2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-3748-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/BIBM58861.2023.10385861

Gaussian Process-based Active Learning for Efficient Cardiovascular Disease Inference

1st Stavroula C. Tassi
Department of Material Science and Engineering,
Unit of Medical Technology and Intelligent Information Systems (MEDLAB),
University of Ioannina
Ioannina, Greece &
Department of Mechanical and Aeronautics Engineering
University of Patras,
Patras, Greece
tassistav@uoi.gr

2nd Konstantinos D. Polyzos
Department of Electrical and Computer Engineering,
University of Minnesota
Minnesota, USA
polyz003@umn.edu

3rd Dimitrios I. Fotiadis, IEEE Fellow
Department of Material Science and Engineering,
Unit of Medical Technology and Intelligent Information Systems (MEDLAB),
University of Ioannina &
Biomedical Reasearch Institute,
Foundation for Research and Technology-Hellas(FORTH)
Ioannina, Greece
fotiadis@uoi.gr

4th Antonis I. Sakellarios
Laboratory of Biomechanics and Biomedical Engineering,
Department of Mechanical and Aeronautics Engineering
University of Patras,
Patras, Greece &
Unit of Medical Technology and Intelligent Information Systems (MEDLAB),
University of Ioannina
Ioannina, Greece
ansakel13@gmail.com

***Abstract*—Cardiovascular disease (CVD) poses a significant global health challenge, and accurate inference methods are vital for early detection and intervention. However, the quality of prediction relies heavily on the availability of labeled data, which are often limited in medical applications. To cope with the challenge of limited labeled data, we are the first to propose an active learning (AL) approach that leverages a weighted ensemble of Gaussian processes to effectively infer CVD by strategically selecting the few most informative data points to label. Through experiments conducted on the SMARTool dataset, we demonstrate the effectiveness of the advocated approach, achieving superior performance in CVD inference compared to baseline methods. Our findings highlight the potential impact of the proposed AL framework in CVD diagnosis and treatment clinical cases, particularly in scenarios where labeled data are scarce, due to data confidentiality concerns or high sampling costs.**

***Index Terms*—CVD inference, Gaussian Processes (GPs), Ensemble of Gaussian Processes (EGPs), FFR index**

*The project is co-financed by Greece and the European Union—European Regional Development Fund (ERDF) under the Operational Program "Competitiveness Entrepreneurship Innovation" (EPAnEK), NSRF 2014-2020 (Project code: 82863, MIS: 5047133), and has also received funding from the European Union's Horizon 2020 research and innovation program TO_AITION under grant agreement No 848146. The data used in the analysis were acquired by the SMARTool project (GA: 689068).

I. INTRODUCTION

Cardiovascular diseases (CVDs) encompass a range of heart and vascular dysfunctions, including coronary heart disease, heart failure, and arrhythmia. These conditions can become a leading cause of death, accounting for approximately one-third of all fatalities and significantly impacting individuals' quality of life. The severity of coronary artery disease is primarily evaluated through imaging techniques like X-ray coronary angiography. Fractional Flow Reserve (FFR) is a measurement taken through invasive means, indicating the severity of coronary stenosis. It serves as a valuable diagnostic tool for clinicians to determine if revascularization is necessary when the FFR is below 0.8. Although FFR is an invasive method, there have been recent advancements in non-invasive approaches using computational fluid dynamics [1], [2].

Significant attention has been given to the diagnosis and prevention of CVDs, with most of the studies focusing on classification models [3]–[5]. However, obtaining a sufficient number of labeled data regarding CVD-related patients may be challenging [6], [7] due to medical confidentiality and high annotation costs, since the latter entails the involvement

Authorized licensed use limited to: University of Ioannina. Downloaded on May 31,2025 at 19:07:31 UTC from IEEE Xplore. Restrictions apply.
979-8-3503-3748-8/23/$31.00 ©2023 IEEE

of specialized experts and (possibly) costly examinations. Therefore, building reliable machine learning (ML) and deep learning (DL) models for accurate CVD prediction with only *few* available labeled training data is a challenging task since many models require a sufficiently large number of training data. The objective of this work is to identify the few most informative unlabeled data that can be incrementally labeled to enhance ML regression tasks related to CVD diagnosis and prevention.

Active learning (AL) provides a principled approach within the realm of human-in-the-loop computing [8], allowing the chosen ML model and the training (labeled) set to evolve over time through the prudent labeling of the most informative samples from a pool of unlabeled instances. Given the objective of minimizing the expenses associated with acquiring labeled data by strategically selecting new data points, AL methods have found application in diverse domains such as natural language processing [9], and biological systems [10] to list a few; also within the context of multi-label classification [11]. Additionally, AL has also been utilized in prediction of risk of heart disease [12] and in the field of coronary artery disease using an ensemble of classifiers [13]. Various strategies, such as uncertainty sampling, query-by-committee, and variance reduction [14], [15] have been established to select the most informative and diverse samples for labeling.

In an effort to reduce data labeling costs and enhance model efficiency with well-quantifiable uncertainty, AL with Gaussian processes (GPs) has been applied. Particularly, the work in [16] utilizes a fully Bayesian GP model to improve model efficiency and in [17] proposed an AL approach with weak labels for GPs. Recently, AL has also been applied in a framework that uses an ensemble (E) of GP experts with weights adapted to the labeled data collected incrementally [18]. This EGP model has also been employed in several learning domains including graph-guided learning [19]–[22], reinforcement learning [23], transfer learning [24], and Bayesian optimization [25]–[27]. Building on the EGP model, the work in [18] presents a number of acquisition functions based on the uncertainty and disagreement rules, and adaptively learns the proper acquisition functions as new data are processed online. While the EGP-based AL framework has also been employed for learning over graphs [28], it has not been utilized for several healthcare domains including cardiovascular disease (CVD)-related tasks.

Contributions. In this paper, we advocate a well-motivated AL framework to assess the severity of coronary stenosis in patients undergoing computed tomography angiography (CTCA) by predicting the gold standard FFR with only a few yet informative training data. This work capitalizes on an ensemble of GPs learning models to build uncertainty-based acquisition criteria used to judiciously select the few most informative unlabeled data to label. Although the notion of EGPs has been employed for FFR index online inference in [29] using a *passively* selected labeled set, in the present work we are the first to introduce an EGP-based AL framework to effectively and efficiently guide CVD related regression tasks. The present work can markedly assist clinicians who can have access to only a few medical records, offering easy, accurate, and cost-effective diagnoses. Numerical tests on the CVD-related SMARTool dataset showcase the impressive merits of adopting the advocated EGP-based AL method for the FFR prediction with uncertainty measures and few data at hand.

II. MATERIALS AND METHODS

A. Data Description and Prepossessing

The present study utilized data from the SMARTool (Simulation Modeling of coronary ARTery disease: a tool for clinical decision support) project [30]. The resulting dataset consists of 187 patients, with input data representing the medical features of each patient including clinical risk factors, lipidomic data, bio-humoral variables, and CTCA imaging data. The SmartFFR index, computed for the most affected coronary artery, was used as the target variable in this study. Data cleaning and data transformation of the raw data were utilized, including the imputation of missing values taking into account the diverse variable types (categorical and continuous features), as well as outliers identification. Nevertheless, the outliers were retained in the analysis due to their potential clinical significance.

B. Problem Formulation

Conventional supervised learning methods aim to estimate a learning function $f(\cdot)$ that maps the input feature vector $\mathbf{x}_n$ to the corresponding output y_n; i.e $\mathbf{x}_n \rightarrow f(\mathbf{x}_n) \rightarrow y_n$. In a regression task, the value of y_n belongs to the set of real numbers, while in the classification task, it belongs to a finite set of options. To accurately identify f, a sufficient number of labeled training samples $(\mathbf{x}_n, y_n)_{n=1}^{t}$ may be necessary, which may not be available though in practical healthcare domains including the CVD-related tasks, thus motivating well the AL paradigm.

C. Active Learning

The AL process begins with a limited collection of labeled samples L_0, and a larger pool of unlabeled instances U_0. Given the corresponding sets, L_t and U_t at time slot t, probabilistic model-based AL approaches involve a statistical function model, represented by the probability density function (pdf) $p(f(x)|L_t)$, which is then utilized by the so-termed acquisition function (AF) $\alpha(\cdot)$ to select the next datum, $x_{t+1} \in U_t$ to label as:

$$\mathbf{x}_{t+1} = \underset{\mathbf{x} \in U_t}{\arg\max}\, \alpha\left(\mathbf{x}; L_t\right). \tag{1}$$

Upon obtaining the label y_{t+1} through an oracle, which could be for example a costly medical exam in a clinical task, the labeled set is augmented with $(\mathbf{x}_{t+1},\ y_{t+1})$, $\mathbf{x}_{t+1}$ is removed from the unlabeled set and this process is repeated iteratively. Hence, the two critical components that affect the AL performance are the selection of the learning model for f and the design of the AF α.

Authorized licensed use limited to: University of Ioannina. Downloaded on May 31,2025 at 19:07:31 UTC from IEEE Xplore. Restrictions apply.

D. Active Learning with single GP

Targeting a well-quantified uncertainty to guide the AL process, GP-based AL utilizes Gaussian Processes (GPs) to estimate a non-parametric function model efficiently, while also providing a measure of model uncertainty. Learning with GPs begins with the assumption that $f \sim \mathcal{GP}(0, \kappa(\mathbf{x}, \mathbf{x}'))$ with $\kappa(\cdot)$ denoting a positive-definite kernel function measuring similarity between inputs. This is equivalent to write that $\mathbf{f}_t := [f(\mathbf{x}_1) \dots f(\mathbf{x}_t)]^\top \sim \mathcal{N}(\mathbf{f}_t; \mathbf{0}_t, \mathbf{K}_t) \ \forall t$, where the (m, m') element of the $t \times t$ covariance (kernel) matrix $\mathbf{K}_t$ is $[\mathbf{K}_t]_{m,m'} = \mathrm{cov}(f(\mathbf{x}_m), f(\mathbf{x}_{m'})) := \kappa(\mathbf{x}_m, \mathbf{x}_{m'})$ [31].

The next assumption is that the batch likelihood $p(\mathbf{y}_t|\mathbf{f}_t; \mathbf{X}_t)$, with $\mathbf{X}_t := [\mathbf{x}_1 \dots \mathbf{x}_t]^\top$ and $\mathbf{y}_t := [y_1 \cdots y_t]^\top$, is factored as $p(\mathbf{y}_t|\mathbf{f}_t; \mathbf{X}_t) = \prod_{n=1}^t p(y_n|f(\mathbf{x}_n))$. With the GP prior and the batch likelihood at hand, it can be shown that the posterior of $f(\mathbf{x}) \sim \mathcal{N}(f(\mathbf{x}); \mu_t(\mathbf{x}), \sigma_t^2(\mathbf{x}))$ with mean and variance given in closed-form as:

$$\mu_t(\mathbf{x}) = \mathbf{k}_t^\top(\mathbf{x})(\mathbf{K}_t + \sigma_n^2 \mathbf{I}_t)^{-1}\mathbf{y}_t, \tag{2a}$$

$$\sigma_t^2(\mathbf{x}) = \kappa(\mathbf{x}, \mathbf{x}) - \mathbf{k}_t^\top(\mathbf{x})(\mathbf{K}_t + \sigma_n^2 \mathbf{I}_t)^{-1}\mathbf{k}_t(\mathbf{x}), \tag{2b}$$

where $\mathbf{k}_t(\mathbf{x}) := [\kappa(\mathbf{x}_1, \mathbf{x}), \dots, \kappa(\mathbf{x}_t, \mathbf{x})]^\top$. The mean in (2a) provides a point estimate of $f(\mathbf{x})$ and the variance in (2b) quantifies the uncertainty of this estimate.

Following the intuition that labeling the most uncertain instances and adding them to the labeled set can aid the prediction performance, a well-known GP-based AF is [32]:

$$\mathbf{x}_{t+1} = \underset{\mathbf{x} \in U_t}{\arg\max} \ \sigma_t^2(\mathbf{x}) \ . \tag{3}$$

Although interesting, its performance relies on a pre-selected $\kappa(\cdot)$ whose selection is a non-trivial task, especially when the number of initially available labeled data is small. To cope with this challenge, in the next section, we will outline an ensemble (E) of GPs framework to adaptively learn the proper kernel function, along with the corresponding AFs to guide the AL process.

E. Active Learning with EGPs

Accounting for a more expressive function space than that offered by a single GP, we advocate an ensemble (E) of GPs with each GP model $m \in \mathcal{M} := 1, \dots, M$ placing a unique GP prior $f \sim \mathcal{GP}(0, \kappa^m(\mathbf{x}, \mathbf{x}'))$ where the distinct kernel $\kappa^m(\mathbf{x}, \mathbf{x}')$ is selected from a given dictionary $K := \{\kappa^1, \dots, \kappa^M\}$. Then the ensemble learner combines all GP priors as:

$$f \sim \sum_{m=1}^M w_0^m \mathcal{GP}\left(0, \kappa^m\left(\mathbf{x}, \mathbf{x}'\right)\right) \ , \tag{4}$$

where w_0^m captures the significance of model m in the ensemble and $\sum_{m=1}^M w_0^m = 1$. Leveraging the sum-product rule, it can be shown that the function posterior pdf of the ensemble learner can be expressed as:

$$p\left(f(\mathbf{x}) \mid \mathbf{y}_t; \mathbf{X}_t\right) = \sum_{m=1}^M w_t^m p\left(f(\mathbf{x}) \mid m, \mathbf{y}_t; \mathbf{X}_t\right) \ , \tag{5}$$

where $w_t^m := \Pr(m|\mathbf{y}_t; \mathbf{X}_t)$. Next, we will introduce the random feature (RF) approximation that offers online model updates, which are particularly appealing in AL scenarios where data arrive and are processed on-the-fly.

RF-based approximation. To efficiently update the EGP function model as new labeled data become available, a parametric function approximant based on random features (RFs) is employed. When dealing with shift-invariant kernels, the RF-based approximation uses the kernel approximant $\check{\kappa}(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}_\mathbf{v}^\top(\mathbf{x})\boldsymbol{\phi}_\mathbf{v}(\mathbf{x}')$, where $\boldsymbol{\phi}_\mathbf{v}(\mathbf{x})$ is the RF vector defined as [33]:

$$\boldsymbol{\phi}_\mathbf{v}(\mathbf{x}) := \frac{1}{\sqrt{D}}\left[\sin(\mathbf{v}_1^\top\mathbf{x}), \cos(\mathbf{v}_1^\top\mathbf{x}), \dots, \sin(\mathbf{v}_D^\top\mathbf{x}), \cos(\mathbf{v}_D^\top\mathbf{x})\right]^\top, \tag{6}$$

and $\{\mathbf{v}\}_{i=1}^D$ are drawn from the power spectral density $\pi_\kappa(\mathbf{v})$ of the standardized kernel κ/σ_θ^2. The RF vector allows for a parametric linear function approximant:

$$\check{f}(\mathbf{x}) = \boldsymbol{\phi}_\zeta^\top(\mathbf{x})\boldsymbol{\theta}, \quad \boldsymbol{\theta} \sim \mathcal{N}\left(\boldsymbol{\theta}; \mathbf{0}_{2D}, \sigma_\theta^2 \mathbf{I}_{2D}\right), \tag{7}$$

that allows for online model updates in a recursive Bayes fashion as shown next.

RF-based parametric EGP model. Adopting the RF approximation for each GP model $m \in \mathcal{M}$ in the ensemble as in (7), yields the following generative model:

$$\begin{aligned} p\left(\boldsymbol{\theta}^m\right) &= \mathcal{N}\left(\boldsymbol{\theta}^m; \mathbf{0}_{2D}, \sigma_{\theta^m}^2 \mathbf{I}_{2D}\right), \\ p\left(f\left(\mathbf{x}_n\right) \mid m, \boldsymbol{\theta}^m\right) &= \delta\left(f\left(\mathbf{x}_n\right) - \boldsymbol{\phi}_\mathbf{v}^{m\top}\left(\mathbf{x}_n\right)\boldsymbol{\theta}^m\right), \\ p\left(y_n \mid \boldsymbol{\theta}^m, \mathbf{x}_n\right) &= \mathcal{N}\left(y_n; \boldsymbol{\phi}_\mathbf{v}^{m\top}\left(\mathbf{x}_n\right)\boldsymbol{\theta}^m, \sigma_n^2\right). \end{aligned} \tag{8}$$

Then, the ensemble learner combines the posterior pdf of all GP models yielding the Gaussian mixture (GM):

$$p(f(\mathbf{x})|L_t) = \sum w_t^m \mathcal{N}(f(\mathbf{x}); \mu_t^m(\mathbf{x}), (\sigma_t^m(\mathbf{x}))^2), \tag{9}$$

where $\mathcal{N}(f(\mathbf{x}); \mu_t^m(\mathbf{x}), (\sigma_t^m(\mathbf{x}))^2)$ is the per model GP posterior pdf with:

$$\mu_t^m(\mathbf{x}) = \boldsymbol{\phi}_\mathbf{v}^{m\top}(\mathbf{x})\hat{\boldsymbol{\theta}}_t^m, \tag{10a}$$

$$(\sigma_t^m(\mathbf{x}))^2 = \boldsymbol{\phi}_\mathbf{v}^{m\top}(\mathbf{x})\boldsymbol{\Sigma}_t^m\boldsymbol{\phi}_\mathbf{v}^m(\mathbf{x}) \ . \tag{10b}$$

Using the minimum-mean square error (MMSE) estimator of the GM in (9) yields the ensemble predictor along with the corresponding variance as follows:

$$\mu_t(\mathbf{x}) = \sum_{m=1}^M w_t^m \mu_t^m(\mathbf{x}), \tag{11a}$$

$$(\sigma_t(\mathbf{x}))^2 = \sum_{m=1}^M w_t^m[(\sigma_t^m(\mathbf{x}))^2 + (\mu_t(\mathbf{x}) - \mu_t^m(\mathbf{x}))^2]. \tag{11b}$$

RF-based EGP model update. Upon optimizing a certain AF that relies on the aforementioned EGP model, $\mathbf{x}_{t+1}$ is obtained and the oracle reveals the corresponding label y_{t+1}. Then the weight of each GP model $w_{t+1}^m := \Pr(m|L_{t+1})$ is updated via

Authorized licensed use limited to: University of Ioannina. Downloaded on May 31,2025 at 19:07:31 UTC from IEEE Xplore. Restrictions apply.

Bayes rule and can be expressed as [18]:

$$w_{t+1}^m = \frac{w_t^m \mathcal{N}\left(y_{t+1}; \hat{y}_{t+1|t}^m, (\sigma_{t+1|t}^m)^2\right)}{\sum_{m'=1}^M w_t^{m'} \mathcal{N}\left(y_{t+1}; \hat{y}_{t+1|t}^{m'}, (\sigma_{t+1|t}^{m'})^2\right)} , \quad (12)$$

where $\mathcal{N}(y_{t+1}; \hat{y}_{t+1|t}^m, (\sigma_{t+1|t}^m))^2$ denotes the per-GP model predictive pdf of y_{t+1} with:

$$\hat{y}_{t+1|t}^m = \boldsymbol{\phi}_{\mathbf{v}}^{m\top}(\mathbf{x}_{t+1})\hat{\boldsymbol{\theta}}_t^m , \quad (13a)$$

$$(\sigma_{t+1|t}^m)^2 = \boldsymbol{\phi}_{\mathbf{v}}^{m\top}(\mathbf{x}_{t+1})\boldsymbol{\Sigma}_t^m \boldsymbol{\phi}_{\mathbf{v}}^m(\mathbf{x}_{t+1}) + \sigma_n^2 \quad (13b)$$

Meanwhile, the posterior of $\boldsymbol{\theta}^m$ of each model m is updated via Bayes rule as:

$$p(\boldsymbol{\theta}^m|L_{t+1}) = \mathcal{N}(\boldsymbol{\theta}^m; \hat{\boldsymbol{\theta}}_{t+1}^m, \boldsymbol{\Sigma}_{t+1}^m) , \quad (14)$$

where the mean $\hat{\boldsymbol{\theta}}_{t+1}^m$ and covariance matrix $\boldsymbol{\Sigma}_{t+1}^m$ are:

$$\hat{\boldsymbol{\theta}}_{t+1}^m = \hat{\boldsymbol{\theta}}_t^m + (\sigma_{t+1|t}^m)^{-2}\boldsymbol{\Sigma}_t^m \boldsymbol{\phi}_{\mathbf{v}}^m(\mathbf{x}_{t+1})(y_{t+1} - \hat{y}_{t+1|t}^m) , \quad (15a)$$

$$\boldsymbol{\Sigma}_{t+1}^m = \boldsymbol{\Sigma}_t^m - (\sigma_{t+1|t}^m)^{-2}\boldsymbol{\Sigma}_t^m \boldsymbol{\phi}_{\mathbf{v}}^m(\mathbf{x}_{t+1})\boldsymbol{\phi}_{\mathbf{v}}^{m\top}(\mathbf{x}_{t+1})\boldsymbol{\Sigma}_t^m . \quad (15b)$$

Next, we will outline a number of EGP uncertainty-based AFs used to select new unlabeled instances to label at each iteration of the AL process.

F. Acquisition rules for EGP-based AL

Using the EGP posterior in (9), we present the following uncertainty-based criteria.

1) Weighted variance. The first Acquisition Function (AF) utilizes the uncertainty expressed by the variance. With GP expert m contributing to the function posterior with variance $(\sigma_t^m(x))^2$, the weighted combination of all M models yields the AF:

$$\alpha^{\text{wVar}}(\mathbf{x}; L_t) := \sum_{m=1}^M w_t^m (\sigma_t^m(\mathbf{x}))^2 . \quad (16)$$

2) Weighted entropy. Alternatively relying on entropy as the uncertainty measure, a weighted sum of the entropy values of all M GP models yields:

$$\alpha^{\text{wEnt}}(\mathbf{x}; L_t) := \frac{1}{2}\sum_{m=1}^M w_t^m \log(\sigma_t^m(\mathbf{x})^2). \quad (17)$$

3) Variance of GP mixtures. Instead of using a weighted combination of all GP model variances, one can directly use the variance of the GM in (11b) as an AF to guide the AL process.

Remark. The EGP model updates in Sec. IIE and the corresponding AFs in the present section are specifically tailored for the *regression task*, since the gold standard FFR index to be predicted does *not* belong to a finite alphabet that pertains to the classification task.

III. RESULTS

We considered a small labeled set ($L_0 = 20$) and a larger set of unlabeled data ($U_0 = 130$). We selected 37 data for testing. We employed the EGP regression model, which incorporates the intuitive uncertainty-based acquisition criteria in Sec. IIF yielding the "EGP_Var", "EGP_wVar" and "EGP_wEnt" approaches. A performance evaluation of the advocated EGP-AL methods on the SMARTool dataset has been carried out. As a baseline, we conducted a comparison using a single "GP_var" model that utilizes the maximum variance criterion of a single GP with a pre-selected kernel, as well as the "EGP_Random", which considers random sampling from the unlabeled data. In all approaches, we utilized the initially labeled data to estimate the kernel hyperparameters for each GP expert by maximizing the marginal likelihood. For the RF-based GPs, we set $D = 50$. The EGP-based approaches employed a kernel dictionary K consisting of 11 radial basis functions (RBFs) with characteristic lengthscales chosen from $\{10^c\}_{c=-4}^6$. All methods underwent 25 iterations, and their average performance, along with the corresponding standard deviation, is reported.

Figure 1 illustrates the Normalized Mean Squared Error (NMSE) performance of all competing approaches, defined similarly as in [18]. It is evident that all advocated EGP-based AL methods exhibit superior performance compared to the "EGP_Random" approach [21], [29], showcasing the benefits of the well-motivated acquisition criteria. It is worth mentioning that all EGP-based AL methods upon iteration 15 achieve comparable or even better NMSE performance than that of "EGP_Random" at iteration 25, demonstrating their ability to achieve satisfactory performance with less labeled data compared to EGPs with passively selected data.

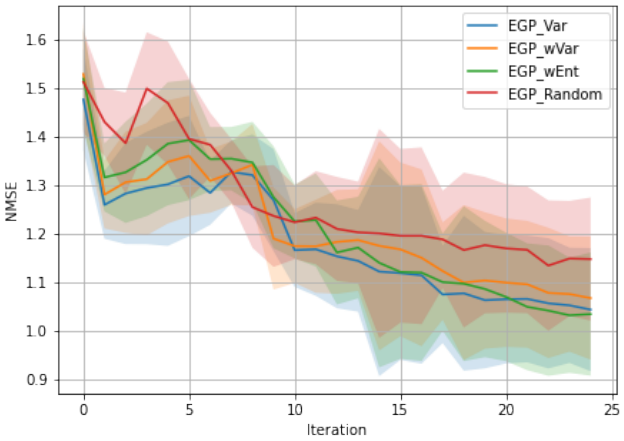

Fig. 1: NMSE perfomance on SMARToll dataset.

To further account for the associated uncertainty besides the accuracy of predictions, Figure 2 depicts the Negative Log Likelihood (NLL) of all competing approaches, defined similarly as in [28]. It is evident that all EGP-AL approaches enjoy the lowest NLL, clearly indicating their notable advantages over the "EGP_Random" approach.

Authorized licensed use limited to: University of Ioannina. Downloaded on May 31,2025 at 19:07:31 UTC from IEEE Xplore. Restrictions apply.

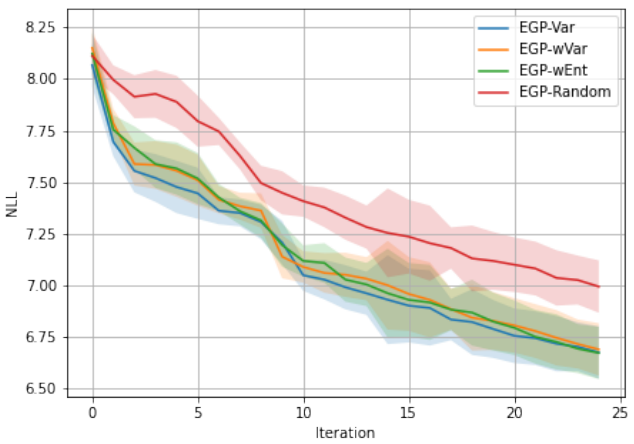

Fig. 2: NLL perfomance on SMARToll dataset

Also note that we have additionally compared the EGP-based AL methods with the single GP_Var baseline [32], with the latter showing a substantially inferior performance (average NMSE $>$ 3.1 and NLL $>$ 13.5 in all iterations, and that is the reason it was omitted in Fig. 1 and Fig. 2 respectively). This highlights the advantages of using EGP models with adaptive weights as new data become available.

IV. DISCUSSION AND CONCLUSIONS

The task of predicting and diagnosing cardiovascular diseases (CVD) poses significant challenges. Driven by the need to protect the confidentiality of patient's medical records in healthcare fields, the disclosure of labels is frequently unattainable. In light of that, AL-based approaches offer a principled solution by prudently selecting the most informative instances to label from an unlabeled set, to efficiently and effectively infer the FFR index of different patients. This work is based on an ensemble (E) of GPs, which enhances the expressiveness of the model by utilizing unique GP models with distinct kernel functions. The RF-based approximation facilitates efficient online updates of the EGP function model in a recursive Bayes manner. This framework allows for adaptive learning and uncertainty quantification in AL settings.

The advocated EGP-based AL methods focus on efficiently assessing the severity of coronary artery disease in patients undergoing CTCA for treatment selection (revascularization vs. medical therapy). Our work stands out by accurately predicting the FFR index with limited labeled data, showing a major impact in clinical cases considering the challenges posed by data limitations, confidentiality, and sampling costs. For future directions, a graph-adaptive setup can be considered, where the input is a graph capturing the similarity among patients, as demonstrated in [28].

REFERENCES

[1] P. K. Siogkas, L. Lakkas, A. I. Sakellarios, G. Rigas, S. Kyriakidis, K. A. Stefanou, C. D. Anagnostopoulos, A. Clemente, S. Rocchiccioli, G. Pelosi *et al.*, "Smartffr, a new functional index of coronary stenosis: comparison with invasive ffr data," *Frontiers in Cardiovascular Medicine*, vol. 8, p. 714471, 2021.

[2] P. S. Douglas, G. Pontone, M. A. Hlatky, M. R. Patel, B. L. Norgaard, R. A. Byrne, N. Curzen, I. Purcell, M. Gutberlet, G. Rioufol *et al.*, "Clinical outcomes of fractional flow reserve by computed tomographic angiography-guided diagnostic strategies vs. usual care in patients with suspected coronary artery disease: the prospective longitudinal trial of ffrct: outcome and resource impacts study," *European heart journal*, vol. 36, no. 47, pp. 3359–3367, 2015.

[3] A. M. Alaa, T. Bolton, E. Di Angelantonio, J. H. Rudd, and M. Van der Schaar, "Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 uk biobank participants," *PloS one*, vol. 14, no. 5, p. e0213653, 2019.

[4] R. G. Nadakinamani, A. Reyana, S. Kautish, A. Vibith, Y. Gupta, S. F. Abdelwahab, A. W. Mohamed *et al.*, "Clinical data analysis for prediction of cardiovascular disease using machine learning techniques," *Computational intelligence and neuroscience*, vol. 2022, 2022.

[5] S. Subramani, N. Varshney, M. V. Anand, M. E. M. Soudagar, L. A. Al-Keridis, T. K. Upadhyay, N. Alshammari, M. Saeed, K. Subramanian, K. Anbarasu *et al.*, "Cardiovascular diseases prediction by machine learning incorporation with deep learning," *Frontiers in Medicine*, vol. 10, p. 1150933, 2023.

[6] L. J. Laslett, P. Alagona, B. A. Clark, J. P. Drozda, F. Saldivar, S. R. Wilson, C. Poe, and M. Hart, "The worldwide environment of cardiovascular disease: Prevalence, diagnosis, therapy, and policy issues," *Journal of the American College of Cardiology*, vol. 60, no. 25_Supplement, pp. S1–S49, 2012. [Online]. Available: https://www.jacc.org/doi/abs/10.1016/j.jacc.2012.11.002

[7] M. Humbert-Droz, P. Mukherjee, O. Gevaert *et al.*, "Strategies to address the lack of labeled data for supervised machine learning training with electronic health records: case study for the extraction of symptoms from clinical notes," *JMIR Medical Informatics*, vol. 10, no. 3, p. e32903, 2022.

[8] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Á. Fernández-Leal, "Human-in-the-loop machine learning: A state of the art," *Artificial Intelligence Review*, vol. 56, no. 4, pp. 3005–3054, 2023.

[9] F. Olsson, "A literature survey of active machine learning in the context of natural language processing," 2009.

[10] A. Pandi, C. Diehl, A. Yazdizadeh Kharrazi, S. A. Scholz, E. Bobkova, L. Faure, M. Nattermann, D. Adam, N. Chapin, Y. Foroughijabbari *et al.*, "A versatile active learning workflow for optimization of genetic and metabolic networks," *Nature Communications*, vol. 13, no. 1, p. 3876, 2022.

[11] I. M. El-Hasnony, O. M. Elzeki, A. Alshehri, and H. Salem, "Multi-label active learning-based machine learning model for heart disease prediction," *Sensors*, vol. 22, no. 3, p. 1184, 2022.

[12] P. Yuan, "Research on predicting heart attack through active learning," in *2021 2nd International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*. IEEE, 2021, pp. 12–18.

[13] F. Khozeimeh, R. Alizadehsani, M. Shirani, M. Tartibi, A. Shoeibi, H. Alinejad-Rokny, C. Harlapur, S. J. Sultanzadeh, A. Khosravi, S. Nahavandi *et al.*, "Alec: Active learning with ensemble of classifiers for clinical diagnosis of coronary artery disease," *Computers in Biology and Medicine*, vol. 158, p. 106841, 2023.

[14] P. Kumar and A. Gupta, "Active learning query strategies for classification, regression, and clustering: a survey," *Journal of Computer Science and Technology*, vol. 35, pp. 913–945, 2020.

[15] A. Tharwat and W. Schenck, "A survey on active learning: State-of-the-art, practical challenges and research directions," *Mathematics*, vol. 11, no. 4, p. 820, 2023.

[16] C. Riis, F. Antunes, F. Hüttel, C. Lima Azevedo, and F. Pereira, "Bayesian active learning with fully bayesian gaussian processes," *Advances in Neural Information Processing Systems*, vol. 35, pp. 12 141–12 153, 2022.

[17] A. Olmin, J. Lindqvist, L. Svensson, and F. Lindsten, "Active learning with weak labels for gaussian processes," *arXiv preprint arXiv:2204.08335*, 2022.

[18] K. D. Polyzos, Q. Lu, and G. B. Giannakis, "Weighted ensembles for active learning with adaptivity," 2022.

[19] ——, "Graph-adaptive incremental learning using an ensemble of gaussian process experts," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5220–5224.

Authorized licensed use limited to: University of Ioannina. Downloaded on May 31,2025 at 19:07:31 UTC from IEEE Xplore. Restrictions apply.

[20] ——, "Online graph-guided inference using ensemble gaussian processes of egonet features," in *2021 55th Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2021, pp. 182–186.

[21] ——, "Ensemble gaussian processes for online learning over graphs with adaptivity and scalability," *IEEE Transactions on Signal Processing*, vol. 70, pp. 17–30, 2021.

[22] Q. Lu and K. D. Polyzos, "Gaussian process dynamical modeling for adaptive inference over graphs," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

[23] K. D. Polyzos, Q. Lu, A. Sadeghi, and G. B. Giannakis, "On-policy reinforcement learning via ensemble Gaussian processes with application to resource allocation," *Proc. Asilomar Conf. Sig., Syst., Comput.*, pp. 1018–1022, 2021.

[24] M. K. Singh, K. D. Polyzos, P. A. Traganitis, S. V. Dhople, and G. B. Giannakis, "Physics-informed transfer learning for voltage stability margin prediction," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[25] K. D. Polyzos, Q. Lu, and G. B. Giannakis, "Bayesian optimization with ensemble learning models and adaptive expected improvement," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

[26] Q. Lu, K. D. Polyzos, B. Li, and G. B. Giannakis, "Surrogate modeling for bayesian optimization beyond a single gaussian process," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 11 283–11 296, 2023.

[27] A. Bacharis, K. D. Polyzos, H. J. Nelson, G. B. Giannakis, and N. Papanikolopoulos, "3d reconstruction in noisy agricultural environments: A bayesian optimization perspective for view planning," *arXiv preprint arXiv:2310.00145*, 2023.

[28] K. D. Polyzos, Q. Lu, and G. B. Giannakis, "Active sampling over graphs for bayesian reconstruction with gaussian ensembles," in *2022 56th Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2022, pp. 58–64.

[29] S. C. Tassi, V. Kigka, P. Siogkas, S. Rocchiccioli, G. Pelosi, D. I. Fotiadis, and A. I. Sakellarios, "Graph-guided gaussian process-based diagnosis of cvd severity with uncertainty measures," *Proccedings of the 45th EMBC Conference*, 2023.

[30] A. I. Sakellarios, P. Tsompou, V. Kigka *et al.*, "Non-invasive prediction of site-specific coronary atherosclerotic plaque progression using lipidomics, blood flow, and ldl transport modeling," *Applied Sciences*, vol. 11, no. 5, p. 1976, 2021.

[31] C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006.

[32] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell, "Active learning with Gaussian processes for object categorization," *Proc. Intl. Conf. Comp. Vision*, 2007.

[33] M. Lázaro-Gredilla, J. Quiñonero Candela, C. E. Rasmussen, and A. Figueiras-Vidal, "Sparse spectrum Gaussian process regression," vol. 11, no. Jun, pp. 1865–1881, 2010.

Authorized licensed use limited to: University of Ioannina. Downloaded on May 31,2025 at 19:07:31 UTC from IEEE Xplore. Restrictions apply.