Knowledge-Infused Global-Local Data Fusion for Spatial Predictive Modeling in Precision Medicine

Lujia Wang, *Member, IEEE*, Andrea Hawkins-Daarud, Kristin R. Swanson, Leland S. Hu, and Jing Li, *Member, IEEE*

***Abstract*— The automated capability of generating spatial prediction for a variable of interest is desirable in various science and engineering domains. Take precision medicine of cancer as an example, in which the goal is to match patients with treatments based on molecular markers identified in each patient's tumor. A substantial challenge, however, is that the molecular markers can vary significantly at different spatial locations of a tumor. If this spatial distribution could be predicted, the precision of cancer treatment could be greatly improved by adapting treatment to the spatial molecular heterogeneity. This is a challenging task because no technology is available to measure the molecular markers at each spatial location within a tumor. Biopsy samples provide direct measurement, but they are scarce/local. Imaging, such as MRI, is global, but it only provides proxy/indirect measurement. Also available are mechanistic models or domain knowledge, which are often approximate or incomplete. This article proposes a novel machine learning framework to fuse the three sources of data/information to generate a spatial prediction, namely, the knowledge-infused global-local (KGL) data fusion model. A novel mathematical formulation is proposed and solved with theoretical study. 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 six MRI contrast images from each patient and biological knowledge encoded by a PDE simulator-based mechanistic model called proliferation-invasion (PI). 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.**

***Note to Practitioners*—This article proposes a machine learning framework to fuse local data, global imaging, and domain knowledge to generate a spatial prediction for a variable of interest. This methodology is relevant to multiple application domains. In precision medicine, it will allow for mapping the spatial distribution of important, treatment-informing molecular markers across each tumor by integrating biopsy data, MRI, and biological knowledge. This capability can help resolve the spatial heterogeneity of molecular characteristics and greatly improve the precision of cancer treatment. Other applications include early detection of regional fire risk across a forest by integrating ground/aerial survey data, satellite imagery, and fire simulator output, as well as regional poverty estimation for resource allocation.**

***Index Terms*— Health care, machine learning, precision medicine, statistical modeling.**

I. Introduction

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.

1) In precision medicine of cancer, one leading cause of treatment failure is intratumor heterogeneity [1], [2]. 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 suboptimal treatment outcomes. If the spatial molecular distribution could be precisely mapped out for each tumor, cancer treatments could be greatly improved.
2) In forest fire management, the ability for predicting regional fire risk across the forest is important for early detection and prevention [3].
3) 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 [4].

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, the 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, the direct measurement for fire risk must be done through aerial or ground survey, which can only sample a few locations

Manuscript received December 13, 2020; revised April 1, 2021; accepted April 19, 2021. This article was recommended for publication by Associate Editor Z. Kong and Editor X. Xie upon evaluation of the reviewers' comments. This work was supported in part by NIH under Grant U01 CA220378-01 and in part by NSF under Grant DMS-2053170. *(Corresponding author: Jing Li.)*

Lujia Wang and Jing Li are with the H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: jing.li@isye.gatch.edu; lwang724@gatech.edu).

Andrea Hawkins-Daarud and Kristin R. Swanson are with the Mathematical Neuro-Oncology Laboratory, Department of Neurosurgery, Mayo Clinic Arizona, Phoenix, AZ 85054 USA (e-mail: Hawkins-Daarud.Andrea@mayo.edu; Swanson.Kristin@mayo.edu).

Leland S. Hu is with the Department of Radiology, Mayo Clinic Arizona, Phoenix, AZ 85054 USA (e-mail: Hu.Leland@mayo.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TASE.2021.3076117.

Digital Object Identifier 10.1109/TASE.2021.3076117

TABLE I
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)

of the forest. For the same reason, survey data that directly reflect 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 noninvasively 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 are 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 [5], [6]. 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 [7]. 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.

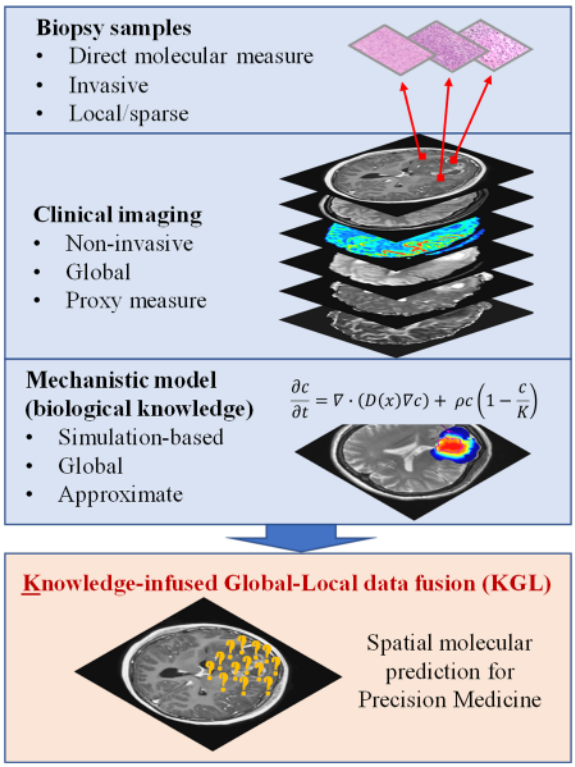

Fig. 1. Schematic overview of the multidata/information fusion framework by the proposed KGL methodology. The framework is illustrated using the application of Precision Medicine, but it is generalizable to other applications given in Table I.

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 I 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 article 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 Fig. 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); and 3) being consistent with the trend or patterns conveyed by domain knowledge.

The contributions of this article are summarized as follows.

1) *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.
2) *Novel Machine Learning Development:* KGL primarily intersects with two subfields in statistical modeling and

machine learning: semisupervised 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 nonparametric, nonlinear, 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 Section II, the existing models in SSL and GP do not provide the capability of KGL.

3) *Theoretical Insight:* We demonstrate that the formulation of KGL belongs to the machine learning paradigm called posterior regularization (PostReg) [8], [9]. 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 the specification of the prior. However, in many applications such as the examples mentioned in Table I, 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 article is the first effort that demonstrates its practical utility in integrating local data, global data, and domain knowledge for spatial prediction.
4) *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.

II. Related Works

KGL primarily intersects with two sub-fields in machine learning: SSL and GP.

A. Semisupervised Learning

SSL is used in situations where labeled samples are scarce but unlabeled samples are available in a large quantity. A typical supervised learning model would only utilize the labeled samples to build a predictive model, whereas SSL can leverage the unlabeled samples. The problem we are targeting in this article has the same nature: the local data with direct measurement for the variable of interest such as biopsy samples and survey samples are labeled and scarce; the imagery data with indirect measurement are unlabeled and available global-wide.

The existing SSL algorithms fall into several main categories. Self-training is a type of wrapper algorithm that repeatedly adds those unlabeled samples predicted with the highest confidence to the training set [10]. Co-training is an extension of self-training, which leverages two views of the data. It assumes that there are two separate datasets which contain conditionally independent feature sets. Two classifiers are built on the two datasets but with information exchange with each other [11], [12]. Low-density separation aims to find the decision boundary in low-density regions in the feature space based on labeled and unlabeled samples [13]. Graph-based models define a graph in which nodes represent labeled and unlabeled samples, and edges reflect the similarity between nodes. Label smoothness is assumed over the graph to allow label diffusion to unlabeled samples [5], [14].

B. GP Model

GP belongs to Bayesian nonparametric kernel-based probabilistic models [15]. Compared to other predictive models, GP has some unique aspects: First, GP makes few assumptions about the shape of the estimator function beyond the assumptions associated with the choice of the covariance function [16]. Another major benefit is its inherently probabilistic nature. GP can generate a predictive distribution for the response variable based on features, instead of just a point estimator of the prediction. This allows for uncertainty quantification and more informed decision-making based on the prediction result [16]. In this article, we are targeting a prediction problem which in theory might use some other predictive models as a baseline. However, GP is chosen due to its advantages of being nonparametric, nonlinear, and most importantly the capacity for generating a predictive probability distribution for the variable of interest. This allows for uncertainty quantification and reduction.

The standard GP is a predictive model. However, due to the aforementioned advantages, GP has been extended to impact multiple subfields of machine learning, such as multitask learning [17], SSL [18], and time series modeling [19]. In terms of application domains, GP and extensions have been used for medical decision-making [20], financial analysis [21], and computer experiments [22].

C. Gaps of the Existing Research

Given the problem we aim to solve, as described in Introduction, none of the existing methods alone would suffice. Here, we discuss some options of applying existing methods directly to our problem and why they are insufficient.

1) One option is to build a predictive model to link local features extracted from imagery with local direct measurement for the variable of interest (i.e., labeled samples). This model can then be used to predict the areas of the imagery where direct measurement is not available. This is the typical procedure when applying a supervised learning model. The limitations are multifold.
    a) Even though we could use GP to build the predictive model, the model can only quantify the predictive uncertainty but not reduce it.

b) Domain knowledge is not integrated in model training.
c) A large portion of the imagery is unlabeled, whose data are not leveraged in the training process.

2) An SSL model can be used to leverage the unlabeled imagery, which, however, still does not tackle the first two limitations in (a) and (b) as mentioned above.

In all, we will need to develop a new model that can simultaneously leverage labeled, local direct measurement and unlabeled, global imagery, reduce the uncertainty of the prediction, as well as integrate domain knowledge with data-driven model training. This capacity does not currently exist and we aim to provide this capacity by the new KGL model.

III. Preliminaries

Let f be a random variable corresponding to an input vector $\mathbf{x}$. GP is a collection of the random variables, any finite number of which have a joint Gaussian distribution. Consider a set that includes L labeled samples, $\{\mathbf{x}_i, y_i\}_{i=1}^{L}$, and an unlabeled sample, $\mathbf{x}^* \in \{\mathbf{x}_i\}_{i=L+1}^{L+U}$. The joint Gaussian distribution of this set is

$$\begin{bmatrix} \mathbf{f} \\ f^* \end{bmatrix} = N\left(\mathbf{0}, \begin{bmatrix} K(\mathbf{X}_L, \mathbf{X}_L) & K(\mathbf{X}_L, \mathbf{x}^*) \\ K(\mathbf{X}_L, \mathbf{x}^*)^T & K(\mathbf{x}^*, \mathbf{x}^*) \end{bmatrix}\right) \tag{1}$$

where K contains covariances between the corresponding samples, computed based on the input variables using kernels. Furthermore, introducing the noise term, the joint distribution of response variables corresponding to the labeled and unlabeled samples is

$$\begin{bmatrix} \mathbf{y}_L \\ y^* \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ f^* \end{bmatrix} + \sigma^2 \mathbf{I}_{(L+1)\times(L+1)}. \tag{2}$$

To predict the response of the unlabeled sample $\mathbf{x}^*$, we can obtain the predictive distribution of f^* by combining (1) and (2), i.e.,

$$f^* | \mathbf{X}_L, \mathbf{y}_L, \mathbf{x}^* \sim N\left(\mu^*, \sigma^{*^2}\right) \tag{3}$$

where

$$\begin{aligned} \mu^* &= K(\mathbf{X}_L, \mathbf{x}^*)^T \left(K(\mathbf{X}_L, \mathbf{X}_L) + \sigma^2 \mathbf{I}_{L\times L}\right)^{-1} \mathbf{y}_L \\ \sigma^{*^2} &= K(\mathbf{x}^*, \mathbf{x}^*) - K(\mathbf{X}_L, \mathbf{x}^*)^T \left(K(\mathbf{X}_L, \mathbf{X}_L) + \sigma^2 \mathbf{I}_{L\times L}\right)^{-1} \\ &\quad \times K(\mathbf{X}_L, \mathbf{x}^*). \end{aligned}$$

Equation (3) contains parameters to be estimated, including parameters in the kernel function and σ^2. Let $\boldsymbol{\theta}$ be the set of all parameters. $\hat{\theta}$ can be estimated by maximizing the marginal likelihood of the labeled samples, i.e.,

$$\min_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} -\log p(\mathbf{y}_L | \mathbf{X}_L, \boldsymbol{\theta}). \tag{4}$$

IV. KGL Data Fusion Model

A. Mathematical Formulation

Adopt the notation in the Preliminaries section 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

$$\min_{\boldsymbol{\theta}} \frac{1}{L} l(\boldsymbol{\theta}) \tag{5}$$

$$\text{s.t. } \frac{1}{U} \sum_{i=L+1}^{L+U} \text{Var}(f_i) \leq t \tag{6}$$

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 (4), 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 (6) on the predictive means of unlabeled samples, i.e., (10)–(12)

$$\min_{\boldsymbol{\theta}} \frac{1}{L} l(\boldsymbol{\theta}) \tag{7}$$

$$\text{s.t. } \frac{1}{U} \sum_{i=L+1}^{L+U} \text{Var}(f_i) \leq t \tag{8}$$

$$\mathbf{g}(E(f_{L+1}), \ldots, E(f_{L+U})) \leq \boldsymbol{\xi} \tag{9}$$

$$\boldsymbol{\xi} \geq 0 \tag{10}$$

$$\boldsymbol{\xi}^T \mathbf{1} \leq \epsilon \tag{11}$$

where $\mathbf{1}$ is a vector of m ones. $\mathbf{g}(\cdot)$ contains m different functions, $g_1(\cdot), \ldots, g_m(\cdot)$. Each $g_j(\cdot)$ is a function of the predictive means of unlabeled samples, $j = 1, \ldots, m$. $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_m)$ contains the upper bounds of these functions. A special case is when $m = 1$. Then, (9) reduces to a single function of $g(E(f_{L+1}), \ldots, E(f_{L+U})) \leq \xi$. Sometimes, a single function is not enough to represent different kinds of domain knowledge. Thus, we use a general notation in (9) 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., $g(E(f_{L+1}), \ldots, E(f_{L+U})) = \xi$, the equation can always been represented by two inequalities of $g(E(f_{L+1}), \ldots, E(f_{L+U})) \leq \xi$ and $-g(E(f_{L+1}), \ldots, E(f_{L+U})) \leq -\xi$, which can be added to the constraint set in (9). 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 (9)–(11). 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,

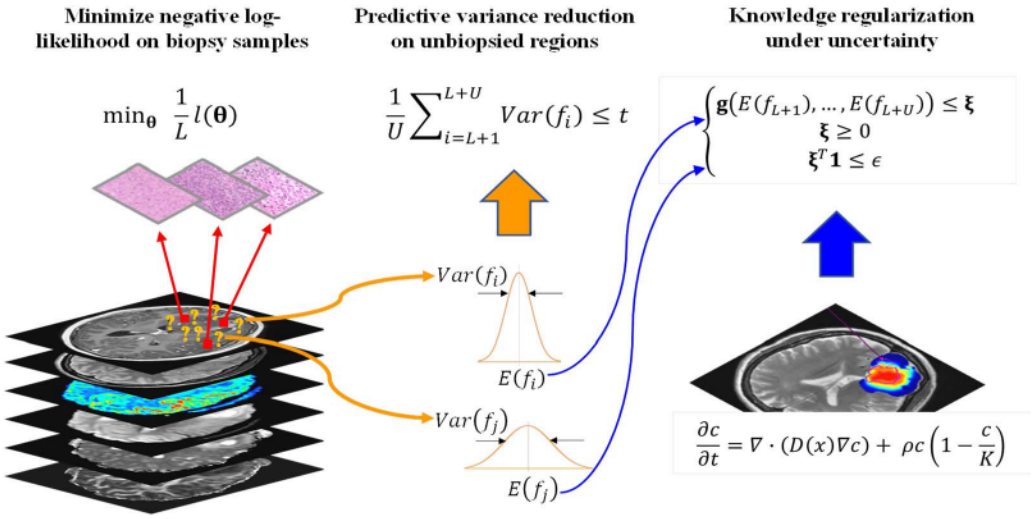

Fig. 2. Mathematical formulation of KGL as a constrained optimization.

please see Fig. 2 for a graphical illustration of the afore-described constrained optimization framework for KGL.

B. Optimization Algorithm for KGL Model Estimation

To solve the optimization problem in (7)–(11), we first write the corresponding Lagrangian function, i.e.,

$$\mathcal{L} = \frac{1}{L} l(\theta) + \alpha_1 \left(\frac{1}{U} \sum_{i=L+1}^{L+U} \mathrm{Var}(f_i) - t \right) + \sum_{j=1}^{m} \mu_j (g_j(\cdot) - \xi_j) - \sum_{j=1}^{m} \upsilon_j \xi_j + \alpha_2 \left(\sum_{j=1}^{m} \xi_j - \epsilon \right) \quad (12)$$

with Lagrange multipliers $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)$, $\boldsymbol{\upsilon} = (\upsilon_1, \ldots, \upsilon_m)$, $\alpha_1 \in \mathbb{R}$ and $\alpha_2 \in \mathbb{R}$, and $g_j(\cdot)$ used to represent $g_j(E(f_{L+1}), \ldots, E(f_{L+U}))$ for notation simplicity. Then, the optimal solution of the primal problem in (7)–(11) is equivalent to the solution of the following optimization:

$$\inf_{\theta,\xi} \sup_{\mu \ge 0, \upsilon \ge 0, \alpha_1 \ge 0, \alpha_2 \ge 0} \mathcal{L}. \quad (13)$$

Theorem 1: Let $\mathcal{L}' = (1/L) l(\theta) + \lambda_1((1/U)\sum_{i=L+1}^{L+U} \mathrm{Var}(f_i)) + \sum_{j=1}^{m} \mu_j (g_j(\cdot) - \xi_j) - \sum_{j=1}^{m} \upsilon_j \xi_j + \lambda_2 (\sum_{j=1}^{m} \xi_j)$, 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 $\inf_{\theta,\xi} \sup_{\mu \ge 0, \upsilon \ge 0, \alpha_1 \ge 0, \alpha_2 \ge 0} \mathcal{L}$ is equal to that of $\inf_{\theta,\xi} \sup_{\mu \ge 0, \upsilon \ge 0} \mathcal{L}'$ and vice versa (proof in Appendix A).

According to Theorem 1, (13) can be further simplified as

$$\inf_{\theta,\xi} \sup_{\mu \ge 0, \upsilon \ge 0} \mathcal{L}'. \quad (14)$$

Since $\mathcal{L}'$ is a convex function of $\xi_j, \boldsymbol{\mu}, \boldsymbol{\upsilon}$ (nonconvex of $\boldsymbol{\theta}$), (14) is equivalent to

$$\inf_{\theta} \sup_{\mu \ge 0, \upsilon \ge 0} \inf_{\xi} \mathcal{L}'. \quad (15)$$

Focus on the inner minimization in (15). The minimizer of ξ_j must satisfy

$$\frac{\partial \mathcal{L}'}{\partial \xi_j} = \lambda_2 - \mu_j - \upsilon_j = 0, \quad j = 1, \ldots, m. \quad (16)$$

From (16), we can write $\upsilon_j = \lambda_2 - \mu_j$. Inserting this into (15), we get

$$\inf_{\theta} \sup_{\mu \ge 0} \mathcal{J}(\mu_j; j = 1, \ldots, m) \quad (17)$$

$$\text{s.t. } 0 \le \mu_j \le \lambda_2, \quad j = 1, \ldots, m \quad (18)$$

where

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

It is clear that the solution of the inner maximization of (17) with (18) is

$$\mu_j = \begin{cases} \lambda_2, & \text{if } g_j(\cdot) > 0 \\ \text{any value in } [0, \lambda_2], & \text{if } g_j(\cdot) = 0 \\ 0, & \text{if } g_j(\cdot) < 0. \end{cases}$$

Then, the final objective function becomes

$$\inf_{\theta} L(\theta) = \inf_{\theta} \frac{1}{L} l(\theta) + \lambda_1 \left(\frac{1}{U} \sum_{i=L+1}^{L+U} \mathrm{Var}(f_i) \right) + \lambda_2 \sum_{j=1}^{m} g_j(\cdot) I(g_j(\cdot) > 0). \quad (19)$$

The gradient of objective function in (19) can be written as

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

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

Discussion on the Insight of the Optimization: Note that the optimization in (19) 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 (4)]; minimizing the predictive variances/uncertainty on unlabeled samples; optimizing the consistency with domain knowledge. The last term in (19) 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 (7)–(11), 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 $I(g_j(\cdot) > 0) = 0$. For the remaining constraints corresponding to $g_j(\cdot) > 0$ or equivalently $I(g_j(\cdot) > 0) = 1$, 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 prespecifying 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 (19) is nonconvex, the converged solution may not be the global optimal. This is a common problem for nonconvex optimization problems. A typical strategy is to use different initial values. More sophisticated nonconvex optimization algorithms may be used but are left for future investigation.

V. Another View: KGL as 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 [9]. The basic idea of PostReg is to use a variational distribution $q(M \mid D)$ to approximate the posterior $p(M|D)$, while at the same time regularizing $q(M \mid D)$ according to domain knowledge. That is, PostReg aims to find the solution $q^*(M \mid D)$ for the following optimization:

$$\inf_{q \in \mathcal{P}_{\text{prob}}} KL(q(M \mid D) \| p(M|D)) + \Omega(q(M \mid D)). \quad (20)$$

The first term is the Kullback–Leibler (KL)-divergence, defined as the expected log-difference between the posterior and approximate distributions. $\Omega(.)$ is a function of the approximate distribution, which regularizes this distribution to comply with domain knowledge. Because of the regularization effect, $q(M \mid D)$ 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. $\mathcal{P}_{\text{prob}}$ denotes a proper variational family of distributions. The PostReg optimization in (20) is a general formulation. It has been realized for specific models such as latent variable models under the EM framework [8], multiview learning [9], and infinite support vector machines [23].

We demonstrate that solving the optimization in (7)–(11) is equivalent to solving a specific form of the PostReg optimization. In this specific form, the choice of the regularizer $\Omega(q(M \mid D))$ corresponds to variance minimization and consistency with domain knowledge in expectation. This theoretical result is summarized in Theorem 2.

Theorem 2: The optimization in (7)–(11) is equivalent to a PostReg optimization taking the form of, i.e.,

$$\inf_{q \in \mathcal{P}_{\text{prob}}} KL(q(M \mid D) \| p(M|D)) + \Omega(q(M \mid D)) \quad (21)$$

with the following specific definitions for the notations: $M = (f, \boldsymbol{\theta})$ is the model; $D = (\{\mathbf{x}_i, y_i\}_{i=1}^{L}, \{\mathbf{x}_i\}_{i=L+1}^{L+U})$ is the data; $\mathcal{P}_{\text{prob}} = \{q | q(f, \boldsymbol{\theta} \mid D) = p(f \mid \boldsymbol{\theta}, D)\delta_{\bar{\boldsymbol{\theta}}}(\boldsymbol{\theta} \mid D), \bar{\boldsymbol{\theta}} \in \Theta\}$ is a variational family of distributions where $q(f \mid \boldsymbol{\theta}, D) = p(f \mid \boldsymbol{\theta}, D)$ and $q(\boldsymbol{\theta} \mid D) = \delta_{\bar{\boldsymbol{\theta}}}(\boldsymbol{\theta} \mid D)$ which is a Dirac delta function centered on $\bar{\boldsymbol{\theta}}$ in the parameter space Θ; $\Omega(q(f, \boldsymbol{\theta} \mid D))$, denoted by a simple form of $\Omega(q)$ hereafter, is given by (22), as shown at the bottom of the page.

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 the practical utility of using the concept of PostReg to integrate local data, global data, and domain knowledge for spatial estimation.

VI. Experiments

A. Data Collection and Preprocessing

GBM is the most aggressive type of brain tumor with median survival of 15 months [24]. Intratumor molecular heterogeneity has been found to be one of the leading causes of treatment failure. TCD is an important molecular marker to

$$\Omega(q) = \inf_{t,\boldsymbol{\xi}} \left\{ (\lambda_1 t + \lambda_2 \sum_{j=1}^{m} \xi_j) \;\middle|\; \begin{array}{c} \frac{1}{U} \sum_{i=L+1}^{L+U} \left(\int_{f,\boldsymbol{\theta}} q \times \left(f(\mathbf{x}_i) - E_q[f(\mathbf{x}_i)] \right)^2 d\eta(f, \boldsymbol{\theta}) \right) \leq t; \\ \mathbf{g}\left(\int_{f,\boldsymbol{\theta}} q \times f(\mathbf{x}_{L+1}) d\eta(f, \boldsymbol{\theta}), \ldots, \int_{f,\boldsymbol{\theta}} q \times f(\mathbf{x}_{L+U}) d\eta(f, \boldsymbol{\theta}) \right) \leq \boldsymbol{\xi}; \\ \boldsymbol{\xi} \geq \mathbf{0} \end{array} \right\} \quad (22)$$

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 subregions of each tumor [1], [2]. 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 subregion of a tumor, a 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 noninvasively. But MRI does not directly measure TCD while only providing proxy data. In this experiment, we apply KGL to predict the regional TCD of each tumor by integrating MRI, biopsy samples, and mechanistic model/domain knowledge.

1) 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. Preoperative 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.

2) MRI Preprocessing 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 coregistration can be found in our prior publications [2], [5]. To extract features, an 8×8 pixel^2 window was placed at each pixel as the center within a presegmented 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 8×8 pixel^2 window from each of the six contrast images and used as features. Therefore, six image features were included in model training.

3) 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.

4) Mechanistic Model: We integrate a well-known mechanistic model called proliferation-invasion (PI) [5], [6]. 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

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

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 $2(D\rho)^{1/2}$ and 2) the gradient of the wavefront, 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 TCD profile [25], [26]. In line with previous articles, 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% TCD threshold and the $T2$ image to the 16% TCD. 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 Section VI-B).

B. Application of KGL

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., $\mathbf{g}(E(f_{L+1}), \ldots, E(f_{L+U})) \leq \boldsymbol{\xi}$. Due to the aforementioned properties of the PI map, we propose to use it to regularize the general spatial trend of the TCD predictions.

$$\begin{cases} g_1(E(f_{L+1}), \ldots, E(f_{L+U})) \triangleq |E(f_{L+1}) - PI_{L+1}| \leq \xi_1 \\ \vdots \\ g_U(E(f_{L+1}), \ldots, E(f_{L+U})) \triangleq |E(f_{L+U}) - PI_{L+U}| \leq \xi_U \end{cases} \tag{23}$$

$$g_{U+1}(E(f_{L+1}), \ldots, E(f_{L+U})) \triangleq \sum_{i=L+1,\ldots,L+U;\, j>i} w_{ij}\left(E(f_i) - E\left(f_j\right)\right)^2 \leq \xi_{U+1} \tag{24}$$

$$\xi_1, \ldots, \xi_{U+1} \geq 0, \quad \sum_{i=1}^{U+1} \xi_i \leq \epsilon \tag{25}$$

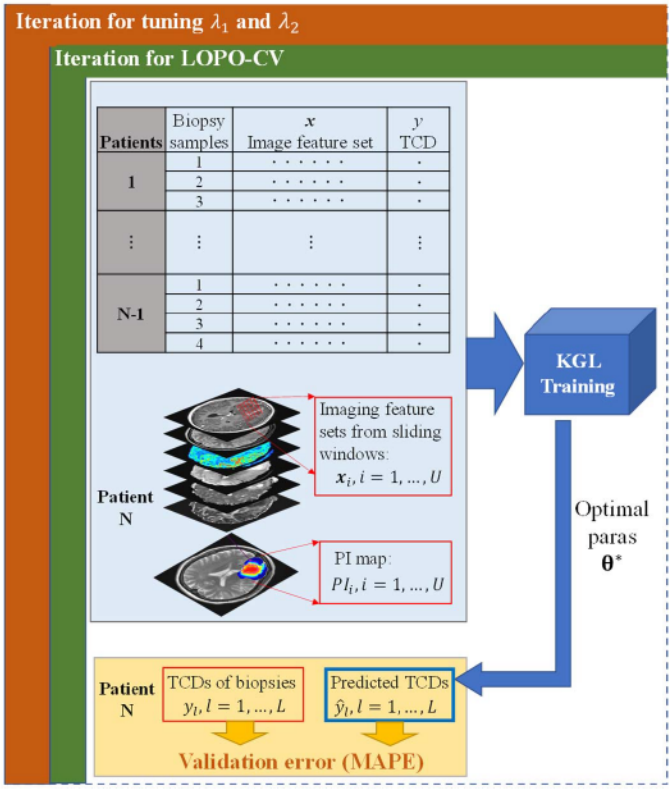

Fig. 3. Model training procedure for KGL.

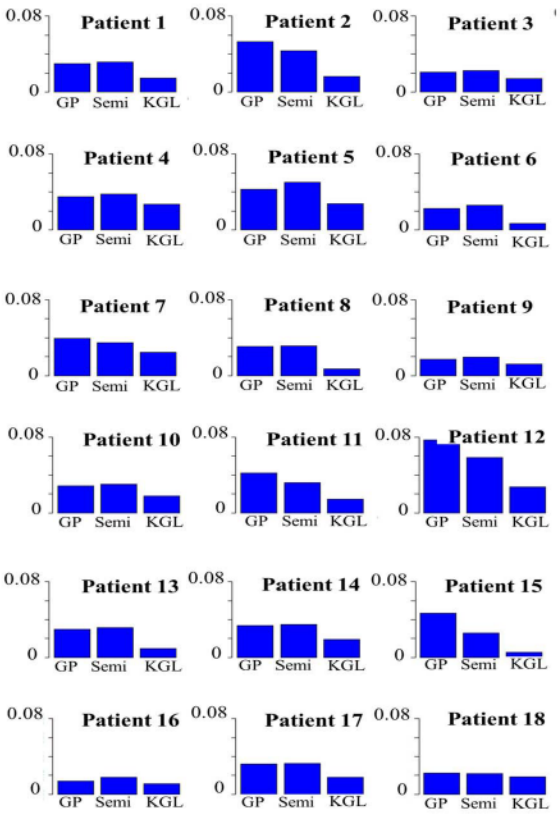

Fig. 4. Comparison of methods on average predictive variance of unlabeled samples for each patient. Averages across all patients: Standard GP = 0.032; Semi-GP = 0.032; KGL = 0.014 (56% variance reduction compared with the other two methods).

Specifically, based on the pixel-wise estimates of TCD generated by PI, we compute the average estimate over 64 pixels within each 8×8 pixel^2 window corresponding to an unlabeled sample. Denote this average estimate for each unlabeled sample i by $PI_i, i = L+1, \ldots, L+U$. The proposed constraints are (23)–(25), as shown at the bottom of the previous page, where $w_{ij} = e^{-(PI_i - PI_j)^2}$. The constraints in (23) encourage similarity between the predictive mean and the PI estimate at the same location (unbiopsied sample). Additionally, the constraint in (24) 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 (25) to make these constraints soft instead of hard constraints.

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 Fig. 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 [15], i.e., a GP model trained using only biopsy samples.
3) *Semi-GP:* A semisupervised GP model based on a data-dependent covariance function for unlabeled data [18].
4) *Co-training SVR-KNN:* An SSL algorithm based on co-training with support vector regression (SVR) and k-nearest neighbors (KNN) [12].
5) *SSRR-AGLP:* Semi-supervised ridge regression with adaptive graph-based label propagation [14].
6) *SS-RT:* Semi-supervised regression trees [13].
7) *SAFER:* SAFE semisupervised regression [27].
8) *KGL With No Variance Reduction:* This is a special case of KGL without the constraint on predictive variances.

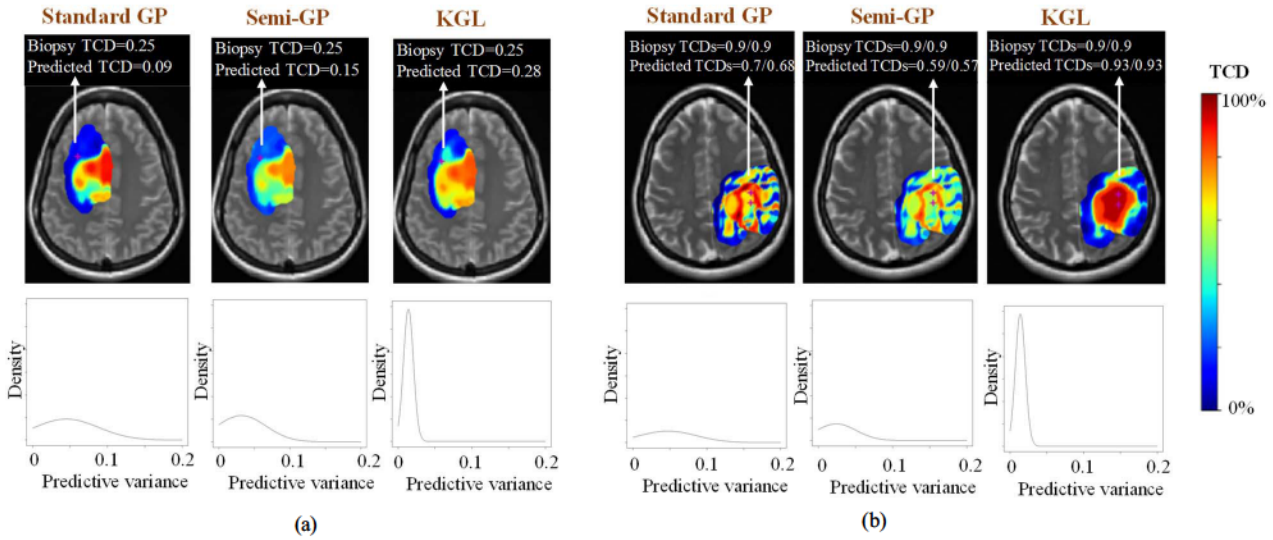

Fig. 5. Predicted means of TCD within an ROI are shown as a color map overlaid on the patient's $T2$ MRI; predicted variances are shown in distribution. (a) Patient A. (b) Patient B.

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 and 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.

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.

C. Results

Table II 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.

TABLE II
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**

Fig. 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.

Furthermore, Fig. 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 align better with the expected tumor cell distributions from known biology, especially for the color map of patient B. This is a benefit due to the 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.

D. 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.

VII. Conclusion

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 nonnumerical response variables, optimal selection of unlabeled samples, and development of more efficient optimization solvers.

Appendix A
Proof of Theorem 1

According to the derivation process from (15) to (19), $\inf_{\theta,\xi}\sup_{\mu\geq 0,v\geq 0}\mathcal{L}'$ can be simplified as

$$\inf_{\theta}\frac{1}{L}l(\theta)+\lambda_1\left(\frac{1}{U}\sum_{i=L+1}^{L+U}\mathrm{Var}(f_i)\right)+\lambda_2\sum_{j=1}^{m}g_j(\cdot)I(g_j(\cdot)>0). \tag{26}$$

$$\inf_{q\in\mathcal{P}_{\text{prob}}}\inf_{t,\xi}\sup_{\alpha_1,\mu,\mathbf{v}\geq 0}\left\{\begin{array}{c}KL(q\|p(f,\theta\mid D))+\lambda_1 t+\lambda_2\left(\sum_{j=1}^{m}\xi_j\right)+\alpha_1\left(\frac{1}{U}\sum_{i=L+1}^{L+U}\Omega_i^1(q)-t\right)\\ +\sum_{j=1}^{m}\mu_j\left(g_j\left(\Omega_{L+1}^2(q),\ldots,\Omega_{L+U}^2(q)\right)-\xi_j\right)-\sum_{j=1}^{m}v_j\xi_j\end{array}\right\} \tag{28}$$

$$\inf_{q\in\mathcal{P}_{\text{prob}}}\sup_{\alpha_1,\mu,\mathbf{v}\geq 0}\inf_{t,\xi}\left\{\begin{array}{c}KL(q\|p(f,\theta\mid D))+\lambda_1 t+\lambda_2\left(\sum_{j=1}^{m}\xi_j\right)+\alpha_1\left(\frac{1}{U}\sum_{i=L+1}^{L+U}\Omega_i^1(q)-t\right)\\ +\sum_{j=1}^{m}\mu_j\left(g_j\left(\Omega_{L+1}^2(q),\ldots,\Omega_{L+U}^2(q)\right)-\xi_j\right)-\sum_{j=1}^{m}v_j\xi_j\end{array}\right\}. \tag{29}$$

$$\inf_{q\in\mathcal{P}_{\text{prob}}}\sup_{\mu\geq 0}\left\{KL(q\|p(f,\theta\mid D))+\lambda_1\left(\frac{1}{U}\sum_{i=L+1}^{L+U}\Omega_i^1(q)\right)+\sum_{j=1}^{m}\mu_j g_j\left(\Omega_{L+1}^2(q),\ldots,\Omega_{L+U}^2(q)\right)\right\}$$
$$\text{s.t. } 0\leq\mu_j\leq\lambda_2,\quad j=1,\ldots,m.$$

$$\inf_{q\in\mathcal{P}_{\text{prob}}}\left\{\begin{array}{c}KL(q\|p(f,\theta\mid D))+\lambda_1\left(\frac{1}{U}\sum_{i=L+1}^{L+U}\Omega_i^1(q)\right)\\ +\lambda_2\sum_{j=1}^{m}g_j\left(\Omega_{L+1}^2(q),\ldots,\Omega_{L+U}^2(q)\right)I\left(g_j\left(\Omega_{L+1}^2(q),\ldots,\Omega_{L+U}^2(q)\right)>0\right)\end{array}\right\} \tag{30}$$

$$\begin{aligned}\inf_{q\in\mathcal{P}_{\text{prob}}}\chi&=\inf_{q\in\mathcal{P}_{\text{prob}}}\left\{\begin{array}{c}\int_{f,\theta}q\log\frac{q}{p(f,\theta|D)}d\eta(f,\theta)+\lambda_1\left(\frac{1}{U}\sum_{i=L+1}^{L+U}\Omega_i^1(q)\right)\\ +\lambda_2\sum_{j=1}^{m}g_i\left(\Omega_{L+1}^2(q),\ldots,\Omega_{L+U}^2(q)\right)I\left(g_i\left(\Omega_{L+1}^2(q),\ldots,\Omega_{L+U}^2(q)\right)>0\right)\end{array}\right\}\\ &=\inf_{q\in\mathcal{P}_{\text{prob}}}\left\{\begin{array}{c}\int_{f,\theta}p(f\mid\theta,D)\delta_{\tilde{\theta}}(\theta\mid D)\log\frac{p(f\mid\theta,D)\delta_{\tilde{\theta}}(\theta\mid D)}{p(f,\theta|D)}d\eta(f,\theta)+\lambda_1\left(\frac{1}{U}\sum_{i=L+1}^{L+U}\Omega_i^1(q)\right)\\ +\lambda_2\sum_{j=1}^{m}g_i\left(\Omega_{L+1}^2(q),\ldots,\Omega_{L+U}^2(q)\right)I\left(g_i\left(\Omega_{L+1}^2(q),\ldots,\Omega_{L+U}^2(q)\right)>0\right)\end{array}\right\}\end{aligned} \tag{31}$$

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

$$\inf_{\theta}\sup_{\alpha_1\geq 0,\alpha_2\geq 0}\left\{\frac{1}{L}l(\theta)+\alpha_1\left(\frac{1}{U}\sum_{i=L+1}^{L+U}\mathrm{Var}(f_i)-t\right)+\alpha_2\left(\sum_{j=1}^{m}g_j(\cdot)I\big(g_j(\cdot)>0\big)-\epsilon\right)\right\}. \quad (27)$$

To prove Theorem 1, we need to prove (26) and (27) are equivalent.

1) For any choice of $\lambda_1 > 0$ and $\lambda_2 > 0$, consider the optimal solution $\boldsymbol{\theta}^*$ from (26). It is not hard to see that $\boldsymbol{\theta}^*$ will also be the optimal solution to (27) if $t = (1/U)\sum_{i=L+1}^{L+U}\mathrm{Var}_{\boldsymbol{\theta}^*}(f_i)$, and $\epsilon = \sum_{j=1}^{m} g_{j;\boldsymbol{\theta}^*}(\cdot)I(g_{j;\boldsymbol{\theta}^*}(\cdot)>0)$; otherwise, if there is some other $\boldsymbol{\theta}'$ with $(1/U)\sum_{i=L+1}^{L+U}\mathrm{Var}(f_i)\leq t$ and $\sum_{j=1}^{m}g_j(\cdot)I(g_j(\cdot)>0)\leq\epsilon$, but a better objective value than $\boldsymbol{\theta}^*$ (note that since t and ϵ are preset, it becomes a hard-constraint optimization and it is easy to know that $\alpha_1 = 0$, $\alpha_2 = 0$, and $(1/L)l(\boldsymbol{\theta}') < (1/L)l(\boldsymbol{\theta}^*)$), then $(1/L)l(\boldsymbol{\theta}') + \lambda_1((1/U)\sum_{i=L+1}^{L+U}\mathrm{Var}_{\boldsymbol{\theta}'}(f_i)) + \lambda_2\sum_{j=1}^{m}g_{j,\boldsymbol{\theta}'}(\cdot)I(g_{j,\boldsymbol{\theta}'}(\cdot)>0) < (1/L)l(\boldsymbol{\theta}^*)+\lambda_1 t+\lambda_2\epsilon = (1/L)l(\boldsymbol{\theta}^*)+\lambda_1(\frac{1}{U}\sum_{i=L+1}^{L+U}\mathrm{Var}_{\boldsymbol{\theta}^*}(f_i))+\lambda_2\sum_{j=1}^{m}g_{j,\boldsymbol{\theta}^*}(\cdot)I(g_{j,\boldsymbol{\theta}^*}(\cdot)>0)$.

This contradicts the optimality of $\boldsymbol{\theta}^*$ in (26). Hence $\boldsymbol{\theta}^*$ is also optimal in (27).

Conversely, for any choice of $t > 0$ and $\epsilon > 0$, let $\boldsymbol{\theta}^*$ the optimal solution from (27), accompanied with the optimal α_1^* and α_2^*. Hence $\boldsymbol{\theta}^*$ is optimal in

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

$$\begin{aligned}
&\inf_{\bar{\theta}}\left\{\begin{array}{c}\int_{f,\theta}p(f\mid\theta,D)\delta_{\bar{\theta}}(\theta\mid D)\log\frac{\delta_{\bar{\theta}}(\theta\mid D)}{p(\theta\mid D)}d\eta(f,\theta)+\\ \lambda_1\left(\frac{1}{U}\sum_{i=L+1}^{L+U}\mathrm{Var}_p(f_i)\right)+\\ \lambda_2\sum_{j=1}^{m}g_j(\cdot)I(g_j(\cdot)>0)\end{array}\right\}\\
&=\inf_{\bar{\theta}}\left\{\int_{\theta}\delta_{\bar{\theta}}(\theta\mid D)\log\frac{\delta_{\bar{\theta}}(\theta\mid D)}{p(\theta\mid D)}d\eta(\theta)+\lambda_1\left(\frac{1}{U}\sum_{i=L+1}^{L+U}\mathrm{Var}_p(f_i)\right)+\lambda_2\sum_{j=1}^{m}g_j(\cdot)I(g_j(\cdot)>0)\right\}\\
&=\inf_{\bar{\theta}}\left\{-\int_{\theta}\delta_{\bar{\theta}}(\theta\mid D)\log p(\theta\mid D)d\eta(\theta)+\lambda_1\left(\frac{1}{U}\sum_{i=L+1}^{L+U}\mathrm{Var}_p(f_i)\right)+\lambda_2\sum_{j=1}^{m}g_j(\cdot)I(g_j(\cdot)>0)\right\}\\
&=\inf_{\bar{\theta}}\left\{-\int_{\theta}\delta_{\bar{\theta}}(\theta\mid D)\log\frac{p(\mathbf{y}_L,\theta\mid\mathbf{X}_L)}{p(\mathbf{y}_L\mid\mathbf{X}_L)}d\eta(\theta)+\lambda_1\left(\frac{1}{U}\sum_{i=L+1}^{L+U}\mathrm{Var}_p(f_i)\right)+\lambda_2\sum_{j=1}^{m}g_j(\cdot)I(g_j(\cdot)>0)\right\}\\
&=\inf_{\bar{\theta}}\left\{-\int_{\theta}\delta_{\bar{\theta}}(\theta\mid D)\log p(\mathbf{y}_L,\theta\mid\mathbf{X}_L)d\eta(\theta)+\lambda_1\left(\frac{1}{U}\sum_{i=L+1}^{L+U}\mathrm{Var}_p(f_i)\right)+\lambda_2\sum_{j=1}^{m}g_j(\cdot)I(g_j(\cdot)>0)\right\}\\
&=\inf_{\bar{\theta}}\left\{-\log p\big(\mathbf{y}_L,\bar{\theta}\mid\mathbf{X}_L\big)+\lambda_1\left(\frac{1}{U}\sum_{i=L+1}^{L+U}\mathrm{Var}_p(f_i)\right)+\lambda_2\sum_{j=1}^{m}g_j(\cdot)I(g_j(\cdot)>0)\right\}\\
&=\inf_{\bar{\theta}}\left\{-\log p\big(\mathbf{y}_L\mid\mathbf{X}_L,\bar{\theta}\big)p\big(\bar{\theta}\mid\mathbf{X}_L\big)+\lambda_1\left(\frac{1}{U}\sum_{i=L+1}^{L+U}\mathrm{Var}_p(f_i)\right)+\lambda_2\sum_{j=1}^{m}g_j(\cdot)I(g_j(\cdot)>0)\right\}\\
&=\inf_{\bar{\theta}}\left\{-\log p\big(\mathbf{y}_L\mid\mathbf{X}_L,\bar{\theta}\big)-\log p\big(\bar{\theta}\mid\mathbf{X}_L\big)+\lambda_1\left(\frac{1}{U}\sum_{i=L+1}^{L+U}\mathrm{Var}_p(f_i)\right)+\lambda_2\sum_{j=1}^{m}g_j(\cdot)I(g_j(\cdot)>0)\right\}\\
&=\inf_{\bar{\theta}}\left\{-\log p\big(\mathbf{y}_L\mid\mathbf{X}_L,\bar{\theta}\big)+\lambda_1\left(\frac{1}{U}\sum_{i=L+1}^{L+U}\mathrm{Var}_p(f_i)\right)+\lambda_2\sum_{j=1}^{m}g_j(\cdot)I(g_j(\cdot)>0\right\}\\
&=\inf_{\bar{\theta}}\left\{-\frac{1}{L}\log p\big(\mathbf{y}_L\mid\mathbf{X}_L,\bar{\theta}\big)+\frac{\lambda_1}{L}\left(\frac{1}{U}\sum_{i=L+1}^{L+U}\mathrm{Var}_p(f_i)\right)+\frac{\lambda_2}{L}\sum_{j=1}^{m}g_j(\cdot)I(g_j(\cdot)>0)\right\}\\
&=\inf_{\bar{\theta}}\left\{-\frac{1}{L}\log p\big(\mathbf{y}_L\mid\mathbf{X}_L,\bar{\theta}\big)+\lambda_1\left(\frac{1}{U}\sum_{i=L+1}^{L+U}\mathrm{Var}_p(f_i)\right)+\lambda_2\sum_{j=1}^{m}g_j(\cdot)I(g_j(\cdot)>0)\right\}. \quad (34)
\end{aligned}$$

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 (26). ■

APPENDIX B
PROOF OF THEOREM 2

Our proof aims to show that the optimization in (21) is equivalent to (7)–(11). For notation simplicity, define $\Omega_i^1(q) \triangleq \int_{f,\theta} q \times (f(\mathbf{x}_i) - E_q[f(\mathbf{x}_i)])^2 d\eta(f,\theta)$ and $\Omega_i^2(q) \triangleq \int_{f,\theta} q \times f(\mathbf{x}_i) d\eta(f,\theta)$. Then the constraints in (22) become $(1/U)\sum_{i=L+1}^{L+U} \Omega_i^1(q) \leq t$ and $g_j(\Omega_{L+1}^2(q), \ldots, \Omega_{L+U}^2(q)) \leq \xi_j, j = 1, \ldots, m$. Using the Lagrange multiplier method, we know that (21) is equivalent to (28) as shown at the bottom of page 10.

Since (28) is a convex function of $\boldsymbol{\xi}, t, \alpha_1, \boldsymbol{\mu}, \boldsymbol{v}$, (28) is equivalent to (29) as shown at the bottom of page 10.

Denote the function within the { } in (29) 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.,

$$\frac{\partial \varphi}{\partial t} = \lambda_1 - \alpha_1 = 0$$
$$\frac{\partial \varphi}{\partial \xi_j} = \lambda_2 - \mu_j - v_j = 0, \quad j = 1, \ldots, m.$$

From these equations we can get $\alpha_1 = \lambda_1$ and $v_j = \lambda_2 - u_j$. Putting these back to (29), we get the equation as shown at the bottom of page 10.

That can be simplified as (30), shown at the bottom of page 10.

Denote the function within { } in (30) by χ. Comparing (30) to that in Theorem 1, we know that the remaining task of this proof is to show that $\inf_{q \in \mathcal{P}_{\text{prob}}} \chi$ is equivalent to

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

Next, we show steps to prove this equivalency. (30) can be re-formed as (31), as shown at the bottom of page 10.

Now focus on the third term within the inf{ } in (31)

$$\begin{aligned} \Omega_i^2(q) &= \int_{f,\theta} p(f \mid \boldsymbol{\theta}, D) \delta_{\bar{\theta}}(\boldsymbol{\theta} \mid D) f(\mathbf{x}_i) d\eta(f,\theta) \\ &= \int_f f(\mathbf{x}_i) \int_{\theta} p(f \mid \boldsymbol{\theta}, D) \delta_{\bar{\theta}}(\boldsymbol{\theta} \mid D) d\eta(f,\theta) \\ &= \int_f f(\mathbf{x}_i) p(f \mid \bar{\boldsymbol{\theta}}, D) d\eta(f) = E_p[f(\mathbf{x}_i)] \end{aligned} \tag{32}$$

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

$$\begin{aligned} &\Omega_i^1(q) \\ &= \int_{f,\theta} p(f \mid \boldsymbol{\theta}, D) \delta_{\bar{\theta}}(\boldsymbol{\theta} \mid D) \times \left(f(\mathbf{x}_i) - E_p[f(\mathbf{x}_i)] \right)^2 d\eta(f,\theta) \\ &= \int_f \left(f(\mathbf{x}_i) - E_p[f(\mathbf{x}_i)] \right)^2 \int_{\theta} p(f \mid \boldsymbol{\theta}, D) \delta_{\bar{\theta}}(\boldsymbol{\theta} \mid D) d\eta(f,\theta) \\ &= \int_f \left(f(\mathbf{x}_i) - E_p[f(\mathbf{x}_i)] \right)^2 p(f \mid \bar{\boldsymbol{\theta}}, D) d\eta(f) = \text{Var}_p(f_i). \end{aligned} \tag{33}$$

Then, inserting (32) and (33) into (31), it becomes (34) as shown at the bottom of the previous page. ■

REFERENCES

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

[2] 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, Nov. 2015, Art. no. e0141506.

[3] O. Ghorbanzadeh, T. Blaschke, K. Gholamnia, and J. Aryal, "Forest fire susceptibility and risk mapping using social/infrastructural vulnerability and environmental variables," *Fire*, vol. 2, no. 3, p. 50, Sep. 2019.

[4] N. Jean, S. M. Xie, and S. Ermon, "Semi-supervised deep kernel learning: Regression with unlabeled data by minimizing predictive variance," in *Proc. Adv. Neural Inf. Process. Syst.*, Montréal, QC, Canada, 2018, pp. 5322–5333.

[5] 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, Dec. 2019.

[6] H. L. P. Harpold, E. C. Alvord, and K. R. Swanson, "The evolution of mathematical modeling of glioma proliferation and invasion," *J. Neuropathol. Exp. Neurol.*, vol. 66, no. 1, pp. 1–9, Jan. 2007.

[7] M. Denham, A. Cortés, T. Margalef, and E. Luque, "Applying a dynamic data driven genetic algorithm to improve forest fire spread prediction," in *Proc. Int. Conf. Comput. Sci.* Heidelberg, Germany: Springer-Verlag, 2008, pp. 36–45.

[8] K. Ganchev, B. Taskar, and J. Gama, "Expectation maximization and posterior constraints," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 20, 2007, pp. 569–576.

[9] K. Ganchev, J. Gillenwater, and B. Taskar, "Posterior regularization for structured latent variable models," *J. Mach. Learn. Res.*, vol. 11, pp. 2001–2049, Jul. 2010.

[10] N. Fazakis, S. Karlos, S. Kotsiantis, and K. Sgarbas, "Self-trained LMT for semisupervised learning," *Comput. Intell. Neurosci.*, vol. 2016, p. 10, Dec. 2016.

[11] S. Samiappan and R. J. Moorhead, "Semi-supervised co-training and active learning framework for hyperspectral image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 401–404.

[12] Y. Q. Li and M. Tian, "A semi-supervised regression algorithm based on co-training with SVR–KNN," *Adv. Mater. Res.*, vols. 926–930, pp. 2914–2918, May 2014.

[13] J. Levatić, M. Ceci, T. Stepišnik, S. Džeroski, and D. Kocev, "Semi-supervised regression trees with application to QSAR modelling," *Expert Syst. Appl.*, vol. 158, Nov. 2020, Art. no. 113569.

[14] Y. Yi *et al.*, "Semi-supervised ridge regression with adaptive graph-based label propagation," *Appl. Sci.*, vol. 8, no. 12, p. 2636, Dec. 2018.

[15] C. E. Rasmussen, "Gaussian processes in machine learning," in *Proc. Summer School Mach. Learn.*, 2003, pp. 63–71.

[16] M. S. Caywood, D. M. Roberts, J. B. Colombe, H. S. Greenwald, and M. Z. Weiland, "Gaussian process regression for predictive but interpretable machine learning models: An example of predicting mental workload across tasks," *Frontiers Hum. Neurosci.*, vol. 10, p. 647, Jan. 2017.

[17] E. V. Bonilla, K. M. Chai, and C. Williams, "Multi-task Gaussian process prediction," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 153–160.

[18] V. Sindhwani, W. Chu, and S. S. Keerthi, "Semi-supervised Gaussian process classifiers," in *Proc. IJCAI*, 2007, pp. 1059–1064.

[19] A. Damianou, M. K. Titsias, and N. D. Lawrence, "Variational Gaussian process dynamical systems," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 2510–2518.

[20] L. L. Duan, X. Wang, J. P. Clancy, and R. D. Szczesniak, "Joint hierarchical Gaussian process model with application to personalized prediction in medical monitoring," *Stat*, vol. 7, no. 1, p. e178, 2018.

[21] J. Gonzalvez, E. Lezmi, T. Roncalli, and J. Xu, "Financial applications of Gaussian processes and Bayesian optimization," 2019, *arXiv:1903.04841*. [Online]. Available: http://arxiv.org/abs/1903.04841

[22] P. Z. G. Qian, H. Wu, and C. F. J. Wu, "Gaussian process models for computer experiments with qualitative and quantitative factors," *Technometrics*, vol. 50, no. 3, pp. 383–396, 2008.

[23] J. Zhu, N. Chen, and E. P. Xing, "Bayesian inference with posterior regularization and applications to infinite latent SVMs," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1799–1847, 2014.

[24] S. De Vleeschouwer, *Glioblastoma*. Singapore: Codon Publications, 2017.
[25] 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," *Brit. J. Cancer*, vol. 98, no. 1, pp. 113–119, 2008.
[26] 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.
[27] Y.-F. Li, H.-W. Zha, and Z.-H. Zhou, "Learning safe prediction for semi-supervised regression," in *Proc. AAAI*, 2017, pp. 2217–2223.

Lujia Wang (Member, IEEE) received the B.S. degree in mathematics and applied mathematics from Nankai University, Tianjin, China, in 2013, and the M.S. degree in probability and mathematical statistics from the Chinese Academy of Sciences, Beijing, China, in 2016. She is currently pursuing the Ph.D. degree with the School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA.

Her research interests include machine learning and biomedical imaging analytics.

Andrea Hawkins-Daarud received the B.S. degree in mathematics from Walla Walla University, College Place, WA, USA, in 2005, and the M.S. and Ph.D. degrees in computational sciences engineering and mathematics from The University of Texas at Austin, Austin, TX, USA, in 2008 and 2011, respectively.

She then completed her postdoctoral training with Dr. Kristin Swanson. She is currently an Assistant Director with the Mathematical NeuroOncology Group, Mayo Clinic Arizona, Phoenix, AZ, USA. Her areas of research interest span mathematical oncology, cancer biology, parameter estimation, and uncertainty quantification.

Kristin R. Swanson received the B.S. degree in mathematics from Tulane University, New Orleans, LA, USA, in 1996, and the M.S. and Ph.D. degrees in mathematical biology from the University of Washington, Washington, DC, USA, in 1998 and 1999, respectively.

She was a Post-Doctoral Researcher at the University of California at San Francisco (UCSF). She is currently a Co-Director of the Precision NeuroTherapeutics Program as well as a Professor and a Vice Chair for Research for the Department of Neurosurgery at Mayo Clinic, Phoenix, AZ, USA. She is an internationally recognized mathematical oncologist focused on delivering optimal treatment to patients with brain cancer. Her expertise bridges mathematical modeling, oncology, artificial intelligence, and cancer biology.

Leland S. Hu received the M.D. degree from the University of Texas–Southwestern Medical Center, Dallas, TX, USA, in 2001.

He completed his clinical internship and residency in diagnostic radiology at the University of Texas–Southwestern Medical Center. He completed his two-year clinical fellowship in diagnostic neuroradiology from the Barrow Neurological Institute, Phoenix, AZ, USA. He received his Board Certification in diagnostic radiology and his subspecialty certification (CAQ) in neuroradiology from the American Board of Radiology (ABR). He is currently a Consultant Physician with the Department of Radiology, Mayo Clinic Arizona, Phoenix, and holds an academic appointment as an Assistant Professor of Radiology with the Mayo Clinic School of Medicine, Rochester, MN, USA. He oversees the clinical imaging component of the neuro-oncology program at Mayo Clinic Arizona. His research interests focus on image-based modeling and clinical imaging applications for the study of brain tumors.

Jing Li (Member, IEEE) received the Ph.D. degree in industrial and operations engineering from the University of Michigan at Ann Arbor, Ann Arbor, MI, USA, in 2007.

She is currently a Professor with the School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA. Her research interests are statistical modeling and machine learning for health care applications.

Dr. Li is a member of IISE and INFORMS. She was a recipient of the NSF CAREER Award.