sample size is large, we can use Sequential minimal optimization (SMO) algorithm for efficient computation [101].

4.3 Application

4.3.1 Data Collection

This study used data from a cohort of 74 GBM patients with IRB approval from Barrow Neurological Institute (BNI) and Mayo Clinic Arizona (MCA). A total of 318 biopsy samples were acquired from these patients (average: 4; range: 1-13). Array CGH data on a subset of biopsy samples was available as previously described [15] [16]. Whole exome sequencing (WES) was performed on the remaining biopsy samples and their paired blood samples. Quality control was performed on raw sequencing data using the MultiQC toolkit. Paired-end clean reads were aligned to GRCh37/hg19 human reference using Burrows-Wheeler Aligner2, and further processed by GATK3 to remove low mapping quality reads and to re-align around the indels. Somatic SNVs and indels were identified by integrating the results from six algorithms for variants calling: Freebayes5, MuTect26, TNhaplotyper7, TNscope7, TNsnv7, and VarScan28. Somatic copy number and tumor purity were estimated from WES by PureCN12. GISTIC213 analysis was then applied to

integrate results from individual patients and identify genomic regions recurrently amplified or deleted in glioma samples.

We focused on three GBM driver genes: EGFR, PDGFRA, and PTEN. For each gene, we considered the gene is altered (class 1) if it has an abnormal CNV or is mutated, and non-altered (class 2) otherwise. For EGFR and PDGFRA, we followed the literature [78] and considered amplification as abnormal CNV; for PTEN, deletion or loss was considered as abnormal CNV [102]. To maximize the sample size in ML training, we included all available samples for each gene. There are 130/171, 53/238, and 206/109 biopsy samples with altered/non-altered EGFR, PDGFRA, and PTEN, respectively.

4.3.2 Image Pre-processing and Feature Extraction

Each patient went through a pre-operative multiparametric MRI exam, from which we obtained five contrast images: T1-weighted contrast-enhanced image (T1+C), T2-weighted image (T2), mean diffusivity (MD), fractional anisotropy (FA), and relative cerebral blood volume (rCBV). Detailed MRI protocols and pre-processing steps can be found in our prior publications [56] [55]. It is well-known that a GBM tumor contains a contrast-enhancing portion (CE) and a non-enhancing portion (NE). CE and NE were manually segmented following standard procedures [56] [78]. The union of CE and NE composes the whole tumoral Region of Interest (t-ROI) for which we wanted to ultimately generate predictions of sub-regional genetic alterations to characterize the intra-tumoral heterogeneity for each patient. Additionally, we ran an automatic algorithm to find the contralateral ROI (c-ROI) of the t-ROI to represent the normal brain.

A sliding window of 8×8 pixel^2 (the size approximately of a biopsy sample) was

placed at each pixel within the t-ROI. For each sliding window, we computed texture features using Gray-Level Co-occurrence Matrix (GLCM) [103] and Gabor Filters [104], two commonly used texture analysis algorithms, and 1st-order statistical features. Collectively, our pipeline generated 280 features from multiparametric MRI images for each sliding window. Additionally, we included two location-related features: the (x, y, z) coordinates for the center of a sliding window; a binary variable that indicates whether a sliding window is in NE or CE. We applied the same sliding window approach and extracted features from the c-ROI.

4.3.3 Application of WSO-SVM

4.3.3.1 Training and cross validation (CV)

All biopsy samples were divided into 10 folds. In each iteration of the CV, WSO-SVM was trained based on 9 folds of the biopsy samples, together with randomly selected unlabeled samples and samples from c-ROI of the same size. After the model was trained, it was applied to the remaining fold of biopsy samples to compute the classification accuracy of class 1 vs. class 2. It was also applied to all samples from t-ROI and c-ROI excluding those used in training to compute the classification accuracy of class 1&2 vs. class 3. This latter classification was not hard and multiple tuning parameters could achieve >80% accuracy for every patient. Among these parameters, we chose those with the highest CV accuracy on biopsy samples. We ran this CV scheme for 30 times and reported the averaged accuracy.

4.3.3.2 Predictive map generation within t-ROI

For each patient, we re-trained WSO-SVM by including randomly selected unlabeled samples and samples from c-ROI to personalize the model under the previously found optimal tuning parameters. Then, we applied the model to predict the gene status for every sample/sliding window within the t-ROI of the patient, which composed the predictive maps.

4.3.3.3 Comparative methods

We compared the performance of WSO-SVM with a range of existing algorithms, such as:

- Supervised learning, including typical algorithms such as SVM, random forest, and logistic elastic net regression;
- SSL, including popular algorithms such as T-SVM [105], Laplacian SVM [106], and co-training [107];
- Multi-task learning (MTL), including regularized MTL [108] and multi-task Gaussian Process (GP) [109], which couple the model trainings of the three genes together.

4.3.3.4 Results

Table 4.1-4.3 show the CV classification accuracy of each gene by different algorithms. WSO-SVM achieved the highest accuracy, sensitivity, and specificity for all three genes. For EGFR, the accuracy of WSO-SVM is 0.8, whereas the range of accuracies by the other algorithms is 0.60-0.74. For PTEN, the accuracy of WSO-SVM is 0.8, whereas the range of accuracies by the other algorithms is 0.51-0.67. For PDGFRA, the accuracy

of WSO-SVM is 0.71, whereas the range of accuracies by the other algorithms is 0.52-0.65. Among the three genes, WSO-SVM achieved a better performance for classifying EGFR and PTEN than PDGFRA, due to the class imbalance of PDGFRA.

Table 4.1 – Classification accuracy of EGFR

ML algorithms		Accuracy	Sensitivity	Specificity
Supervised learning	SVM	0.69	0.56	0.79
	Random forest	0.74	0.66	0.79
	Logistic elastic net regression	0.60	0.54	0.64
Semi-supervised learning	T-SVM	0.60	0.61	0.59
	Laplace-SVM	0.68	0.64	0.70
	Co-training	0.69	0.73	0.66
Multi-task learning	Multi-task GP	0.68	0.67	0.68
	Regularized MTL	0.64	0.64	0.64
WSO-SVM		**0.80**	**0.79**	**0.81**

Table 4.2 – Classification accuracy of PDGFRA

ML algorithms		Accuracy	Sensitivity	Specificity
Supervised learning	SVM	0.65	0.63	0.65
	Random forest	0.65	0.60	0.66
	Logistic elastic net regression	0.61	0.50	0.63
Semi-supervised learning	T-SVM	0.54	0.57	0.53
	Laplace-SVM	0.58	0.64	0.57
	Co-training	0.62	0.62	0.62
Multi-task learning	Multi-task GP	0.59	0.69	0.57
	Regularized MTL	0.52	0.65	0.49
WSO-SVM		**0.71**	**0.70**	**0.72**

Table 4.3 – Classification accuracy of PTEN

ML algorithms		Accuracy	Sensitivity	Specificity
Supervised learning	SVM	0.60	0.57	0.66
	Random forest	0.67	0.65	0.70
	Logistic elastic net regression	0.51	0.45	0.61
Semi-supervised learning	T-SVM	0.57	0.56	0.58
	Laplace-SVM	0.58	0.56	0.60
	Co-training	0.62	0.62	0.62
Multi-task learning	Multi-task GP	0.61	0.62	0.61
	Regularized MTL	0.54	0.47	0.65
WSO-SVM		**0.80**	**0.78**	**0.83**

Our data suggest that leveraging both normal samples in contralateral ROI and ordering relationship can further optimize model performance and applicability. When stratifying WSO-SVM sample predictions based on predictive uncertainty, we observed a striking difference in model performance. For PDGFRA modeling, the subgroup of sample predictions with lower uncertainties achieved the better predictive performance as shown in Figure 4.1: compared to the entire cohort as a whole (71% accuracy, 100% biopsy samples), 71% biopsy samples achieved 78% accuracy with predictive probabilities greater than 0.6; 45% biopsy samples achieved 82% accuracy with predictive probabilities greater than 0.7; 26% biopsy samples achieved 85% accuracy with predictive probabilities greater than 0.8; the subgroup of sample predictions with the lowest uncertainty (predictive probabilities greater than 0.9) achieved the highest predictive performance (94% accuracy). Similar findings can be observed from the modeling of EGFR and PTEN in Figure 4.1. These results suggest that predictive uncertainty can inform the likelihood of achieving an accurate sample prediction.

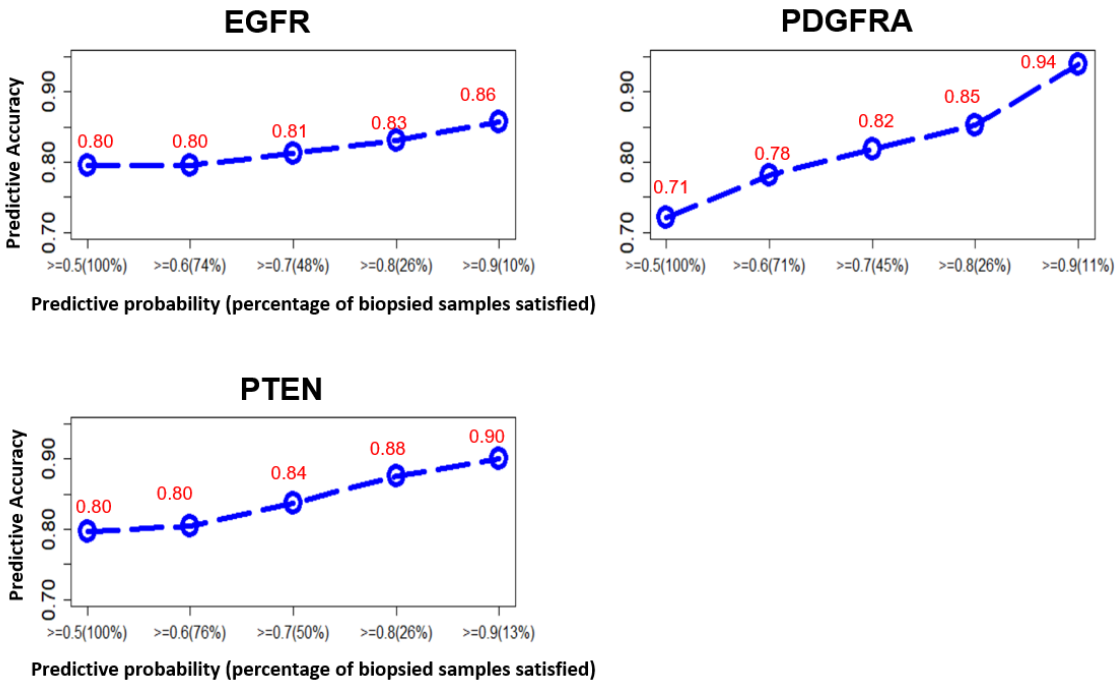

Figure 4.1 – Differences in predictive accuracy related to certain versus uncertain sample predictions by WSO-SVM

Finally, the trained WSO-SVM models were used to generate prediction maps for each patient. For demonstration, Figure 4.2 shows the prediction maps for four different patients (Yellow dots represent biopsy samples whose predicted gene statuses by WSO-SVM are reported underneath the maps (all predictions are correct)). The alterations in EGFR and PDGFRA promote tumor growth. Thus, we showed their co-alteration patterns in one map. PTEN is a tumor suppressor gene, whose alteration is shown in a separate map. Patient A demonstrates predominant regions with EGFR alteration, with scattered regions of PDGFRA co-alteration; the PTEN map shows largely non-alteration. For patient B, the PTEN map shows an opposite pattern, whereas the EGFR & PDGFRA map demonstrates a similar pattern as patient A. In contrast to patient A and B, patient C demonstrates

predominant regions with PDGFRA alteration. For patient D, the regions with EGFR & PDGFRA co-alteration are relatively concentrated compared to the other patients. These examples demonstrated the great extent of intra-tumoral genetic heterogeneity for each patient.

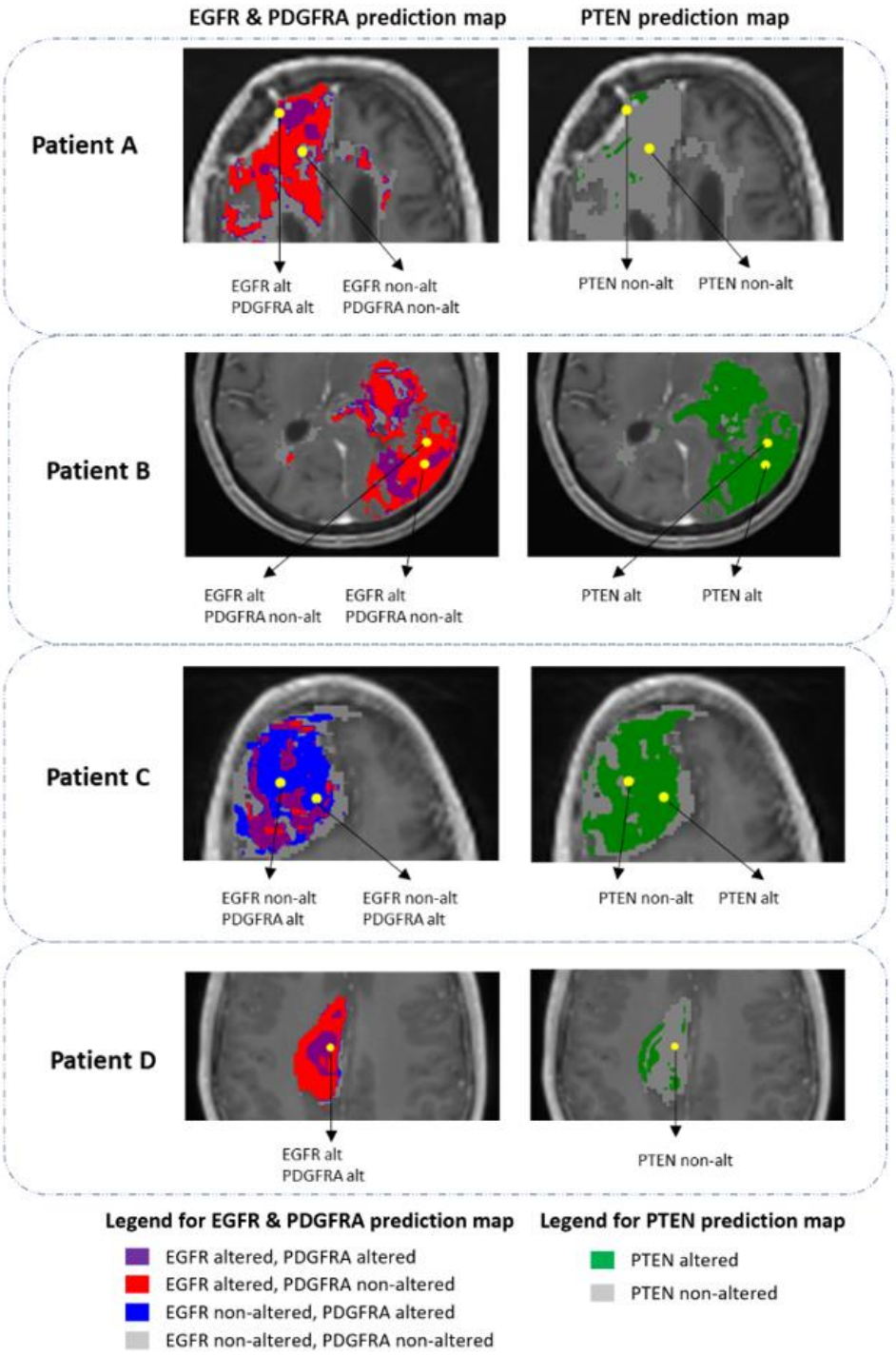

Figure 4.2 – EGFR & PDGFRA prediction map (left column) and PTEN prediction map (right column) in t-ROI for four patients (rows)

(a) EGFR

(b) PDGFRA

(c) PTEN

Figure 4.3 – Patient-wise proportions of alteration vs. non-alteration for (a) EGFR, (b) PDGFRA, and (c) PTEN within t-ROI

4.3.3.5 Discussion

We proposed a new ML algorithm, WSO-SVM, to predict regional genetic alteration within each GBM tumor using MRI. WSO-SVM outperformed supervised learning, SSL, and multi-task learning algorithms, which suggested that the data-inclusive strategy adopted by WSO-SVM was effective.

Due to space limit, we only showed the predictions maps for four patients in Figure 4.2. In Fig. 4.3, we aggregated the prediction map of every patient in our dataset into a pie chart. These results revealed a great amount of variability between patients in terms of the genetic alteration patterns. Within each individual patient's tumor, there is also region-to-region variation for the genetic alteration patterns. This is consistent with findings that point out the intra-tumoral genetic heterogeneity as a contributing factor to the ineffectiveness of current treatment approaches [56] [66] [67] [23] [24].

This study has several limitations. First, there are uncertainties in the upstream processes to ML, such as image pre-processing, spatial matching of biopsy locations to the MRI space, and genetic analysis. While we have attempted to mitigate these uncertainties by following well-established protocols, it will be beneficial to evaluate the robustness of WSO-SVM to these uncertainties in future study. Secondly, we randomly selected unlabeled samples an samples from c-ROI. More efficient sample selection strategies may be tried to enhance classification performance. Thirdly, we noted that the PDGFRA classification was not as accurate as EGFR\PTEN due to the heavy class imbalance of PDGFRA. More PDGFRA altered samples need to be collected to improve performance. Lastly, our model was trained and validated using a dataset composed of 318 biopsy

samples from 74 patients. Further validation will be needed to assess reproducibility in the context of other independent datasets.

4.4 Conclusion

We developed a data-inclusive WSO-SVM model to predict regional genetic alteration status within each GBM tumor using MRI. This study demonstrated the feasibility of using MRI and WSO-SVM to enable non-invasive prediction of regional genetic alteration for each patient, which can inform future adaptive therapies for individualized oncology.

CHAPTER 5: CONCLUSION

The goal of this work is to study model uncertainty of image-based ML in the context of precision medicine of brain cancer. Specifically, we focus on developing ML models to predict intra-tumor heterogeneity of genomic and molecular markers based on multi-contrast MRI data for GBM.

In Chapter 2, we developed a Semi-supervised Gaussian Process with Uncertainty-minimizing Feature-selection (SGP-UF), which can incorporate selected unlabeled samples (i.e. unbiopsied regions of a tumor) in the model training, and integrates feature selection with a new criterion of seeking features that minimize the prediction uncertainty. The model generated predictions for regional EGFR amplification status to resolve the intratumoral genetic heterogeneity across each individual tumor. The model used probability distributions for each sample prediction to quantify uncertainty, and used transductive learning to reduce the overall uncertainty. We demonstrated that SGP-UP significantly reduces prediction uncertainty while at the same time achieving higher accuracy. This should help integrate more reliable radiogenomics models for improved medical decision-making.

In Chapter 3, we proposed a novel machine learning framework, KGL data fusion model, to fuse the three sources of data/information to generate a spatial prediction. A novel mathematical formulation was proposed and solved with theoretical study. We presented a real-data application of predicting the spatial distribution of tumor cell density —an important molecular marker for brain cancer. A total of 82 biopsy samples were acquired from 18 patients with glioblastoma, together with six MRI contrast images from

each patient and biological knowledge encoded by a PDE simulator-based mechanistic model – PI model. KGL achieved the highest prediction accuracy and minimum prediction uncertainty compared with a variety of competing methods. The result has important implications for providing individualized, spatially optimized treatment for each patient.

In Chapter 4, we proposed a data-inclusive machine learning model, Weakly Supervised Ordinal (WSO) learning, to predict regional genetic alterations of each tumor. The novelty of WSO is to leverage the vast amount of MRI data including unlabeled data outside the sparsely sampled biopsies within brain tumor and normal brain samples outside the brain tumor to improve accuracies. Our study included a unique dataset of 318 biopsies with spatially matched MRI from 71 GBM patients. 10-fold cross validation accuracies for predicting alterations of driver genes including EGFR, PDGFRA and PTEN outperformed a variety of existing ML methods. We generated regional genetic alteration maps for each patient within the tumoral areas.

The work presented in this thesis demonstrates that these proposed novel methods significantly reduce prediction uncertainty while at the same time achieving higher accuracy in precision medicine, which can inform personalized targeted treatment decisions that potentially improve clinical outcome. The proposed models can be applied to multiple application domains including robotics and autonomous systems, prosthetics and human enhancement, economics and so on.

APPENDICES

Appendix A: Proof of Theorem 2.1

Let $D_U = \{\mathbf{X}_U\}$ denote samples from a test set (e.g., a patient of interest). Consider the prediction for a test sample $\mathbf{x}^*$. Let σ^{*2} and σ_{tran}^{*2} denote the predictive variances from the standard GP and SGP models, respectively. Our objective is to prove $\sigma_{tran}^{*2} < \sigma^{*2}$.

Let $M = \begin{pmatrix} K(\mathbf{X}_L, \mathbf{X}_L) + \sigma^2 \mathbf{I} & K(\mathbf{X}_U, \mathbf{X}_L)^T \\ K(\mathbf{X}_U, \mathbf{X}_L) & K(\mathbf{X}_U, \mathbf{X}_U) + \sigma^2 \mathbf{I} \end{pmatrix} \triangleq \begin{pmatrix} A & B \\ B^T & D \end{pmatrix}$. Its inverse matrix M^{-1} can be partitioned into 4 sub-matrices

$$M \triangleq \begin{pmatrix} E & F \\ F^T & H \end{pmatrix}$$

where $E = A^{-1} + A^{-1}B(D - B^T A^{-1} B)^{-1} B^T A^{-1}$; $F = -A^{-1}B(D - B^T A^{-1} B)^{-1}$; $H = (D - B^T A^{-1} B)^{-1}$.

$$\sigma_{tran}^{*2} = K(\mathbf{x}^*, \mathbf{x}^*) - \begin{pmatrix} K(\mathbf{x}^*, \mathbf{X}_L)^T \\ K(\mathbf{x}^*, \mathbf{X}_U)^T \end{pmatrix}^T \begin{pmatrix} K(\mathbf{X}_L, \mathbf{X}_L) + \sigma^2 \mathbf{I} & K(\mathbf{X}_U, \mathbf{X}_L)^T \\ K(\mathbf{X}_U, \mathbf{X}_L) & K(\mathbf{X}_U, \mathbf{X}_U) + \sigma^2 \mathbf{I} \end{pmatrix}^{-1} \begin{pmatrix} K(\mathbf{x}^*, \mathbf{X}_L)^T \\ K(\mathbf{x}^*, \mathbf{X}_U)^T \end{pmatrix}$$

$$= K(\mathbf{x}^*, \mathbf{x}^*) - \begin{pmatrix} K(\mathbf{x}^*, \mathbf{X}_L)^T \\ K(\mathbf{x}^*, \mathbf{X}_U)^T \end{pmatrix}^T \begin{pmatrix} E & F \\ F^T & H \end{pmatrix} \begin{pmatrix} K(\mathbf{x}^*, \mathbf{X}_L)^T \\ K(\mathbf{x}^*, \mathbf{X}_U)^T \end{pmatrix}$$

$$= K(\mathbf{x}^*, \mathbf{x}^*) - (K(\mathbf{x}^*, \mathbf{X}_L) E K(\mathbf{x}^*, \mathbf{X}_L)^T + 2\, K(\mathbf{x}^*, \mathbf{X}_L) F K(\mathbf{x}^*, \mathbf{X}_U)^T + K(\mathbf{x}^*, \mathbf{X}_U) H K(\mathbf{x}^*, \mathbf{X}_U)^T)$$

$$= K(\mathbf{x}^*, \mathbf{x}^*) - K(\mathbf{x}^*, \mathbf{X}_L) A K(\mathbf{x}^*, \mathbf{X}_L)^T - K(\mathbf{x}^*, \mathbf{X}_L) A^{-1} B (D - B^T A^{-1} B)^{-1} B^T A^{-1} K(\mathbf{x}^*, \mathbf{X}_L)^T$$

$$+ 2K(\mathbf{x}^*, \mathbf{X}_L) A^{-1} B (D - B^T A^{-1} B)^{-1} K(\mathbf{x}^*, \mathbf{X}_U)^T - K(\mathbf{x}^*, \mathbf{X}_U)(D - B^T A^{-1} B)^{-1} K(\mathbf{x}^*, \mathbf{X}_U)^T$$

$$= \sigma^{*2} - (K(\mathbf{x}^*, \mathbf{X}_L) A^{-1} B - K(\mathbf{x}^*, \mathbf{X}_U))(D - B^T A^{-1} B)^{-1} (B^T A^{-1} K(\mathbf{x}^*, \mathbf{X}_L)^T - K(\mathbf{x}^*, \mathbf{X}_U)^T)$$

$$= \sigma^{*2} - Z^T (D - B^T A^{-1} B)^{-1} Z$$

where $Z = B^T A^{-1} K(\mathbf{x}^*, \mathbf{X}_L)^T - K(\mathbf{x}^*, \mathbf{X}_U)^T$.

Since the kernel matrix is positive semi-definite, then its inverse matrix M is also positive semi-definite. Then its leading principal sub-matrix H is also positive semi-definite. That is, $(D - B^T A^{-1} B)^{-1}$ is positive semi-definite. Then we can know $Z^T (D - B^T A^{-1} B)^{-1} Z \geq 0$. Hence, ${\sigma^*_{tran}}^2 \leq {\sigma^*}^2$.

Appendix B: Proof of Theorem 2.2

Let $D_{KL}(f^*, f^*_{tran})$ denote Kullback–Leibler divergence of f^*_{tran} from f^*.

$$D_{KL}(f^*, f^*_{tran}) = \int f log\left(\frac{f}{f_{tran}}\right) d\boldsymbol{x} \leq \int f|log\left(\frac{f}{f_{tran}}\right)|d\boldsymbol{x}$$

$$= \int f|\frac{1}{2}log(\sigma^{*2}) + \frac{\left(y^* - K(\mathbf{x}^*, \mathbf{X}_L)\left(K(\mathbf{X}_L, \mathbf{X}_L) + \sigma^2\mathbf{I}\right)^{-1}\mathbf{y}_L\right)^2}{\sigma^{*2}} - \frac{1}{2}log(\sigma^{*}_{tran}{}^2)$$

$$- \frac{\left(y^* - K_{tran}\left(K(\mathbf{X}_L, \mathbf{X}_L) + \sigma^2\mathbf{I}\right)^{-1}\mathbf{y}_L\right)^2}{\sigma^{*}_{tran}{}^2}|d\boldsymbol{x}$$

$$\leq \int f(\frac{1}{2}\left|log(\sigma^{*2}) - log(\sigma^{*}_{tran}{}^2)\right| + |\frac{\left(y^* - K(\mathbf{x}^*, \mathbf{X}_L)\left(K(\mathbf{X}_L, \mathbf{X}_L) + \sigma^2\mathbf{I}\right)^{-1}\mathbf{y}_L\right)^2}{\sigma^{*2}}$$

$$- \frac{\left(y^* - K_{tran}\left(K(\mathbf{X}_L, \mathbf{X}_L) + \sigma^2\mathbf{I}\right)^{-1}\mathbf{y}_L\right)^2}{\sigma^{*}_{tran}{}^2}|)d\boldsymbol{x}$$

According to Theorem 2.1, $\left|log(\sigma^{*2}) - log(\sigma^{*}_{tran}{}^2)\right| = \left|log(\sigma^{*2}) - log(\sigma^{*2} - Z^T(D - B^TA^{-1}B)^{-1}Z)\right|$.

From the known conditions, i.e., $K(\mathbf{X}_U, \mathbf{X}_L) \to \mathbf{0}$; $K(\mathbf{X}_U, \mathbf{x}^*) \to \mathbf{0}$, then it is obvious that $Z \to 0$. It's easy to infer $Z^T(D - B^TA^{-1}B)^{-1}Z \to 0$. Hence, $\left|log(\sigma^{*2}) - log(\sigma^{*}_{tran}{}^2)\right| \to 0$. Similarly, we can know $K(\mathbf{x}^*, \mathbf{X}_U)\left(K(\mathbf{X}_U, \mathbf{X}_U) + \sigma^2\mathbf{I}\right)^{-1}K(\mathbf{X}_U, \mathbf{X}_L) \to \mathbf{0}$. Then $K_{tran}\left(K(\mathbf{X}_L, \mathbf{X}_L) + \sigma^2\mathbf{I}\right)^{-1}\mathbf{y}_L \to K(\mathbf{x}^*, \mathbf{X}_L)\left(K(\mathbf{X}_L, \mathbf{X}_L) + \sigma^2\mathbf{I}\right)^{-1}\mathbf{y}_L$.

Hence, $|\frac{\left(y^* - K(\mathbf{x}_j, \mathbf{X}_L)\left(K(\mathbf{X}_L, \mathbf{X}_L) + \sigma^2\mathbf{I}\right)^{-1}\mathbf{y}_L\right)^2}{\sigma^{*2}} - \frac{\left(y^* - K_{tran}\left(K(\mathbf{X}_L, \mathbf{X}_L) + \sigma^2\mathbf{I}\right)^{-1}\mathbf{y}_L\right)^2}{\sigma^{*}_{tran}{}^2}| \to 0$.

Finally $D_{KL}(f^*, f^*_{tran}) \to 0$. Thus $f^*_{tran} \xrightarrow{d} f^*$.

Appendix C: Proof of Theorem 3.1

According to derivation process from (3.11) to (3.15), $\inf_{\boldsymbol{\theta},\xi}\sup_{\boldsymbol{\mu}\geq 0,\boldsymbol{\nu}\geq 0}\mathcal{L}'$ can be simplified as

$$\inf_{\boldsymbol{\theta}} \frac{1}{L} l(\boldsymbol{\theta}) + \lambda_1\left(\sum_{i=L+1}^{L+U} Var(f_i)\right) + \lambda_2 \sum_{j=1}^{m} g_j(\cdot) I(g_j(\cdot) \geq 0)\,. \qquad \text{(C.1)}$$

Similarly, $\inf_{\boldsymbol{\theta},\xi} \sup_{\boldsymbol{\mu}\geq 0,\boldsymbol{\nu}\geq 0,\alpha_1\geq 0,\alpha_2\geq 0} \mathcal{L}$ can be simplified as

$$\inf_{\boldsymbol{\theta}} \sup_{\alpha_1\geq 0,\alpha_2\geq 0} \{\frac{1}{L} l(\boldsymbol{\theta}) + \alpha_1\left(\sum_{i=L+1}^{L+U} Var(f_i) - t\right) + \alpha_2(\sum_{j=1}^{m} g_j(\cdot) I\left(g_j(\cdot) \geq 0\right) - \epsilon)\}. \qquad \text{(C.2)}$$

To prove Theorem 3.1, it means to prove (C.1) and (C.2) are equivalent.

(a) For any choice of λ_1 and λ_2, consider the optimal solution $\boldsymbol{\theta}^*$ from (C.1). It is not hard to see that $\boldsymbol{\theta}^*$ will also be the optimal solution to (C.2) if $t = \frac{1}{U}\sum_{i=L+1}^{L+U} Var_{\boldsymbol{\theta}^*}(f_i)$, and $\epsilon = \sum_{j=1}^{m} \xi_j{}^*$; else if there is some other $\boldsymbol{\theta}'$ with $\frac{1}{U}\sum_{i=L+1}^{L+U} Var(f_i) \leq t$ and $\sum_{j=1}^{m} \xi_j \leq \epsilon$, but a better objective value than $\boldsymbol{\theta}^*$ (Note since the t and ϵ is pre-setting, it becomes a hard-constraint optimization. It's easy to know that $\alpha_1 = 0$; $\alpha_2 = 0$; and $\frac{1}{L} l(\boldsymbol{\theta}') \leq \frac{1}{L} l(\boldsymbol{\theta}^*)$). Then

$$\frac{1}{L} l(\boldsymbol{\theta}') + \lambda_1\left(\sum_{i=L+1}^{L+U} Var_{\boldsymbol{\theta}'}(f_i)\right) + \lambda_2 \sum_{j=1}^{m} g_{j,\boldsymbol{\theta}'}(\cdot) I\left(g_{j,\boldsymbol{\theta}'}(\cdot) \geq 0\right)$$

$$\leq \frac{1}{L} l(\boldsymbol{\theta}^*) + \lambda_1 t + \lambda_2 \epsilon$$

$$= \frac{1}{L} l(\boldsymbol{\theta}^*) + \lambda_1\left(\sum_{i=L+1}^{L+U} Var_{\boldsymbol{\theta}^*}(f_i)\right) + \lambda_2 \sum_{j=1}^{m} g_{j,\boldsymbol{\theta}^*}(\cdot) I(g_{j,\boldsymbol{\theta}^*}(\cdot) \geq 0).$$

This contradicts the optimality of $\boldsymbol{\theta}^*$ in (C.1). Hence $\boldsymbol{\theta}^*$ is also optimal in (C.2).

(b) Conversely, for any choice of t and ϵ, let $\boldsymbol{\theta}^*$ the optimal solution from (C.2), accompanied with the optimal α_1^* and α_2^*. Hence $\boldsymbol{\theta}^*$ is optimal in

$$\inf_{\boldsymbol{\theta}} \frac{1}{L} l(\boldsymbol{\theta}) + \alpha_1^* \left(\sum_{i=L+1}^{L+U} Var(f_i) - t \right) + \alpha_2^* \left(\sum_{j=1}^{m} g_j(\cdot) I(g_j(\cdot) \geq 0) - \epsilon \right).$$

Removing the constant term $\alpha_1^* t$ and $\alpha_2^* \epsilon$, and setting $\lambda_1 = \alpha_1^*$ and $\lambda_2 = \alpha_2^*$, we have that $\boldsymbol{\theta}^*$ is the optimal solution for (C.1). ∎

Appendix D: Proof of Theorem 3.2

Proof: Our proof aims to show that the optimization in (3.17) is equivalent to (3.3)- (3.7). For notation simplicity, define $\Omega_i^1(q) \triangleq \int_{f,\boldsymbol{\theta}} q \times \left(f(\mathbf{x}_i) - E_q[f(\mathbf{x}_i)]\right)^2 d\eta(f,\boldsymbol{\theta})$ and $\Omega_i^2(q) \triangleq \int_{f,\boldsymbol{\theta}} q \times f(\mathbf{x}_i)\, d\eta(f,\boldsymbol{\theta})$. Then the constraints in

(3.18) become $\frac{1}{U}\sum_{i=L+1}^{L+U} \Omega_i^1(q) \le t$ and $g_j\left(\Omega_{L+1}^2(q), \dots, \Omega_{L+U}^2(q)\right) \le \xi_j, j = 1, \dots, m.$ Using the Lagrange multiplier method, we know that (3.17) is equivalent to

$$\inf_{q\in\mathcal{P}_{prob}} \inf_{t,\boldsymbol{\xi}} \sup_{\alpha_1,\boldsymbol{\mu},\mathbf{v}\ge 0} \left\{ \begin{array}{c} KL\left(q\|p(f,\boldsymbol{\theta}|D)\right) + \lambda_1 t + \lambda_2\left(\sum_{j=1}^m \xi_j\right) + \alpha_1\left(\frac{1}{U}\sum_{i=L+1}^{L+U}\Omega_i^1(q) - t\right) + \\ \sum_{j=1}^m u_j\left(g_j\left(\Omega_{L+1}^2(q), \dots, \Omega_{L+U}^2(q)\right) - \xi_j\right) - \sum_{j=1}^m v_j\xi_j \end{array} \right\}. \tag{D.1}$$

Since (D.1) is a convex function of $\boldsymbol{\xi}, t, \alpha_1, \boldsymbol{\mu}, \boldsymbol{v}$, it is equivalent to

$$\inf_{q\in\mathcal{P}_{prob}} \sup_{\alpha_1,\boldsymbol{\mu},\mathbf{v}\ge 0} \inf_{t,\boldsymbol{\xi}} \left\{ \begin{array}{c} KL\left(q\|p(f,\boldsymbol{\theta}|D)\right) + \lambda_1 t + \lambda_2\left(\sum_{j=1}^m \xi_j\right) + \alpha_1\left(\frac{1}{U}\sum_{i=L+1}^{L+U}\Omega_i^1(q) - t\right) + \\ \sum_{j=1}^m u_j\left(g_j\left(\Omega_{L+1}^2(q), \dots, \Omega_{L+U}^2(q)\right) - \xi_j\right) - \sum_{j=1}^m v_j\xi_j \end{array} \right\}. \tag{D.2}$$

Denote the function within the { } in (D.2) by φ. Focus on solving the inner-most optimization with respect to $t, \boldsymbol{\xi}$ by equating the derivatives of φ to zeros, i.e.,

$$\frac{\partial\varphi}{\partial t} = \lambda_1 - \alpha_1 = 0,$$

$$\frac{\partial\varphi}{\partial \xi_j} = \lambda_2 - u_j - v_j = 0, j = 1, \dots, m.$$

From these equations we can get $\alpha_1 = \lambda_1$ and $v_j = \lambda_2 - u_j$. Putting these back to (D.2), we get

$$\inf_{q\in\mathcal{P}_{prob}} \sup_{\boldsymbol{\mu}\geq 0} \left\{ KL\big(q\|p(f,\boldsymbol{\theta}|D)\big) + \lambda_1 \left(\frac{1}{U}\sum_{i=L+1}^{L+U} \Omega_i^1(q)\right) + \sum_{j=1}^{m} u_j g_j\big(\Omega_{L+1}^2(q),\dots,\Omega_{L+U}^2(q)\big)\right\}$$

$$\text{s.t. } 0 \leq \mu_j \leq \lambda_2, j = 1,\dots,m.$$

That can be simplified as

$$\inf_{q\in\mathcal{P}_{prob}} \left\{ \begin{matrix} KL\big(q\|p(f,\boldsymbol{\theta}|D)\big) + \lambda_1 \left(\frac{1}{U}\sum_{i=L+1}^{L+U} \Omega_i^1(q)\right) + \\ \lambda_2 \sum_{j=1}^{m} g_j\left(\Omega_{L+1}^2(q),\dots,\Omega_{L+U}^2(q)\right) I\left(g_j\left(\Omega_{L+1}^2(q),\dots,\Omega_{L+U}^2(q)\right)>0\right) \end{matrix} \right\}. \quad \text{(D.3)}$$

Denote the function within { } in Eq. (D.3) by χ. Comparing (D.3) to that in Theorem 3.1, we know that the remaining task of this proof is to show that $\inf_{q\in\mathcal{P}_{prob}} \chi$ is equivalent to

$$\min_{\boldsymbol{\theta}} \left\{ \frac{1}{L} l(\boldsymbol{\theta}) + \lambda_1 \left(\frac{1}{U}\sum_{i=L+1}^{L+U} Var(f_i)\right) + \lambda_2 \sum_{j=1}^{m} g_j(\cdot)\, I(g_j(\cdot)>0)\right\}.$$

Next, we show steps to prove this equivalency.

$$\inf_{q\in\mathcal{P}_{prob}} \chi$$

$$= \inf_{q\in\mathcal{P}_{prob}} \left\{ \begin{matrix} \int_{f,\boldsymbol{\theta}} q log \frac{q}{p(f,\boldsymbol{\theta}|D)} d\eta(f,\boldsymbol{\theta}) + \lambda_1 \left(\frac{1}{U}\sum_{i=L+1}^{L+U} \Omega_i^1(q)\right) \\ +\lambda_2 \sum_{j=1}^{m} g_j\left(\Omega_{L+1}^2(q),\dots,\Omega_{L+U}^2(q)\right) I\left(g_j\left(\Omega_{L+1}^2(q),\dots,\Omega_{L+U}^2(q)\right)>0\right) \end{matrix} \right\}$$

$$= \inf_{q\in\mathcal{P}_{prob}} \left\{ \begin{matrix} \int_{f,\boldsymbol{\theta}} p(f|\boldsymbol{\theta},D)\delta_{\bar{\boldsymbol{\theta}}}(\boldsymbol{\theta}|D)\, log \frac{p(f|\boldsymbol{\theta},D)\delta_{\bar{\boldsymbol{\theta}}}(\boldsymbol{\theta}|D)}{p(f,\boldsymbol{\theta}|D)} d\eta(f,\boldsymbol{\theta}) + \\ \lambda_1 \left(\frac{1}{U}\sum_{i=L+1}^{L+U} \Omega_i^1(q)\right) + \\ \lambda_2 \sum_{j=1}^{m} g_j\left(\Omega_{L+1}^2(q),\dots,\Omega_{L+U}^2(q)\right) I\left(g_j\left(\Omega_{L+1}^2(q),\dots,\Omega_{L+U}^2(q)\right)>0\right) \end{matrix} \right\}. \quad \text{(D.4)}$$

Now focus on the third term within the inf { } in (D.4):

$$\Omega_i^2(q) = \int_{f,\boldsymbol{\theta}} p(f|\boldsymbol{\theta}, D)\delta_{\bar{\boldsymbol{\theta}}}(\boldsymbol{\theta}|D) f(\mathbf{x}_i) d\eta(f, \boldsymbol{\theta})$$

$$= \int_f f(\mathbf{x}_i) \int_{\boldsymbol{\theta}} p(f|\boldsymbol{\theta}, D)\delta_{\bar{\boldsymbol{\theta}}}(\boldsymbol{\theta}|D)\, d\eta(f, \boldsymbol{\theta}) = \int_f f(\mathbf{x}_i)\, p(f|\bar{\boldsymbol{\theta}}, \mathcal{D}) d\eta(f) = E_p[f(\mathbf{x}_i)],$$

which is not related to f or $\boldsymbol{\theta}$. Similarly,

$$\Omega_i^1(q) = \int_{f,\boldsymbol{\theta}} p(f|\boldsymbol{\theta}, \mathcal{D})\delta_{\bar{\boldsymbol{\theta}}}(\boldsymbol{\theta}|D) \times \left(f(\mathbf{x}_i) - E_p[f(\mathbf{x}_i)]\right)^2 d\eta(f, \boldsymbol{\theta})$$

$$= \int_f \left(f(\mathbf{x}_i) - E_p[f(\mathbf{x}_i)]\right)^2 \int_{\boldsymbol{\theta}} p(f|\boldsymbol{\theta}, D)\delta_{\bar{\boldsymbol{\theta}}}(\boldsymbol{\theta}|D)\, d\eta(f, \boldsymbol{\theta})$$

$$= \int_f \left(f(\mathbf{x}_i) - E_p[f(\mathbf{x}_i)]\right)^2 p(f|\bar{\boldsymbol{\theta}}, \mathcal{D}) d\eta(f) = Var_p(f_i).$$

Then, (D.4) becomes:

$$\inf_{\bar{\boldsymbol{\theta}}} \left\{ \begin{array}{c} \int_{f,\boldsymbol{\theta}} p(f|\boldsymbol{\theta}, \mathcal{D})\delta_{\bar{\boldsymbol{\theta}}}(\boldsymbol{\theta}|D)\, log\, \frac{\delta_{\bar{\boldsymbol{\theta}}}(\boldsymbol{\theta}|D)}{p(\boldsymbol{\theta}|D)} d\eta(f, \boldsymbol{\theta}) + \\ \lambda_1 \left(\frac{1}{U}\sum_{i=L+1}^{L+U} Var_p(f_i)\right) + \\ \lambda_2 \sum_{j=1}^{m} g_j(\cdot)\, I(g_j(\cdot) > 0) \end{array} \right\}$$

$$= \inf_{\bar{\boldsymbol{\theta}}} \left\{ \int_{\boldsymbol{\theta}} \delta_{\bar{\boldsymbol{\theta}}}(\boldsymbol{\theta}|D)\, log\, \frac{\delta_{\bar{\boldsymbol{\theta}}}(\boldsymbol{\theta}|D)}{p(\boldsymbol{\theta}|D)} d\eta(\boldsymbol{\theta}) + \lambda_1 \left(\frac{1}{U}\sum_{i=L+1}^{L+U} Var_p(f_i)\right) + \lambda_2 \sum_{j=1}^{m} g_j(\cdot)\, I(g_j(\cdot) > 0) \right\}$$

$$= \inf_{\bar{\boldsymbol{\theta}}} \left\{ - \int_{\boldsymbol{\theta}} \delta_{\bar{\boldsymbol{\theta}}}(\boldsymbol{\theta}|D)\, log\, p(\boldsymbol{\theta}|D)\, d\eta(\boldsymbol{\theta}) + \lambda_1 \left(\frac{1}{U}\sum_{i=L+1}^{L+U} Var_p(f_i)\right) + \lambda_2 \sum_{j=1}^{m} g_j(\cdot)\, I(g_j(\cdot) > 0) \right\}$$

$$= \inf_{\bar{\boldsymbol{\theta}}} \left\{ - \int_{\boldsymbol{\theta}} \delta_{\bar{\boldsymbol{\theta}}}(\boldsymbol{\theta}|D)\, log\, \frac{p(\mathbf{y}_L, \boldsymbol{\theta}|\mathbf{X}_L)}{p(\mathbf{y}_L|\mathbf{X}_L)} d\eta(\boldsymbol{\theta}) + \lambda_1 \left(\frac{1}{U}\sum_{i=L+1}^{L+U} Var_p(f_i)\right) + \lambda_2 \sum_{j=1}^{m} g_j(\cdot)\, I(g_j(\cdot) > 0) \right\}$$

$$= \inf_{\bar{\boldsymbol{\theta}}} \left\{ - \int_{\boldsymbol{\theta}} \delta_{\bar{\boldsymbol{\theta}}}(\boldsymbol{\theta}|D)\, log\, p(\mathbf{y}_L, \boldsymbol{\theta}|\mathbf{X}_L)\, d\eta(\boldsymbol{\theta}) + \lambda_1 \left(\frac{1}{U}\sum_{i=L+1}^{L+U} Var_p(f_i)\right) + \lambda_2 \sum_{j=1}^{m} g_j(\cdot)\, I(g_j(\cdot) > 0) \right\}$$

$$= \inf_{\bar{\boldsymbol{\theta}}} \left\{ -\log p(\mathbf{y}_L, \bar{\boldsymbol{\theta}} | \mathbf{X}_L) + \lambda_1 \left(\frac{1}{U} \sum_{i=L+1}^{L+U} Var_p(f_i) \right) + \lambda_2 \sum_{j=1}^{m} g_j(\cdot)\, I(g_j(\cdot) > 0) \right\}$$

$$= \inf_{\bar{\boldsymbol{\theta}}} \left\{ -\log p(\mathbf{y}_L | \mathbf{X}_L, \bar{\boldsymbol{\theta}}) p(\bar{\boldsymbol{\theta}} | \mathbf{X}_L) + \lambda_1 \left(\frac{1}{U} \sum_{i=L+1}^{L+U} Var_p(f_i) \right) + \lambda_2 \sum_{j=1}^{m} g_j(\cdot)\, I(g_j(\cdot) > 0) \right\}$$

$$= \inf_{\bar{\boldsymbol{\theta}}} \left\{ -\log p(\mathbf{y}_L | \mathbf{X}_L, \bar{\boldsymbol{\theta}}) - \log p(\bar{\boldsymbol{\theta}} | \mathbf{X}_L) + \lambda_1 \left(\frac{1}{U} \sum_{i=L+1}^{L+U} Var_p(f_i) \right) + \lambda_2 \sum_{j=1}^{m} g_j(\cdot)\, I(g_j(\cdot) > 0) \right\}$$

$$= \inf_{\bar{\boldsymbol{\theta}}} \left\{ -\log p(\mathbf{y}_L | \mathbf{X}_L, \bar{\boldsymbol{\theta}}) + \lambda_1 \left(\frac{1}{U} \sum_{i=L+1}^{L+U} Var_p(f_i) \right) + \lambda_2 \sum_{j=1}^{m} g_j(\cdot)\, I(g_j(\cdot) > 0) \right\}$$

$$= \inf_{\bar{\boldsymbol{\theta}}} \left\{ -\frac{1}{L} \log p(\mathbf{y}_L | \mathbf{X}_L, \bar{\boldsymbol{\theta}}) + \frac{\lambda_1}{L} \left(\frac{1}{U} \sum_{i=L+1}^{L+U} Var_p(f_i) \right) + \frac{\lambda_2}{L} \sum_{j=1}^{m} g_j(\cdot)\, I(g_j(\cdot) > 0) \right\}$$

$$= \inf_{\bar{\boldsymbol{\theta}}} \left\{ -\frac{1}{L} \log p(\mathbf{y}_L | \mathbf{X}_L, \bar{\boldsymbol{\theta}}) + \lambda_1 \left(\frac{1}{U} \sum_{i=L+1}^{L+U} Var_p(f_i) \right) + \lambda_2 \sum_{j=1}^{m} g_j(\cdot)\, I(g_j(\cdot) > 0) \right\}$$

Appendix E: Proof of Proposition 4.1

Proof: Let $\alpha_i^{(1)},\ \alpha_i^{(2)}, \beta_j^{(12)}, \beta_k^{(3)}, A_i^{(1)}, A_i^{(2)}, B_j^{(12)}, B_k^{(3)}, \mu \geq 0$ be Lagrangian multipliers and C_1 and C_2 be tuning parameters. The Lagrangian for the primal WSO-SVM optimization in (4.1) is

$$L = \frac{1}{2}w^T w - \sum_{i=1}^{n_1} \alpha_i^{(1)} \left(w^T \phi\left(x_i^{(1)}\right) + b_1 - 1 + \xi_i^{(1)}\right) + \quad \sum_{i=1}^{n_2} \alpha_i^{(2)} \left(w^T \phi\left(x_i^{(2)}\right) + b_1 + 1 - \xi_i^{(2)}\right) - \sum_{j=1}^{m_{12}} \beta_j^{(12)} \left(w^T \phi\left(x_j^{(12)}\right) + b_2 - 1 + \zeta_j^{(12)}\right) + \sum_{k=1}^{m_3} \beta_k^{(3)} (w^T \phi\left(x_k^{(3)}\right) + b_2 + 1 - \zeta_k^{(3)}) + C_1 (\sum_{i=1}^{n_1} \xi_i^{(1)} + \sum_{i=1}^{n_2} \xi_i^{(2)}) + C_2 \left(\sum_{j=1}^{m_{12}} \zeta_j^{(12)} + \sum_{k=1}^{m_3} \zeta_k^{(3)}\right) - \sum_{i=1}^{n_1} A_i^{(1)} \xi_i^{(1)} - \sum_{i=1}^{n_2} A_i^{(2)} \xi_i^{(2)} - \sum_{j=1}^{m_{12}} B_j^{(12)} \zeta_j^{(12)} - \sum_{k=1}^{m_3} B_k^{(3)} \zeta_k^{(3)} + \mu(b_1 - b_2).$$

Then the optimal solution of the primal problem in (4.1) is equivalent to the solution of the following optimization:

$$\max_{\alpha,\beta,A,B,\mu} min_{w,b,\xi,\zeta} L. \tag{E.1}$$

The KKT conditions for the primal problem require the following to hold:

$$\nabla_w L = w - \sum_{i=1}^{n_1} \alpha_i^{(1)} \phi\left(x_i^{(1)}\right) + \sum_{i=1}^{n_2} \alpha_i^{(2)} \phi\left(x_i^{(2)}\right) - \sum_{j=1}^{m_{12}} \beta_j^{(12)} \phi\left(x_j^{(12)}\right) + \sum_{k=1}^{m_3} \beta_k^{(3)} \phi\left(x_k^{(3)}\right) = 0,$$

$$\nabla_{b_1} L = -\sum_{i=1}^{n_1} \alpha_i^{(1)} + \sum_{i=1}^{n_2} \alpha_i^{(2)} + \mu = 0,$$

$$\nabla_{b_2} L = -\sum_{j=1}^{m_{12}} \beta_j^{(12)} + \sum_{k=1}^{m_3} \beta_k^{(3)} - \mu = 0,$$

$$\nabla_{\xi_i^{(1)}} L = -\alpha_i^{(1)} + C_1 - A_i^{(1)} = 0,\ i = 1, \ldots, n_1,$$

$$\nabla_{\xi_i^{(2)}} L = -\alpha_i^{(2)} + C_1 - A_i^{(2)} = 0, i = 1, \ldots, n_2,$$

$$\nabla_{\zeta_j^{(12)}} L = -\beta_j^{(12)} + C_2 - B_j^{(12)} = 0,\ j = 1, \ldots, m_{12},$$

$$\nabla_{\zeta_k^{(3)}} L = -\beta_k^{(3)} + C_2 - B_k^{(3)} = 0, k = 1, \ldots, m_3.$$

Then we have

$$w = \sum_{i=1}^{n_1} \alpha_i^{(1)} \phi\left(x_i^{(1)}\right) - \sum_{i=1}^{n_2} \alpha_i^{(2)} \phi\left(x_i^{(2)}\right) + \sum_{j=1}^{m_{12}} \beta_j^{(12)} \phi\left(x_j^{(12)}\right) - \sum_{k=1}^{m_3} \beta_k^{(3)} \phi\left(x_k^{(3)}\right), \tag{E.2}$$

$$\mu = \sum_{i=1}^{n_1} \alpha_i^{(1)} - \sum_{i=1}^{n_2} \alpha_i^{(2)}, \tag{E.3}$$

$$\mu = -\sum_{j=1}^{m_{12}} \beta_j^{(12)} + \sum_{k=1}^{m_3} \beta_k^{(3)}, \tag{E.4}$$

$$A_i^{(1)} = -\alpha_i^{(1)} + C_1,\ i = 1, \ldots, n_1, \tag{E.5}$$

$$A_i^{(2)} = -\alpha_i^{(2)} + C_1,\ i = 1, \ldots, n_2, \tag{E.6}$$

$$B_j^{(12)} = -\beta_j^{(12)} + C_2,\ j = 1, \ldots, m_{12}, \tag{E.7}$$

$$B_k^{(3)} = -\beta_k^{(3)} + C_2,\ k = 1, \ldots, m_3. \tag{E.8}$$

Inserting (E.3), (E.5)-(E.8) into the optimization in (E.1), after simplification we can get

$$\max_{\alpha,\beta} L = \frac{1}{2} w^T w - \sum_{i=1}^{n_1} \alpha_i^{(1)} \left(w^T \phi\left(x_i^{(1)}\right) - 1\right) + \sum_{i=1}^{n_2} \alpha_i^{(2)} \left(w^T \phi\left(x_i^{(2)}\right) + 1\right) -$$

$$\sum_{j=1}^{m_{12}} \beta_j^{(12)} \left(w^T \phi\left(x_j^{(12)}\right) - 1\right) + \sum_{k=1}^{m_3} \beta_k^{(3)} \left(w^T \phi\left(x_k^{(3)}\right) + 1\right). \tag{E.9}$$

Furthermore, inserting (E.2) into the optimization in (E.9), we can have

$$\max_{\alpha,\beta} L = -\frac{1}{2}\gamma^T YKY\gamma + \sum_{i=1}^{n_1}\alpha_i^{(1)} + \sum_{i=1}^{n_2}\alpha_i^2 + \sum_{j=1}^{m_{12}}\beta_j^{(12)} + \sum_{k=1}^{m_3}\beta_k^{(3)}.$$

Additionally, the conditions in (E.3)- (E.4) give rise to the constraints of

$$\sum_{i=1}^{n_1}\alpha_i^{(1)} - \sum_{i=1}^{n_2}\alpha_i^{(2)} + \sum_{j=1}^{m_{12}}\beta_j^{(12)} - \sum_{k=1}^{m_3}\beta_k^{(3)} = 0,$$

$$\sum_{i=1}^{n_1}\alpha_i^{(1)} - \sum_{i=1}^{n_2}\alpha_i^{(2)} \geq 0.$$

The conditions in (E.5)-(E.8) give rise to the constraints of

$$0 \leq \alpha_i^{(1)} \leq C_1, i = 1,\ldots,n_1; 0 \leq \alpha_i^{(2)} \leq C_1, i = 1,\ldots,n_2,$$

$$0 \leq \beta_j^{(12)} \leq C_2, j = 1,\ldots,m_{12};\ 0 \leq \beta_k^{(3)} \leq C_2, k = 1,\ldots,m_3.$$

Finally, the dual problem becomes

$$\min_{\alpha,\beta} \frac{1}{2}\gamma^T YKY\gamma - \sum_{i=1}^{n_1}\alpha_i^{(1)} - \sum_{i=1}^{n_2}\alpha_i^{(2)} - \sum_{j=1}^{m_{12}}\beta_j^{(12)} - \sum_{j=1}^{m_3}\beta_k^{(3)},$$

subject to

$$\sum_{i=1}^{n_1}\alpha_i^{(1)} - \sum_{i=1}^{n_2}\alpha_i^{(2)} + \sum_{j=1}^{m_{12}}\beta_j^{(12)} - \sum_{k=1}^{m_3}\beta_k^{(3)} = 0,$$

$$\sum_{i=1}^{n_1}\alpha_i^{(1)} - \sum_{i=1}^{n_2}\alpha_i^{(2)} \geq 0,$$

$$0 \leq \alpha_i^{(1)} \leq C_1, i = 1,\ldots,n_1; 0 \leq \alpha_i^{(2)} \leq C_1, i = 1,\ldots,n_2,$$

$$0 \leq \beta_j^{(12)} \leq C_2, j = 1,\ldots,m_{12};\ 0 \leq \beta_k^{(3)} \leq C_2, k = 1,\ldots,m_3.$$

■

REFERENCES

[1] J. C. HELTON and D. E. BURMASTER, “Treatment of aleatory and epistemic uncertainty,” *Reliability engineering & systems safety*, vol. 54, no. 2–3, 1996.

[2] W. L. Oberkampf, “Uncertainty quantification using evidence theory,” 2005.

[3] W. L. Oberkampf and J. C. Helton, “Evidence theory for engineering applications,” in *Engineering design reliability handbook*, CRC Press, 2004, pp. 197–226.

[4] W. L. Oberkampf, S. M. DeLand, B. M. Rutherford, K. V Diegert, and K. F. Alvin, “Error and uncertainty in modeling and simulation,” *Reliability Engineering & System Safety*, vol. 75, no. 3, pp. 333–357, 2002.

[5] J. Liu, J. Paisley, M.-A. Kioumourtzoglou, and B. Coull, “Accurate Uncertainty Estimation and Decomposition in Ensemble Learning,” in *Advances in Neural Information Processing Systems*, 2019, pp. 8950–8961.

[6] G. Ramachandran, *Assessing nanoparticle risks to human health*. William Andrew, 2016.

[7] M. C. Kennedy and A. O’Hagan, “Bayesian calibration of computer models,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 3, pp. 425–464, 2001.

[8] S. De Vleeschouwer, “Epidemiology and Outcome of Glioblastoma--Glioblastoma,” 2017.

[9] "Auspicious machine learning," *Nature Biomedical Engineering*. 2017. doi: 10.1038/s41551-017-0036.

[10] C. G. A. R. Network, "Integrated genomic analyses of ovarian carcinoma," *Nature*, vol. 474, no. 7353, p. 609, 2011.

[11] C. G. A. R. Network, "Comprehensive genomic characterization of squamous cell lung cancers," *Nature*, vol. 489, no. 7417, p. 519, 2012.

[12] D. M. Muzny *et al.*, "Comprehensive molecular characterization of human colon and rectal cancer," *Nature*, 2012, doi: 10.1038/nature11252.

[13] C. G. A. Network, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, p. 61, 2012.

[14] 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.

[15] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[16] D. L. Shrestha and D. P. Solomatine, "Machine learning approaches for estimation of prediction interval for the model output," *Neural Networks*, vol. 19, no. 2, pp. 225–235, 2006.

[17] J. Quinonero-Candela, C. E. Rasmussen, F. Sinz, O. Bousquet, and B. Schölkopf, "Evaluating predictive uncertainty challenge," in *Machine Learning Challenges Workshop*, 2005, pp. 1–27.

[18] D. Beck, L. Specia, and T. Cohn, “Exploring prediction uncertainty in machine translation quality estimation,” *arXiv preprint arXiv:1606.09600*, 2016.

[19] D. P. Solomatine and D. L. Shrestha, “A novel method to estimate model uncertainty using machine learning techniques,” *Water Resources Research*, vol. 45, no. 12, 2009.

[20] S. Wang, L. Zhang, and R. Urtasun, “Transductive Gaussian processes for image denoising,” in *2014 IEEE international conference on computational photography (ICCP)*, 2014, pp. 1–8.

[21] L. S. Hu *et al.*, “Radiogenomics to characterize regional genetic heterogeneity in glioblastoma.,” *Neuro Oncol*, vol. 19, no. 1, pp. 128–137, Jan. 2017, doi: 10.1093/neuonc/now135.

[22] L. S. Hu *et al.*, “Multi-Parametric MRI and Texture Analysis to Visualize Spatial Histologic Heterogeneity and Tumor Extent in Glioblastoma.,” *PLoS One*, vol. 10, no. 11, p. e0141506, 2015.

[23] L. S. Hu *et al.*, “Radiogenomics to characterize regional genetic heterogeneity in glioblastoma.,” *Neuro Oncol*, vol. 19, no. 1, pp. 128–137, Jan. 2017, doi: 10.1093/neuonc/now135.

[24] L. S. Hu *et al.*, “Multi-parametric MRI and texture analysis to visualize spatial histologic heterogeneity and tumor extent in glioblastoma,” *PLoS ONE*, 2015, doi: 10.1371/journal.pone.0141506.

[25] M. J. Borad *et al.*, “Integrated genomic characterization reveals novel, therapeutically relevant drug targets in FGFR and EGFR pathways in sporadic intrahepatic cholangiocarcinoma,” *PLoS Genet*, vol. 10, no. 2, 2014.

[26] D. W. Craig *et al.*, “Genome and transcriptome sequencing in prospective metastatic triple-negative breast cancer uncovers therapeutic vulnerabilities,” *Mol Cancer Ther*, vol. 12, no. 1, pp. 104–116, 2013.

[27] D. Lipson, Y. Aumann, A. Ben-Dor, N. Linial, and Z. Yakhini, “Efficient calculation of interval scores for DNA copy number data analysis,” *Journal of computational biology*, vol. 13, no. 2, pp. 215–228, 2006.

[28] C. W. Brennan *et al.*, “The somatic genomic landscape of glioblastoma,” *Cell*, vol. 155, no. 2, pp. 462–477, 2013.

[29] A. Sottoriva *et al.*, “Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 10, pp. 4009–4014, 2013.

[30] R.-Y. Bai, V. Staedtke, and G. J. Riggins, “Molecular targeting of glioblastoma: drug discovery and therapies,” *Trends Mol Med*, vol. 17, no. 6, pp. 301–312, 2011.

[31] L. S. Hu *et al.*, “Reevaluating the imaging definition of tumor progression: perfusion MRI quantifies recurrent glioblastoma tumor fraction, pseudoprogression, and radiation necrosis to predict survival.,” *Neuro Oncol*, vol. 14, no. 7, pp. 919–30, Jul. 2012, doi: 10.1093/neuonc/nos112.

[32] S. J. Price *et al.*, "Improved delineation of glioma margins and regions of infiltration with the use of diffusion tensor imaging: an image-guided biopsy study.," *AJNR Am J Neuroradiol*, vol. 27, no. 9, pp. 1969–74, Oct. 2006.

[33] N. B. Semmineh *et al.*, "Assessing tumor cytoarchitecture using multiecho DSC-MRI derived measures of the transverse relaxivity at tracer equilibrium (TRATE)," *Magnetic Resonance in Medicine*, vol. 74, no. 3, pp. 772–784, 2015, doi: 10.1002/mrm.25435.

[34] L. S. Hu *et al.*, "Impact of Software Modeling on the Accuracy of Perfusion MRI in Glioma.," *AJNR Am J Neuroradiol*, vol. 36, no. 12, pp. 2242–9, Dec. 2015, doi: 10.3174/ajnr.A4451.

[35] J. L. Boxerman, K. M. Schmainda, and R. M. Weisskoff, "Relative cerebral blood volume maps corrected for contrast agent extravasation significantly correlate with glioma tumor grade, whereas uncorrected maps do not," *American Journal of Neuroradiology*, vol. 27, no. 4, pp. 859–867, 2006.

[36] A. Zwanenburg, S. Leger, M. Vallières, and S. Löck, "Image biomarker standardisation initiative," *arXiv preprint arXiv:1612.07003*, 2016.

[37] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973, doi: 10.1109/TSMC.1973.4309314.

[38] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[39] A. G. Ramakrishnan, S. K. Raja, and H. V. R. Ram, “Neural network-based segmentation of textures using Gabor features,” in *Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing*, 2002, pp. 365–374.

[40] J. M. Furgason *et al.*, “Whole genome sequencing of glioblastoma multiforme identifies multiple structural variations involved in EGFR activation,” *Mutagenesis*, vol. 29, no. 5, pp. 341–350, 2014.

[41] N. E. Wineinger, R. E. Kennedy, S. W. Erickson, M. K. Wojczynski, C. E. Bruder, and H. K. Tiwari, “Statistical issues in the analysis of DNA copy number variations,” *Int J Comput Biol Drug Des*, vol. 1, no. 4, p. 368, 2008.

[42] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li, *Applied linear statistical models*, vol. 5. McGraw-Hill Irwin New York, 2005.

[43] I. Yeo and R. A. Johnson, “A new family of power transformations to improve normality or symmetry,” *Biometrika*, vol. 87, no. 4, pp. 954–959, 2000.

[44] M. B. Kursa and W. R. Rudnicki, “Feature selection with the Boruta package,” *J Stat Softw*, vol. 36, no. 11, pp. 1–13, 2010.

[45] L. S. Hu *et al.*, “Radiogenomics to characterize regional genetic heterogeneity in glioblastoma,” *Neuro Oncol*, vol. 19, no. 1, pp. 128–137, 2016.

[46] M. Long, J. Wang, G. Ding, S. J. Pan, and P. S. Yu, “Adaptation regularization: A general framework for transfer learning,” *IEEE Transactions on Knowledge and Data Engineering*, 2014, doi: 10.1109/TKDE.2013.111.

[47] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, “Transfer feature learning with joint distribution adaptation,” in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2200–2207.

[48] C. C. Aggarwal, *Data classification: algorithms and applications*. CRC press, 2014.

[49] K. Weiss, T. M. Khoshgoftaar, and D. Wang, “A survey of transfer learning,” *Journal of Big data*, vol. 3, no. 1, p. 9, 2016.

[50] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, “Domain adaptation via transfer component analysis,” *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2010.

[51] O. Day and T. M. Khoshgoftaar, “A survey on heterogeneous transfer learning,” *Journal of Big Data*, vol. 4, no. 1, p. 29, 2017.

[52] W. M. Kouw and M. Loog, “An introduction to domain adaptation and transfer learning,” *arXiv preprint arXiv:1812.11806*, 2018.

[53] B. Tran, M. Karimzadehgan, R. K. Pasumarthi, M. Bendersky, and D. Metzler, “Domain Adaptation for Enterprise Email Search,” in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 25–34.

[54] N. R. Draper and H. Smith, *Applied regression analysis*, vol. 326. John Wiley & Sons, 1998.

[55] N. Gaw *et al.*, "Integration of machine learning and mechanistic models accurately predicts variation in cell density of glioblastoma using multiparametric MRI," *Sci Rep*, vol. 9, no. 1, pp. 1–9, 2019.

[56] L. S. Hu *et al.*, "Multi-parametric MRI and texture analysis to visualize spatial histologic heterogeneity and tumor extent in glioblastoma," *PLoS One*, vol. 10, no. 11, p. e0141506, 2015.

[57] H. L. P. Harpold, E. C. Alvord Jr, and K. R. Swanson, "The evolution of mathematical modeling of glioma proliferation and invasion," *Journal of Neuropathology & Experimental Neurology*, vol. 66, no. 1, pp. 1–9, 2007.

[58] P. R. Jackson, J. Juliano, A. Hawkins-Daarud, R. C. Rockne, and K. R. Swanson, "Patient-specific mathematical neuro-oncology: using a simple proliferation and invasion tumor model to inform clinical practice," *Bull Math Biol*, vol. 77, no. 5, pp. 846–856, 2015.

[59] K. R. Swanson, R. Rostomily, and E. C. Alvord Jr, "Predicting survival of patients with glioblastoma by combining a mathematical model and pre-operative MR imaging characteristics: A proof of principle," *British Journal of Cancer*, vol. 98, pp. 113–119, 2008.

[60] U. Garczarek, "Classification rules in standardized partition spaces." Universität Dortmund, 2002.

[61] Y. Q. Li and M. Tian, “A semi-supervised regression algorithm based on co-training with SVR-KNN,” in *Advanced Materials Research*, 2014, vol. 926, pp. 2914–2918.

[62] Y. Yi *et al.*, “Semi-supervised ridge regression with adaptive graph-based label propagation,” *Applied Sciences*, vol. 8, no. 12, p. 2636, 2018.

[63] J. Levatić, M. Ceci, T. Stepišnik, S. Džeroski, and D. Kocev, “Semi-supervised regression trees with application to QSAR modelling,” *Expert Systems with Applications*, vol. 158, p. 113569, 2020.

[64] Y.-F. Li, H.-W. Zha, and Z.-H. Zhou, “Learning safe prediction for semi-supervised regression,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, vol. 31, no. 1.

[65] R. Stupp *et al.*, “Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma,” *New England journal of medicine*, vol. 352, no. 10, pp. 987–996, 2005.

[66] L. S. Hu, A. Hawkins-Daarud, L. Wang, J. Li, and K. R. Swanson, “Imaging of intratumoral heterogeneity in high-grade glioma,” *Cancer Lett*, vol. 477, pp. 97–106, 2020.

[67] C. I. Ene and H. A. Fine, “Many tumors in one: a daunting therapeutic prospect,” *Cancer Cell*, vol. 20, no. 6, pp. 695–697, 2011.

[68] K. R. Swanson, C. Bridge, J. D. Murray, and E. C. Alvord Jr, “Virtual and real brain tumors: using mathematical modeling to quantify glioma growth and invasion,” *J Neurol Sci*, vol. 216, no. 1, pp. 1–10, 2003.

[69] K. R. Swanson, E. C. Alvord, and J. D. Murray, “Virtual brain tumours (gliomas) enhance the reality of medical imaging and highlight inadequacies of current therapy,” *Br J Cancer*, vol. 86, no. 1, pp. 14–18, 2002.

[70] A. L. Baldock *et al.*, “Patient-specific metrics of invasiveness reveal significant prognostic benefit of resection in a predictable subset of gliomas,” *PLoS One*, vol. 9, no. 10, p. e99057, 2014.

[71] A. Marusyk, V. Almendro, and K. Polyak, “Intra-tumour heterogeneity: a looking glass for cancer?,” *Nature Reviews Cancer*, vol. 12, no. 5, pp. 323–334, 2012.

[72] R.-Y. Bai, V. Staedtke, and G. J. Riggins, “Molecular targeting of glioblastoma: drug discovery and therapies,” *Trends Mol Med*, vol. 17, no. 6, pp. 301–312, 2011.

[73] M. Snuderl *et al.*, “Mosaic amplification of multiple receptor tyrosine kinase genes in glioblastoma,” *Cancer Cell*, vol. 20, no. 6, pp. 810–817, 2011.

[74] M. L. Sos *et al.*, “PTEN loss contributes to erlotinib resistance in EGFR-mutant lung cancer by activation of Akt and EGFR,” *Cancer Res*, vol. 69, no. 8, pp. 3256–3261, 2009.

[75] S. Shostak, *Cancer Stem Cells: The Cutting Edge*. BoD–Books on Demand, 2011.

[76] P. D. Brown *et al.*, "Phase I/II trial of erlotinib and temozolomide with radiation therapy in the treatment of newly diagnosed glioblastoma multiforme: North Central Cancer Treatment Group Study N0177," *Journal of Clinical Oncology*, vol. 26, no. 34, p. 5603, 2008.

[77] D. Chow, P. Chang, B. D. Weinberg, D. A. Bota, J. Grinband, and C. G. Filippi, "Imaging Genetic Heterogeneity in Glioblastoma and Other Glial Tumors: Review of Current Methods and Future Directions.," *AJR Am J Roentgenol*, vol. 210, no. 1, pp. 30–38, 2018.

[78] L. S. Hu *et al.*, "Radiogenomics to characterize regional genetic heterogeneity in glioblastoma," *Neuro Oncol*, vol. 19, no. 1, pp. 128–137, 2017.

[79] L. S. Hu *et al.*, "Uncertainty quantification in the radiogenomics modeling of EGFR amplification in glioblastoma," *Sci Rep*, vol. 11, no. 1, pp. 1–14, 2021.

[80] H. Akbari *et al.*, "In vivo evaluation of EGFRvIII mutation in primary glioblastoma patients via complex multiparametric MRI signature," *Neuro Oncol*, vol. 20, no. 8, pp. 1068–1079, 2018.

[81] E. S. Tykocinski *et al.*, "Use of magnetic perfusion-weighted imaging to determine epidermal growth factor receptor variant III expression in glioblastoma," *Neuro Oncol*, vol. 14, no. 5, pp. 613–623, 2012.

[82] P. Kickingereder *et al.*, "Radiogenomics of glioblastoma: machine learning–based classification of molecular characteristics by using multiparametric and multiregional MR imaging features," *Radiology*, vol. 281, no. 3, pp. 907–918, 2016.

[83] H. Chen *et al.*, “Deep Learning Radiomics to Predict PTEN Mutation Status From Magnetic Resonance Imaging in Patients With Glioma,” *Front Oncol*, vol. 11, 2021.

[84] B. Zhang *et al.*, “Multimodal MRI features predict isocitrate dehydrogenase genotype in high-grade gliomas,” *Neuro Oncol*, vol. 19, no. 1, pp. 109–117, 2017.

[85] X. Zhang *et al.*, “IDH mutation assessment of glioma using texture features of multimodal MR images,” in *Medical Imaging 2017: Computer-Aided Diagnosis*, 2017, vol. 10134, p. 101341S.

[86] V. G. Kanas, E. I. Zacharaki, G. A. Thomas, P. O. Zinn, V. Megalooikonomou, and R. R. Colen, “Learning MRI-based classification models for MGMT methylation status prediction in glioblastoma,” *Comput Methods Programs Biomed*, vol. 140, pp. 249–257, 2017.

[87] S. E. Combs *et al.*, “Prognostic significance of IDH-1 and MGMT in patients with glioblastoma: one step forward, and one step back?,” *Radiation oncology*, vol. 6, no. 1, pp. 1–5, 2011.

[88] G. Singh *et al.*, “Radiomics and radiogenomics in gliomas: a contemporary update,” *British Journal of Cancer*, vol. 125, no. 5, pp. 641–657, 2021.

[89] A. P. Becker, B. E. Sells, S. J. Haque, and A. Chakravarti, “Tumor heterogeneity in glioblastomas: from light microscopy to molecular pathology,” *Cancers (Basel)*, vol. 13, no. 4, p. 761, 2021.

[90] X. Zhu and A. B. Goldberg, “Introduction to semi-supervised learning,” *Synthesis lectures on artificial intelligence and machine learning*, vol. 3, no. 1, pp. 1–130, 2009.

[91] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, “Semi-supervised learning with deep generative models,” *Adv Neural Inf Process Syst*, vol. 27, 2014.

[92] O. Chapelle and A. Zien, “Semi-supervised classification by low density separation,” in *International workshop on artificial intelligence and statistics*, 2005, pp. 57–64.

[93] Z.-H. Zhou and M. Li, “Semi-supervised learning by disagreement,” *Knowledge and Information Systems*, vol. 24, no. 3, pp. 415–439, 2010.

[94] Y. Chong, Y. Ding, Q. Yan, and S. Pan, “Graph-based semi-supervised learning: A review,” *Neurocomputing*, vol. 408, pp. 216–230, 2020.

[95] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Proceedings of the 31st international conference on neural information processing systems*, 2017, pp. 4768–4777.

[96] K. H. Ng and A. H. Pooi, “Calibration intervals in linear regression models,” *Communications in Statistics-Theory and Methods*, vol. 37, no. 11, pp. 1688–1696, 2008.

[97] U. Garczarek, “Classification rules in standardized partition spaces,” 2002.

[98] P. N. Bennett, “Using asymmetric distributions to improve text classifier probability estimates,” in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003, pp. 111–118.

[99] J. Platt, “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[100] M. Kull, T. Silva Filho, and P. Flach, “Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers,” in *Artificial Intelligence and Statistics*, 2017, pp. 623–631.

[101] J. Platt, “Sequential minimal optimization: A fast algorithm for training support vector machines,” 1998.

[102] D. Koul, “PTEN signaling pathways in glioblastoma,” *Cancer Biol Ther*, vol. 7, no. 9, pp. 1321–1325, 2008.

[103] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, “Textural features for image classification,” *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610–621, 1973.

[104] H. G. Feichtinger and T. Strohmer, *Gabor analysis and algorithms: Theory and applications*. Springer Science & Business Media, 2012.

[105] R. Collobert, F. Sinz, J. Weston, L. Bottou, and T. Joachims, “Large scale transductive SVMs.,” *Journal of Machine Learning Research*, vol. 7, no. 8, 2006.

[106] M. Belkin, P. Niyogi, and V. Sindhwani, “Manifold regularization: A geometric framework for learning from labeled and unlabeled examples.,” *Journal of machine learning research*, vol. 7, no. 11, 2006.

[107] Y. Zhou and S. Goldman, “Democratic co-learning,” in *16th IEEE International Conference on Tools with Artificial Intelligence*, 2004, pp. 594–602.

[108] H. Cao, J. Zhou, and E. Schwarz, “RMTL: an R library for multi-task learning,” *Bioinformatics*, vol. 35, no. 10, pp. 1797–1798, 2019.

[109] C. Williams, E. v Bonilla, and K. M. Chai, “Multi-task Gaussian process prediction,” *Adv Neural Inf Process Syst*, pp. 153–160, 2007.

[110] M. Aghi, P. Gaviani, J. W. Henson, T. T. Batchelor, D. N. Louis, and F. G. Barker, “Magnetic resonance imaging characteristics predict epidermal growth factor receptor amplification status in glioblastoma,” *Clinical Cancer Research*, vol. 11, no. 24, pp. 8600–8605, 2005.

[111] E. S. Tykocinski *et al.*, “Use of magnetic perfusion-weighted imaging to determine epidermal growth factor receptor variant III expression in glioblastoma,” *Neuro Oncol*, vol. 14, no. 5, pp. 613–623, 2012.

[112] A. Gupta *et al.*, “Pretreatment dynamic susceptibility contrast MRI perfusion in glioblastoma: prediction of EGFR gene amplification,” *Clin Neuroradiol*, vol. 25, no. 2, pp. 143–150, 2015.

[113] I. Ryoo *et al.*, "Cerebral blood volume calculated by dynamic susceptibility contrast-enhanced perfusion MR imaging: preliminary correlation study with glioblastoma genetic profiles," *PLoS One*, vol. 8, no. 8, p. e71704, 2013.