Trustworthy Prediction with Gaussian Process Knowledge Scores

Kurt Butler,† Guanchao Feng,‡ Tong Chen,‡ and Petar M. Djurić‡
†*School of Engineering*, *The University of Edinburgh*, Edinburgh, UK
‡*Department of Electrical and Computer Engineering*, *Stony Brook University*, Stony Brook, NY, USA
kbutler2@ed.ac.uk, {guanchao.feng, tong.chen, petar.djuric}@stonybrook.edu

***Abstract*—Probabilistic models are often used to make predictions in regions of the data space where no observations are available, but it is not always clear whether such predictions are well-informed by previously seen data. In this paper, we propose a knowledge score for predictions from Gaussian process regression (GPR) models that quantifies the extent to which observing data have reduced our uncertainty about a prediction. The knowledge score is interpretable and naturally bounded between 0 and 1. We demonstrate in several experiments that the knowledge score can anticipate when predictions from a GPR model are accurate, and that this anticipation improves performance in tasks such as anomaly detection, extrapolation, and missing data imputation. Source code for this project is available online at https://github.com/KurtButler/GP-knowledge.**

***Index Terms*—anomaly detection, Gaussian processes, regression models, trustworthy machine learning, predictive distributions.**

I. Introduction

The task of prediction is of fundamental importance in many domains. Regression models are designed to predict the values of a function at new locations. In anomaly detection [1], [2], a predictive distribution is used to decide if a newly received set of data contains abnormal samples. In data interpolation problems [3], the goal is to predict data that have been censored, corrupted, or are otherwise missing. While methods that rely on the ideas of prediction are widely deployed, they all require the common assumption that the predictive model is well-informed about what it intends to predict. Although common measures of accuracy on training or testing datasets help build confidence in the model, they fail to address a fundamental issue: the model's predictive reliability depends on where a given input lies within the input space, as shown in Fig. 1. This is evident in extrapolation problems, where the predictive confidence of the model decays as one strays farther out of distribution from the training data.

Recent work has noted that while a classification or regression method can produce a prediction, such predictions are not always well informed by training data [4]. Extrapolation problems [5], as noted above, represent one class of problems where this consideration occurs naturally. Another example is the problem of open-set recognition [6], [7], where new samples that are far out-of-distribution (OOD) are suspected to be new, unseen classes, and so the classifier should withhold assigning the label of a known class to these new samples. These works highlight the growing interest in quantifying the trustworthiness of a model's prediction.

This work was supported by NSF under Award 2212506. We also acknowledge support of the UKRI AI programme, and the Engineering and Physical Sciences Research Council, for CHAI - Causality in Healthcare AI Hub [grant number EP/Y028856/1].

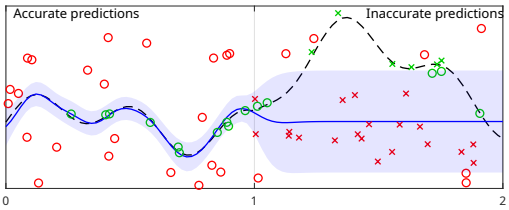

Fig. 1: Anomaly detection with a probabilistic model of the form $y = F(x) + noise$. The model is trained on a set of training data for values $x \leq 1$ (not pictured). The symbol $\times$ represents a point incorrectly identified as an anomaly or normal, and $\circ$ represents a correct assessment. Normal points and anomalies are shown in green and red respectively. The model is well-informed when $x \leq 1$, and the model becomes less useful for $x > 1$ since it lacks data in this region. The problem is to determine the boundary between regions using the model alone.

Recent works have proposed various methods to quantify the trustworthiness of an OOD prediction. Uncertainty quantification frameworks such as conformal prediction [8] or Bayesian credible intervals [9] can assess trustworthiness by the size of a predictive interval. Other approaches define mathematical scores that assess whether a classifier's prediction is OOD by examining ratios of distances between nearby classes [4], [7]. These approaches come with their own pitfalls, and the problem is far from solved in general.

In this paper, we introduce the concept of knowledge score, which can be applied to any probabilistic machine learning (ML) model. This score is derived from uncertainty quantification and, under certain conditions, is bounded between zero and one. The knowledge score quantifies the extent to which a prediction from a probabilistic ML model is informed by the training data. A high score indicates that the model is confident about the expected behavior at a given location, while a low score suggests that the model relies mostly on the prior distribution, suggesting that a prediction is largely uninformed by the previously seen data. In this way, the knowledge score provides a principled measure of how much

trust we should place in the model's predictions.

In this paper, we address the knowledge score of Gaussian processes (GPs) in the context of Gaussian process regression (GPR). Our experiments show that the knowledge score can be applied to many problems, including extrapolation, interpolation, and anomaly detection.

The paper is organized as follows. In Section II, we formulate the problem we address. Section III describes the methodology for estimating the reliability of a GPR prediction, based on the concept of a knowledge score. The following section presents experimental results that demonstrate the application of the knowledge score to synthetic and real data. Finally, Section V provides concluding remarks.

II. Problem Formulation

We begin by recalling the general framework of Gaussian process regression (GPR) as motivation for our proposed knowledge score. Consider a regression model of the form,

$$y = F(\mathbf{x}) + \varepsilon, \tag{1}$$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2), \tag{2}$$

where $\mathbf{x} \in \mathbb{R}^D, y \in \mathbb{R}$, F is an unknown nonlinear function and ε is white Gaussian noise with variance σ^2. To estimate F from data, we first place a GP prior over functions,

$$F \sim \mathcal{GP}(0, k), \tag{3}$$

where k is a kernel or covariance function. We assume that we have previously seen a data set $\mathcal{D}$ of input-output pairs,

$$\mathcal{D} = \{(\mathbf{x}_n, y_n) : n = 1, ..., N\}. \tag{4}$$

Given prior observations at the locations $\mathbf{x}_n$, our predictive task requires us to predict the distribution at a new location $\mathbf{x}$. By conditioning on $\mathcal{D}$, one can obtain a GP posterior distribution over F, i.e., we have

$$F|\mathcal{D} \sim \mathcal{GP}(m_p, k_p), \tag{5}$$

$$m_p(\mathbf{x}) = \mathbf{k}(\mathbf{x})^\top (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{y}, \tag{6}$$

$$k_p(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - \mathbf{k}(\mathbf{x})^\top (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{k}(\mathbf{x}'). \tag{7}$$

where $\mathbf{y} = [y_1, \cdots, y_N]^\top$, and the vector $\mathbf{k}(\mathbf{x})$ and the matrix $\mathbf{K}$ are defined by

$$\mathbf{k}(\mathbf{x}) = \begin{bmatrix} k(\mathbf{x}, \mathbf{x}_1) \\ \vdots \\ k(\mathbf{x}, \mathbf{x}_N) \end{bmatrix}, \mathbf{K} = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \cdots & k(\mathbf{x}_N, \mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_1, \mathbf{x}_N) & \cdots & k(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix}.$$

This posterior distribution over F also induces a corresponding distribution over the observations $y = F(\mathbf{x}) + \varepsilon$, which takes the form:

$$y_*|\mathbf{x}_*, \mathcal{D} \sim \mathcal{N}(m_p(\mathbf{x}_*), k_p(\mathbf{x}_*, \mathbf{x}_*) + \sigma^2), \tag{8}$$

and is used to infer the parameters of the GP kernel by maximizing the likelihood of $\mathcal{D}$.

In (7), it can be seen that the posterior variance is strictly upper bounded by the prior variance:

$$k(\mathbf{x}, \mathbf{x}) - k_p(\mathbf{x}, \mathbf{x}) = \mathbf{k}(\mathbf{x})^\top (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{k}(\mathbf{x}) \geq 0. \tag{9}$$

This observation leads to the development of the knowledge score in the following section.

III. Methodology

We now formulate the problem of quantifying model knowledge as a comparison of the prior distribution $p(F|\mathbf{x})$ and the posterior $p(F|\mathbf{x}, \mathcal{D})$. We begin with a formal definition that defines the *knowledge score* in terms of a general principle,

$$G(\mathbf{x}, \mathcal{D}) \triangleq \frac{\text{Var}(F(\mathbf{x})|\mathbf{x}) - \text{Var}(F(\mathbf{x})|\mathbf{x}, \mathcal{D})}{\text{Var}(F(\mathbf{x})|\mathbf{x})}. \tag{10}$$

The function $G(\mathbf{x}, \mathcal{D})$ quantifies the extent to which conditioning on the data set $\mathcal{D}$ reduces the variance of the prediction $F(\mathbf{x})$ at location $\mathbf{x}$. Thus, we may also refer to G as a variance reduction score.[1] A key insight of our approach is that the trustworthiness of a prediction depends critically on the location in the input space relative to the previously seen data.

In the case of a GPR model, the function G in (10) becomes highly interpretable. From the definition and (9), it can be seen that G is bounded between 0 and 1 for GPR models. A value near 0 indicates that the GP is not knowledgeable at the given location, in the sense that the prior and posterior distributions are nearly identical. Conversely, a value of G near 1 indicates that conditioning on $\mathcal{D}$ has greatly reduced our uncertainty about $F(\mathbf{x})$. Since posterior variances in GPR admit closed-form solutions, as shown in (7), the function G can be expressed as

$$G(\mathbf{x}, \mathcal{D}) = \frac{\mathbf{k}(\mathbf{x})^\top (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{k}(\mathbf{x})}{k(0, 0)}. \tag{11}$$

From (11), it is apparent that besides $\mathcal{D}$ and $\mathbf{x}$, the choice of k also affects the knowledge score. Simple kernel functions such as the radial basis function (RBF) kernel [10] are directly tied to the distances of each $\mathbf{x}_n$ in $\mathcal{D}$ to the new location $\mathbf{x}_*$. Learning the hyperparameters of the kernel can therefore be closely connected to metric learning [11]. Other kernels, such as periodic or linear kernels, or compositions of multiple kernels, can have more complex behavior and lead to less intuitive patterns in the knowledge score. For example, when modeling time series, a kernel that combines a periodic component with an RBF kernel may be used to forecast future data. In such cases, the GP knowledge score provides a principled way to understand how the combination of these two kernels affects the confidence of predictions during extrapolation.

We now discuss properties of the knowledge score, and its applications to problems such as extrapolation, interpolation, and anomaly detection.

A. Properties of the Knowledge Score

Several properties of the knowledge score are worth noting. First, the score is naturally normalized and defined in terms of an interpretable quantity (variance reduction). Also, knowledge scores are agnostic to the distribution of the model inputs,

[1] In this work, we focus on the variance reduction score. Entropy-based alternatives are left for future work.

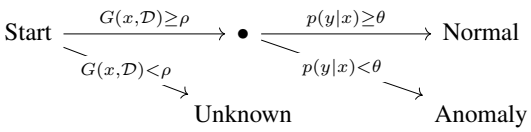

Fig. 2: Flowchart describing a two-stage anomaly detection scheme using the knowledge score. The symbols ρ and θ represent thresholds.

$p(\mathbf{x})$ and do not require an observed output sample y_* when making predictions about $F(\mathbf{x}_*)$. From (10), it is evident that the knowledge score is invariant under normalization of the output variable.

A minor limitation is that the knowledge score is bounded between 0 and 1 for Gaussian models. In general, G can be negative, which would indicate that observing $\mathcal{D}$ has *increased* the uncertainty of a prediction. In such cases, G may still provide useful information, though its interpretation may be different. In this work, we restrict our focus to GPR models.

B. Improving Anomaly Detection with Knowledge Scores

GPs are natural tools for anomaly detection [12]–[14]. Knowledge scores can be useful for withholding classification of new data when the GP model is uninformed by the training data.

To understand how knowledge scores can help with this task, we recall a typical formulation of the anomaly detection problem in the GPR framework. Given the posterior predictive distribution in (8), one can test whether a novel observation pair $(\mathbf{x}_*, y_*)$ is consistent with the previously observed data. If y_* lies far outside the distribution $p(y_*|\mathbf{x}, \mathcal{D})$, then the new pair $(\mathbf{x}_*, y_*)$ may be marked as an outlier, because it is inconsistent with the previously observed data.

To formalize the detection of outliers, we use the following decision rule, which exploits Gaussianity of the posterior predictive distribution. We declare a new pair $(\mathbf{x}_*, y_*)$ to be an anomaly if

$$|y_* - \mathbb{E}(F(\mathbf{x}_*)|\mathbf{x}_*, \mathcal{D})| > \theta(\mathbf{x}_*, \mathcal{D}), \tag{12}$$

where the expectation is taken w.r.t. $F \sim p(F|\mathcal{D})$, and θ is a threshold determined by some criterion. A common heuristic is to set

$$\theta(\mathbf{x}_*, \mathcal{D}) = 3\sqrt{\mathrm{Var}(y_*|\mathbf{x}_*, \mathcal{D})},$$

which asserts that y_* is an anomaly if the probability of a more extreme value is very low (less than 0.3%). Other principles may be used to select θ, although for Gaussian distributions, other principles typically yield tests of the form in (12).

For a new sample $(\mathbf{x}_*, y_*)$, one can determine if the pair is anomalous using an anomaly detection approach such as the one described above. However, this approach implicitly assumes that $p(y_*|\mathbf{x}_*, \mathcal{D})$ is well-informed by the data set $\mathcal{D}$. Even if $\mathcal{D}$ contains many clean input-output pairs that allow us to learn a model with strong predictive performance on the training set, it is possible and sometimes inevitable that the model encounters difficulties when extrapolating at new input locations $\mathbf{x}_*$ that are unfamiliar relative to those in $\mathcal{D}$. One remedy for this issue is to use models that inflate the posterior predictive variance in such regions. However, artificially increasing the predictive variance can also lead to a higher type I error (miss rate) in the anomaly detector.

An alternative solution is to refuse to classify points for which the model is not confident, using a two-stage classification approach. We propose one version of this in Fig. 2, where the knowledge score is used to determine when the class should be considered unknown. This first stage consists of a comparison of the knowledge score $G(\mathbf{x}, \mathcal{D})$ to a threshold parameter ρ, where points are classified as unknown when $G < \rho$. The second stage performs anomaly detection only on points for which $G \geq \rho$. Our empirical results suggest that the two-stage approach can be useful to identify subsets of data where the anomaly detection is performed with high accuracy. The question of how to select ρ is left to future work, although a simple approach is to select ρ such that the empirical performance of the anomaly detector, evaluated on a given training set of data, satisfies some problem-specific criteria.

C. Improving Extrapolation/Interpolation with Knowledge Scores

Both extrapolation and interpolation are examples of problems where we wish to estimate the value of a function at a location where we have not received data. In the extrapolation problem, we want to predict values at locations that lie outside the distribution of the input points [5].

In interpolation problems, there are locations where data are missing and we seek to infer the value of the underlying function [3]. The missing data segments can vary in length, but typically there are observed data on either side of each segment. When a missing segment is short, many standard methods may be used to interpolate the function. However, as the length of the missing segment becomes larger, it becomes less clear whether the model is well-equipped to interpolate accurately. Thus, a critical issue is to determine how long is too long to perform interpolation.

IV. Experiments

In this section, we demonstrate how GP knowledge scores can be applied to several problems of common interest.

A. Trustworthy Anomaly Detection on Toy Data

In this experiment, we investigate how knowledge scores can be used to improve anomaly detection. Data are generated according to a model of the form (1), where x_n is randomly sampled from the range 0 to 1. Anomalies were generated independently according to a large variance distribution, $\mathcal{U}(-5, 5)$, for a small randomly selected minority of points. A total of 50 points were used for training. The training data were split as normal and anomalous using the GP mixture model described previously. The performance of the GPR model for

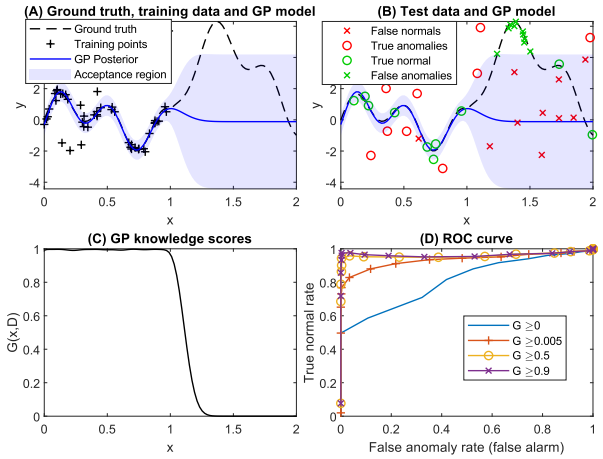

Fig. 3: **(A)** Ground truth function, and training data used to train the GPR model for anomaly detection. **(B)** The GPR model and a subset of the test set predictions. **(C)** The GP knowledge score $G(x, \mathcal{D})$ viewed as a function of the input location x. **(D)** Receiver-operating characteristic (ROC) curve showing the performance of the anomaly detector on subsets of the data. If we restrict ourselves to only analyzing points x_n where $G(x_n, \mathcal{D})$ exceeds some nonzero threshold, we observe that the performance of the model increases.

anomaly detection was then evaluated on 1000 test points. The use of the knowledge score is visualized in Fig. 3.

In the region $x > 1$, the posterior predictive variance of the GPR grows rapidly, indicating that the model is not informed about the test points in that region. The knowledge score quantifies the growth of the variance in this region, by examining how much the posterior and prior variance differ. In the two-stage anomaly detection scheme of Sec. III-B, we propose that the knowledge score can classify the points as "unknown" when the predictive model is not confident enough to classify them as anomalous or normal. The performance of the anomaly detector on the remaining points should then improve. In Fig. 3, we observe that the receiver-operating characteristic (ROC) curves change depending on whether points are included or excluded due to their knowledge scores. As a threshold for G is used to filter out points (i.e., when $\rho > 0$ in the two-stage anomaly detector), we observe that the accuracy of the anomaly detector improves considerably. As a result, the knowledge scores allow us to identify which portions of the test set are likely to be more accurate, even before we observe the test label values.

B. Trustworthy Extrapolation on Electricity Data

In this experiment, we consider the problem of extrapolation using a real data set. We considered the Electricity Load Diagrams data set from the UCI Machine Learning Repository [15], which records the electricity consumption of 370 households in a Portuguese city over several years. With these data, we averaged and normalized the energy consumption across

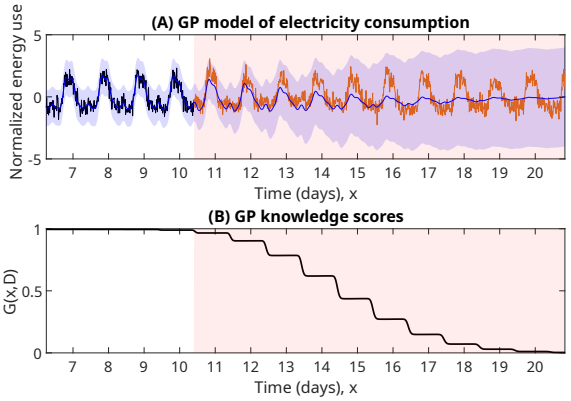

Fig. 4: **(A)** Prediction of average consumption using a GPR model. Black points are observed real data, red points are unobserved (censored) real data, and the blue represents the GPR model. **(B)** GP knowledge scores associated with the predictions. As time increases beyond the point of the data cutoff, the GP knowledge about predictions decreases.

households to yield the energy load pattern for an average house in the city. A GPR model of energy consumption as a function of time was trained using the first 1,000 samples of data. We used a locally periodic kernel [16] for the GPR model, given by the formula,

$$k_{\text{LocPer}}(t, t') = \sigma_f^2 e^{\left(\frac{-2\sin(\pi\rho^{-1}|t-t'|)}{\ell^2}\right)} e^{\left(-\frac{(t-t')^2}{2\ell^2}\right)},$$

where σ_f^2, ρ, and ℓ are hyperparameters. The model was then tasked with extrapolating its predictions several days into the future, as shown in Fig. 4.

It can be observed in the figure that the quality of the prediction degrades as the distance in time from the training data increases. As the distance increases, the posterior predictive distribution of the GP reverts to the prior. The GP knowledge scores track the transition back to the mean distribution. After many days, the predictive variance becomes large enough that it lacks descriptive power, indicating that the extrapolation scheme is only appropriate for a few days. Determining the exact period of time for which extrapolation is useful is a heuristic choice, but the GP knowledge score provides a clear and grounded metric for determining when the predictions have become too vague to be useful.

C. Trustworthy Interpolation on Fetal Heart Rate Recordings

In this experiment, we consider the problem of missing data interpolation using an open-access cardiotocography (CTG) data set [17] in which 552 cardiotocography (CTG) recordings were selected from 9164 recordings acquired between April 2010 and August 2012 at the obstetrics ward of the University Hospital in Brno, Czech Republic. In CTG recordings, both fetal heart rate (FHR) and uterine activity (UA) measurements are usually obtained externally using ultrasound. As a result, various reasons, such as fetal or maternal movements and misplaced electrodes, can cause missing samples for varying

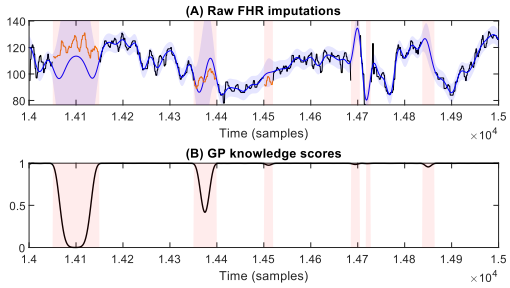

Fig. 5: **(A)** Missing data interpolation with GPR. The red shaded regions represent missing data segments. Some data were manually removed (the red curves) to obtain a ground truth for some missing data segments. **(B)** GP knowledge scores associated with GPR data interpolations. The dip in the GP knowledge score across missing data segments can be useful as a measure of where a missing data segment is too long to warrant interpolation.

periods of time. Since computerized FHR analysis often relies on features extracted from FHR recordings, and many popular FHR features are sensitive to missing samples [18], properly addressing these missing FHR segments is of great importance. In practice, small segments of missing samples are interpolated using linear or cubic spline interpolation, while larger segments are often completely removed [19]. However, there is no consensus on the missing segment length threshold for interpolation. It is worth noting that, similar to missing FHR segments, unreliably recovered/estimated FHR segments can also introduce serious distortions to downstream tasks and analyses. In this section, we show that trust score can be adopted to determine whether or not one should impute missing FHR segments in a principled manner.

In Fig. 5, we visualize a GPR model used to interpolate missing samples from a randomly selected FHR recording. This example contains several missing segments of varying lengths. Longer segments are less reliably interpolated, so a principled metric for determining when a missing data segment should be interpolated is desirable. We see that the GP knowledge score shows a clear distinction in how much the knowledge drops based on the length of the segment. As such, if the GP model is taken to be accurate, then those segments which exhibit only small decreases in knowledge scores can be safely interpolated.

V. Conclusion

In this paper, we introduced the concept of GP knowledge score to quantify the extent to which the GPR posterior distribution is informed by the training data. We applied the knowledge score to a few problems of interest and showed that it can be a powerful tool to understand how models generalize. In future work, we plan to study the relationship between knowledge scores and the empirical performance on specific tasks. We also intend to explore information-theoretic alternatives to our proposed variance reduction score as well as relaxations of the Gaussianity assumption.

References

[1] Biao Wang and Zhizhong Mao, "Outlier detection based on Gaussian process with application to industrial processes," *Applied Soft Computing*, vol. 76, pp. 505–516, 2019.

[2] Alban Siffer, Pierre-Alain Fouque, Alexandre Termier, and Christine Largouet, "Anomaly detection in streams with extreme value theory," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 1067–1075.

[3] Mario Cesarelli, Maria Romano, Paolo Bifulco, Fiammetta Fedele, and Marcello Bracale, "An algorithm for the recovery of fetal heart rate series from CTG data," *Computers in Biology and Medicine*, vol. 37, no. 5, pp. 663–669, 2007.

[4] Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta, "To trust or not to trust a classifier," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[5] Andrew Wilson and Ryan Adams, "Gaussian process kernels for pattern discovery and extrapolation," in *International Conference on Machine Learning*. PMLR, 2013, pp. 1067–1075.

[6] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult, "Toward open set recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1757–1772, 2012.

[7] Pedro R Mendes Júnior, Roberto M De Souza, Rafael de O Werneck, Bernardo V Stein, Daniel V Pazinato, Waldir R De Almeida, Otávio AB Penatti, Ricardo da S Torres, and Anderson Rocha, "Nearest neighbors distance ratio open-set classifier," *Machine Learning*, vol. 106, no. 3, pp. 359–386, 2017.

[8] Xiaofan Zhou, Baiting Chen, Yu Gui, and Lu Cheng, "Conformal prediction: A data perspective," *ACM Computing Surveys*, 2025.

[9] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin, *Bayesian Data Analysis*, Chapman and Hall/CRC, 1995.

[10] Carl Edward Rasmussen and Christopher KI Williams, *Gaussian Processes for Machine Learning*, vol. 2, MIT Press, 2006.

[11] Tong Chen, Guanchao Feng, and Petar M. Djurić, "Improving open-set recognition with Bayesian metric learning," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 6185–6189.

[12] Varun Chandola and Ranga Raju Vatsavai, "A Gaussian process based online change detection algorithm for monitoring periodic time series," in *Proceedings of the 2011 SIAM International Conference on Data Mining*. SIAM, 2011, pp. 95–106.

[13] Liu Yang, Kurt Butler, and Petar M Djurić, "Sequential detection of anomalies in noisy outputs of an unknown function using Gaussian and Yule-Simon processes," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 7205–7209.

[14] Kira Kowalska and Leto Peel, "Maritime anomaly detection using Gaussian process active learning," in *2012 15th International Conference on Information Fusion*. IEEE, 2012, pp. 1164–1171.

[15] Artur Trindade, "ElectricityLoadDiagrams20112014," UCI Machine Learning Repository, 2015, DOI: https://doi.org/10.24432/C58C86.

[16] David Duvenaud, *Automatic Model Construction with Gaussian Processes*, Ph.D. thesis, 2014.

[17] Václav Chudáček, Jiří Spilka, Miroslav Burša, Petr Janků, Lukáš Hruban, Michal Huptych, and Lenka Lhotská, "Open access intrapartum CTG database," *BMC Pregnancy and Childbirth*, vol. 14, pp. 1–12, 2014.

[18] Jirı Spilka, V Chudáček, Miroslav Burša, Lukáš Zach, Michal Huptych, Lenka Lhotská, Petr Janků, and Lukáš Hruban, "Stability of variability features computed from fetal heart rate with artifcially infused missing data," in *2012 Computing in Cardiology*. IEEE, 2012, pp. 917–920.

[19] Guanchao Feng, J Gerald Quirk, and Petar M Djurić, "Recovery of missing samples in fetal heart rate recordings with Gaussian processes," in *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 261–265.