ACS OMEGA

Open Access

This article is licensed under CC-BY-NC-ND 4.0

http://pubs.acs.org/journal/acsodf

Article

Nonlinear Small Sample Data Regression with a New Rational-Quadratic Minkowski Kernel for Tobacco Laser Perforation Process Tar Reduction Estimation

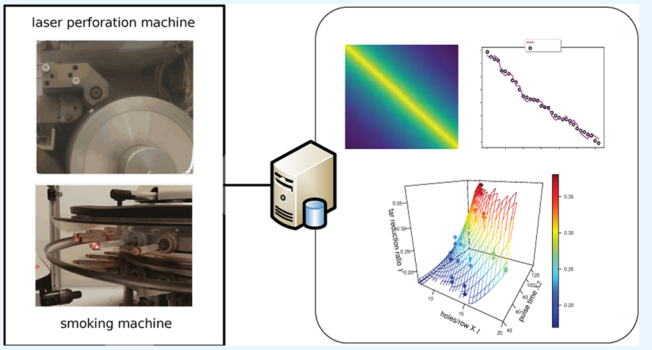
Juan Huo,* Feng He, Changtong Lu, Meng Zhu, Yifan Bu, Di Kang, Rui Wang, Wenning Feng,* and Rong Ma*

Cite This: *ACS Omega* 2025, 10, 2908−2918

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: This paper investigates the nonlinear relationship between tobacco harmful content tar reduction and laser perforation parameters. To find a model to demonstrate the relationship between the laser perforation parameters and the cigarette tar reduction level, an online platform based on Python Streamlit was built to collect and publish related data. After the initial analysis of the collected experimental data, the quadratic nonlinear regression model demonstrates a significant fit to the experimental data. However, although the nonlinear regression has much higher accuracy than the linear regression plane, the prediction normalized root mean squared error (NRMSE) is still high, over 10%, which indicates that the regression relationship is more complex than the simple quadratic function expression. On the other hand, the sample dataset used for modeling is very limited, which restricts its exploration and the development of a model comparable to those built with big data. To address this challenge for small sample size data in modeling this complex nonlinear relationship, a novel rational-quadratic Minkowski (RM)-based kernel was designed. This RM-kernel model acquires higher accuracy than other kernels in both SVM and Gaussian process regression. Furthermore, this new kernel also shows less sensitivity to hyperparameter change, the greater ability to capture complex relationships, and more flexibility than the RBF kernel and RQ kernel. Subsequently, the kernel-based RM regression model was successfully implemented for laser perforation parameter selection, yielding consistent results that align with human sensory test data.



1. INTRODUCTION

Nowadays, Industry 4.0 is revolutionizing the way of manufacturing in companies.[1] The smart factories are equipped with advanced sensors and embedded software to collect and analyze data to improve the production efficiency and decision making. The performance of machine learning in recognizing patterns is closely related to the size of the dataset.[2] There have been numerous rules of thumb suggested for determining the minimum number of samples required to have enough statistical power for regression analysis.[3] However, due to various kinds of limitations, the scarcity of the labeled data sample size is common in the industry, as the sample response always requires costly and time-consuming laboratory testing. This dilemma also exists in wide areas of the tobacco industry, in which the laser perforation technique used for filter ventilation is one of them. Filter ventilation is an important technique in the tobacco industry to reduce the unhealthy content of a single cigarette during smoking. This method involves adjusting the air flow passing through the filter of a cigarette in order to dilute the concentrations of compounds found in mainstream smoke. By doing so, the tobacco industry can create cigarettes that produce lower levels of tar, nicotine, and CO (TNCO) while still maintaining the addictive properties of nicotine. The WHO Study Group on Tobacco Product Regulation (TobReg) has advised the tobacco industry to regulate and lower toxicant levels in cigarette smoke.[4] The microholes perforated by the laser beam dilute the mainstream smoke of the cigarettes during combustion, so as to reduce the unhealthy content and improve the sensory quality.[4−6] Although there is some argument about the effect of the filter ventilation technique due to the smoker's personal compensation smoking behavior,

Received: October 7, 2024
Revised: December 2, 2024
Accepted: December 10, 2024
Published: January 13, 2025

© 2025 The Authors. Published by American Chemical Society

https://doi.org/10.1021/acsomega.4c08978
ACS Omega 2025, 10, 2908−2918

it is widely proved and accepted that smoke dilution can reduce the harmful carbon monoxide and tar in mainstream smoke.[4,7−13] At present, researches have studied many different aspects of related techniques, such as drilling methods, tar reduction process, regular physical and chemical index of smoke and aroma components, etc.[14−22] It has been reported that some tobacco components are known to vary linearly with the amount of ventilation, and some other components vary in relationships that are not linear.[4,8,10,13,23−25] To assist the industry designer for healthier tobacco products, an experiment here was conducted to further investigate the effect of laser perforation parameters on both harmful content reduction and mouth feel.[24]

For the laser perforation process, there are two kinds of laser techniques most widely used for cigarette filter ventilation: (1) online laser perforation directly inside the cigarette machine and (2) offline or preperforation performed by tipping paper manufacturers.[15−18,26] According to ref27 the online perforation technique has an advantage in controlling the stability of the ventilation rate for cigarette filter ventilation. Therefore, this research uses the online laser perforation technique as shown in Figure 1, where high-energy laser beams are used to

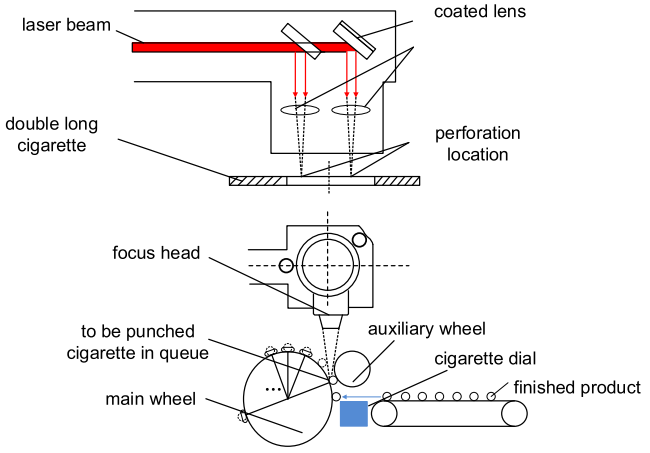


Figure 1. Mechanical structure of cigarette laser perforation.

penetrate and vaporize tipping paper and plug wrap paper during the rolling process of slim cigarettes to create microholes instantly. In Figure 1, the upper part is the inner structure of the laser perforation architecture of the laser head. The laser beam is focused by the coated lens and transferred to be a pulse beam. These coated lenses are controlled by a remote computer to adjust the lens' rotor. The lower part of Figure 1 is the whole cigarette perforation machine's structure. Each cigarette is sent to the fixed position under the laser head through the main wheel before rolling past an auxiliary wheel to create rows of perforation holes. Since these wheels are also remotely controlled, the number of laser holes per row can be manipulated. Because of these flexible components, laser perforation can be easily adapted to the required parameter even remotely. However, the perforation parameter is limited for physical and mechanical reasons, and the lab evaluation experiment is complex, which restricts the sample data size.

To collect sample data and select a proper model to estimate the effect of the various laser perforation parameters on cigarette quality, several models for small sample data machine learning were embedded in the backend of an online system whose web services can be remotely accessed by both the lab investigator and laser perforation controller. The architecture of this online test platform can be seen in Figure 2. After the

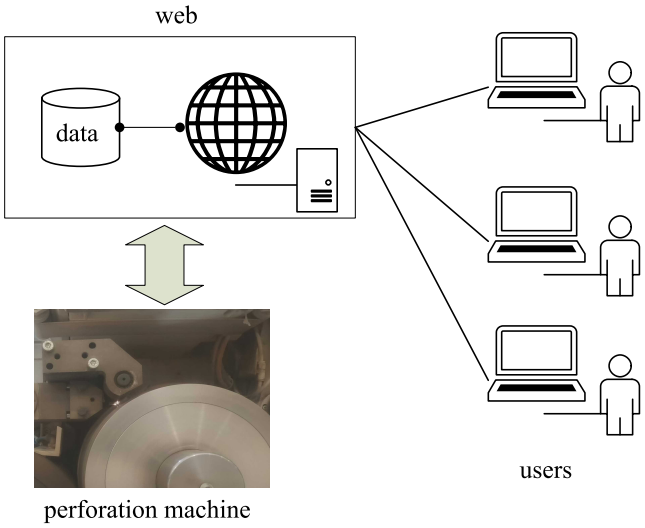


Figure 2. Architecture of the laser perforation estimation system.

perforation parameters were recorded in the online system database, the perforated cigarettes were sampled and their corresponding smoke chemicals and filter ventilation level were evaluated in the lab by a smoking machine and gas chromatograph, among other things, whose results will also be uploaded online.

In our analysis of laser perforation data of brand "H", it is found that the holes/row and pulse time have a close nonlinear relationship with the harmful content reduction level. Initial analysis of the experimental data revealed that a quadratic nonlinear regression model provides a significant fit. Despite achieving higher accuracy than linear regression, the normalized root-mean-square error (NRMSE) remains above 10%, suggesting a more complex relationship beyond a simple quadratic function. The limited size of the dataset restricts further exploration and hinders the development of a model comparable to those built with larger datasets.[2] To overcome this challenge inherent in modeling nonlinear relationships with small sample sizes, we propose a novel rational-quadratic Minkowski (RM)-kernel for kernel methods since kernel methods are popularly used for small dataset modeling because they strike a good balance between capturing nonlinearity and avoiding overfitting. For kernel method-dependent nonlinear regression, the kernel function is used to transform the input data into a higher dimensional space in order to find the best separation between the classes. The choice of kernel thus can have a significant impact on the performance of the model, as it determines how the data are mapped and how nonlinear relationships are captured. Among the kernel methods, support vector regression (SVR) and Gaussian processes (GP) are widely recognized as popular choices. Section 5 demonstrates the implementation of the proposed RM kernel within SVR and Gaussian process regression, revealing superior performance in the laser perforation regression task compared with other kernels. This proposed kernel model is also applied to the laser parameter and mouth feel relationship in Section 5.1. The source python code of this GP-RM (Gaussian process with Rational Minkowski) kernel and related data can be seen in https://www.github.com/JxxxHuo/RMkernel.

https://doi.org/10.1021/acsomega.4c08978
ACS Omega 2025, 10, 2908−2918

2. DATA SOURCE EXPERIMENTS

2.1. Materials and Instruments. *2.1.1. Main Materials.* Fine slim cigarettes from the same brand were used; tipping paper: 72 mm; cigarette paper: 19 mm × 4000 m × 32g 60 CU; and filter rod: 120 mm × 16.9 mm.

2.1.2. Main Instruments. An AB204-S electronic balance, Mettler Toledo, Switzerland; ZJ coiler, Changde Tobacco Industrial Machinery Factory; SODIMAX comprehensive test bench, SODIM; FD240 blast drying oven, BINDER; RM200A rotary smoking machine, BORGWALDT; Agilent 6890N gas chromatograph, Agilent Company of the United States; and LASERZJ online cigarette laser drilling system were used (laser model: CO2 pulsed laser; wavelength: 10.6 ± 0.4 μm; divergence: ≤ 2.0 mrad; laser diameter: 9 ± 1 mm; drilling Pulse time: 2−1000 μs; number of holes: 4−99 holes).

2.2. Experiment Design. *2.2.1. Cigarette Rolling Specifications.* Cigarette circumference: 17.00 ± 0.20 mm; cigarette length: 30 + 67.0 ± 0.4 mm; cigarette weight: 0.55 ± 0.05 g/cigarette; online laser drilling equipment parameter adjustment test, number of punching rows: 2 rows; punching position: 13 mm from the lip end; number of punching holes in each row: 7−20; and pulse duration: 50−130 μs.

2.2.2. Sample Preparation. The test needs to collect data on the number of online laser double row perforations, laser pulse duration time, and tar content of slim cigarettes. The values of perforation quantity and pulse time cover a wide range, and the measured tar content covers the determination of nonperforated samples. Table 1 shows these values in detail. The two laser perforation parameters were directly imported from the perforation machine from the factory product line, which are the main parameters of the laser perforation machine.

2.2.3. Detection of Physical and Chemical Indicators. Test samples were tested for nicotine, tar, carbon monoxide, moisture, and total particulate matter in mainstream smoke according to national standards: GB/T 19609-2004, GB/T 23355-2009, GB/T 23356-2009, and GB/T 23203.1-2008.

3. STATISTICAL ANALYSIS

According to some research, both the ventilation parameter and smoke chemical emissions have close relationship with the laser perforation parameter.[28] To further investigate the effect of laser perforation, in this research, tar values were specifically measured from cigarettes with laser holes and compared to those without laser holes, as shown in Figure 3. The red cross represents the tar value measured from cigarettes without holes, and the black square represents the tar value measured from cigarettes with holes. These samples listed in the index were collected from the same brand "H" with two tar value standards according to different leaf sources. These sample values are shown as two parallel lines in Figure 3. Therefore, to evaluate the effect of laser perforation on tar reduction within one standard, a tar reduction ratio is used instead of tar as the output metric in this paper. For cigarette sample i, its tar reduction ratio y_i is defined as

$$y_i = \frac{\mathrm{Tar}_{i\_without\_holes} - \mathrm{Tar}_{i\_with\_holes}}{\mathrm{Tar}_{i\_without\_holes}} \quad (1)$$

In the following analysis, the laser perforation parameters, the holes per row, and the laser pulse time are the input variables represented as $\boldsymbol{X_1} = \{x_{11}, x_{12}..., x_{1n}\}$ and $\boldsymbol{X_2} = \{X_{21}, X_{22}, ..., X_{2n}\}$ separately. The output is the dependent variable tar reduction ratio as $\boldsymbol{Y} = \{y_1, y_2, ..., \text{and } y_n\}$.

Table 1. Perforation Parameter and Tar Content

holes/row	time (μs)	tar (mg)	tar without a hole (mg)	tar reduction ratio (%)
14	80	4.39	7.09	38.08181
12	100	6.03	9.56	36.92469
14	80	6.08	9.56	36.40167
12	100	4.53	7.09	36.10719
16	60	4.7	7.09	33.70945
14	70	6.35	9.56	33.57741
16	60	6.4	9.56	33.05439
14	70	4.86	7.09	31.45275
12	80	6.59	9.56	31.06695
12	80	5.01	7.09	29.33709
18	50	6.9	9.56	27.82427
7	130	6.93	9.56	27.51046
20	40	5.17	7.09	27.08039
18	50	5.18	7.09	26.93935
20	40	7	9.56	26.77824
7	130	5.28	7.09	25.52891
12	70	5.3	7.09	25.24683
12	70	7.16	9.56	25.1046
14	60	7.27	9.56	23.95397
7	100	5.44	7.09	23.27221
12	60	7.38	9.56	22.80335
12	60	5.5	7.09	22.42595
14	60	5.5	7.09	22.42595
7	110	7.45	9.56	22.07113
7	100	7.53	9.56	21.23431
7	110	5.65	7.09	20.3103
12	50	7.67	9.56	19.76987
7	90	5.71	7.09	19.46403
14	50	7.75	9.56	18.93305
12	50	5.77	7.09	18.61777
7	90	7.89	9.56	17.46862
14	50	5.89	7.09	16.92525

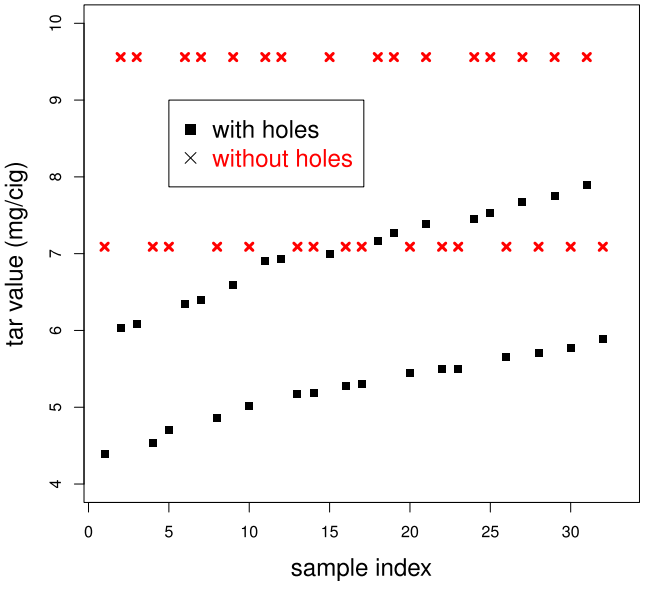


Figure 3. Tar values of samples.

https://doi.org/10.1021/acsomega.4c08978
ACS Omega 2025, 10, 2908−2918

4. DATA MODEL

4.1. Linear Regression. To explore the relationship between the tar reduction ratio and laser perforation parameters, initially, a linear regression plane based on $Y_{pred} = X_1 + X_2$ is generated as shown in Figure 4. In this figure, the

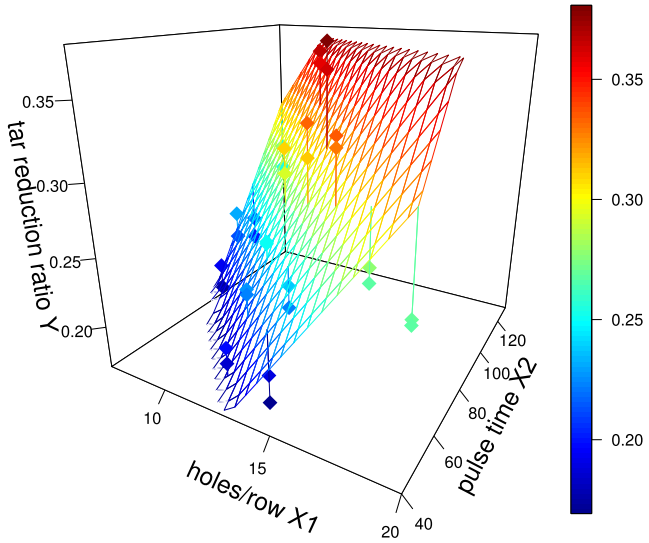


Figure 4. Linear regression between the tar reduction ratio Y and input perforation parameters: hole numbers/row X_1 and pulse time X_2.

grid lines are generated from sequential independent variables of $[\min(X_1), \max(X_1)]$, $[\min(X_2), \max(X_2)]$ and dependent variables of Y_{pred}. As seen in Figure 4, some samples are within the plane. The tar reduction ratio, Y, generally decreases with input variables X_1 and X_2. The residual standard error of this linear model is 0.03777 on 29 degrees of freedom. However, most of the data points are far from the plane, and the prediction error is big.

4.2. Nonlinear Regression. Then, a nonlinear least-squares regression (NLS) model with a second-order polynomial function $f(\mathbf{x}, \boldsymbol{\beta})$, a quadratic function as eq 2, is built, in which $\boldsymbol{\beta} = \{\beta_0, \beta_1, \beta_2, ..., \beta_5\}$ represents the polynomial parameters.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i}^2 + \beta_4 x_{2i}^2 + \beta_5 x_{1i} x_{2i} \quad (2)$$

Within the samples, the residual e_i of sample i is

$$e_i = r_i - f(\mathbf{x}_i, \boldsymbol{\beta}) \quad (3)$$

The nonlinear regression process is to find the vector $\boldsymbol{\beta}$ of parameters such that the curve fits best to the given data in the least-squares that can minimize $\sum_{i=1}^{n} e_i^2$. After the regression process, a grid plane generated from this model and the sample data as diamonds are shown in Figure 5, in which most of the sample data are close to the grid plane. The residual standard error of this nonlinear least-squares regression model is 0.01921 on 26 degrees of freedom, which shows better performance than the linear regression model. However, the accuracy of the nonlinear regression depends on an acceptable parameter estimate and a good model fit. The heavily adjusted parameters after training will also easily induce overfitting especially for models trained from a small sample dataset.[29] As shown in the test results of Table 2 in Section 5, the LOOCV test results of the normalized root mean squared error of

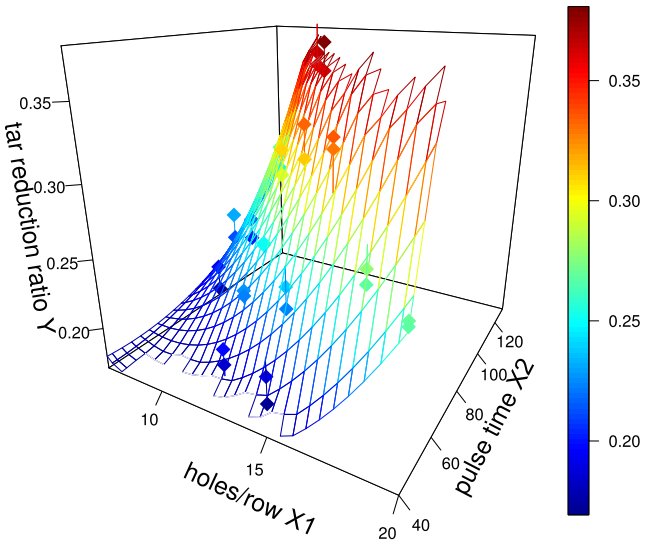


Figure 5. Nonlinear plane generated between the laser perforation parameters and the tar reduction ratio. The diamonds represent the lab data.

nonlinear regression is above 10%, while the R^2 value is less than 0.9.

Table 2. Result Comparison

kernel name	NRMSE (%)	RMSE	R^2	RPD	time cost (s)
linear	19.1	0.040	0.564	1.527	0.0610
nonlinear	10.1	0.021	0.877	2.881	0.0653
SVM-Linear	18.47	0.645	0.584	1.550	91.235
SVM-RBF	8.49	0.296	0.912	3.374	54.447
SVM-RQ	9.74	0.340	0.884	2.940	47.605
SVM-RM	7.00	0.245	0.940	4.089	47.697
GP-Linear	19.06	0.665	0.557	1.503	0.333
GP-RBF	7.02	0.245	0.940	4.078	0.541
GP-RM	7.02	0.245	0.940	4.078	0.968
PLS-regr	19.06	0.6653	0.5573	1.5030	0.0404

4.3. Support Vector Regression. As one of the most used machine learning algorithms for small sample datasets, the kernel-based support vector machine (SVM) is widely considered good at dealing with a small size dataset, for which the deep learning neural networks always fail.[2] The essential idea of SVR is to construct a hyperplane or a set of hyperplanes that can separate the data in a high-dimensional space for use in classification or regression.[30] The SVR method used for solving regression problems is called SVR (support vector regression). The standard SVR is normally to solve

$$\min_{\boldsymbol{w},b,\xi,\xi^*} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{n}\xi_i + C\sum_{i=1}^{n}\xi_i^*$$

$$\text{subject to } \boldsymbol{w}^T\phi(\boldsymbol{x}_i) + b - y_i \le \epsilon + \xi_i,$$

$$y_i - \boldsymbol{w}^T\phi(\boldsymbol{x}_i) - b \le \epsilon + \xi_i^*,$$

$$\xi_i, \xi_i^* \ge 0, i = 1, ..., n, \ \epsilon \ge 0 \quad (4)$$

In this equation, C, ϵ, and γ are important soft margin parameters. As SVR is a supervised machine learning algorithm that works by finding the optimal hyperplane that maximizes

https://doi.org/10.1021/acsomega.4c08978
ACS Omega 2025, 10, 2908−2918

the margin between two classes of data points. The soft margin parameter, also known as the slack variable or regularization parameter, is used to control the complexity of the decision boundary. In eq 4, the function $\boldsymbol{\phi}(x)$ is used to map data to high-dimensional space, which transforms the input vectors $\boldsymbol{x} \in \mathbb{R}^n$ into the feature space $\boldsymbol{\phi}(x) = [\phi_1((\boldsymbol{x})), \phi_2((\boldsymbol{x})), ..., \phi_n((\boldsymbol{x}))]^T \in \mathbb{R}^f$.[31,32] To avoid coordinate computation, the kernel function of $\boldsymbol{\phi}(x)$ as

$$\boldsymbol{K}(x_i, x_j) = \phi(x_i)^T \phi(x_j) \tag{5}$$

is used to operate in a high-dimensional space by simply computing the inner products between all pairs of data. This equivalent dimension mapping process provides more generality for the learning model. All the above algorithms are implemented by the Python machine learning library scikit-learn.[33]

4.4. Gaussian Process Regression. Gaussian process regression (GPR) is a nonparametric Bayesian approach for regression analysis, where the predicted function is represented as a Gaussian process. Instead of estimating the specific parameters of a function, GPR models the entire function space, providing not only predictions but also uncertainty estimates.[34,35] A Gaussian process is uniquely defined by its mean function and covariance function. The kernel function $k(x, x')$ models the covariance between each pair in x. Generally, the Gaussian process can be expressed as

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{6}$$

$$\text{where, } m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] \tag{7}$$

$$k(\mathbf{x}, \mathbf{x}') = \text{Cov}(f(\mathbf{x}), f(\mathbf{x}')) \tag{8}$$

In this expression, $\mathcal{GP}$ denotes that $f(\mathbf{x})$ is a Gaussian process. $m(\mathbf{x})$ is the mean function, representing the expected value of $f(\mathbf{x})$, and $k(\mathbf{x}, \mathbf{x}')$ is the covariance function or called kernel, which determines the covariance between $f(\mathbf{x})$ and $f(\mathbf{x}')$. Gaussian process regression (GPR) is also nonparametric; it is a Bayesian approach to regression analysis, where the predicted function is represented as the Gaussian process. The predictive distribution of GPR can be represented as

$$p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(\mu_*, \sigma_*^2) \tag{9}$$

$$\mu_* = \mathbf{k}_*^T(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{y} \tag{10}$$

$$\sigma_*^2 = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^T(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{k}_* \tag{11}$$

where $\mathbf{K}$ is the covariance matrix between the training inputs, $\mathbf{k}_*$ is the vector of covariances between the training inputs and the new input, $k(\mathbf{x}_*, \mathbf{x}_*)$ is the covariance between the new input and itself, and σ_n^2 is the variance of the noise. The covariance function, namely the kernel function $k(\cdot,\cdot)$ is a crucial component in Gaussian process (GP) regression because it determines the way the data are related to the outputs or the predictor variables are related to the inputs. A kernel function maps the input data into a higher dimensional space and can determine the complexity of the model, the smoothness of the functions, and how well the model fits the data. In GP regression, the choice of kernel function affects the trade-offs in terms of flexibility and interpretability of the model. Commonly used kernels include the squared exponential (SE) kernel, Matérn family of kernels, and the radial basis function (RBF) kernel.[36]

4.5. Rational-Minkowski Kernel. Both GPR and SVR are kernel-based nonparametric machine learning algorithms; namely, they all do not impose any constraints on the shape of the underlying functions. However, it is necessary to customize the kernel for specific data characteristics to improve the performance of these two types of models. Choosing the most appropriate kernel highly depends on the problem at hand because it depends on what we are trying to model. Different kernels work better with different types of data, and selecting the appropriate one can improve the accuracy of the resulting classifier or regressor. As mentioned in the context, one way to customize a kernel for specific data characteristics is to use domain-specific knowledge or problem-specific constraints to guide its design. The high correlation and nonlinearity between the laser perforation parameters and the tar reduction ratio require a proposed kernel to better fit their small sample regression. This is the reason a new rational kernel based on the Minkowski distance is designed for laser perforation data.

The Minkowski distance is a metric in a normed vector space that can be considered a generalization of both the Euclidean distance and the Manhattan distance. Thus, we design a new rational-Minkowski (RM) kernel, which replaces the squared Euclidean distance in the rational-quadratic (RQ) kernel with Minkowski distance. The Minkowski distance of order p between two points $x_a = (x_{a1}, x_{a2}, ..., x_{an})$ and $x_b = (x_{b1}, x_{b2}, ..., x_{bn}) \in \mathbb{R}^n$ is defined as

$$D(x_a, x_b) = \left(\sum_{i=1}^{n} |x_a - x_b|^p\right)^{1/p} \tag{12}$$

Therefore, the rational-Minkowski (RM) kernel is

$$k(x_a, x_b) = \sigma^2\left(1 + \frac{\left(\sum_{i=1}^{n} |x_a - x_b|^p\right)^{1/p}}{2\alpha l^2}\right)^{-\alpha} \tag{13}$$

Note that it is only a quasi-metric if $0 < p < 1$.

σ^2 is the overall variance (σ is also known as amplitude), l is the length scale, and α is the scale mixture ($\alpha > 0$).

The expression of the RM kernel is similar to the rational-quadratic (RQ) kernel except that the RQ kernel uses Euclidean distance. The advantage of the RQ kernel is widely known as it captures both linear and quadratic relationships between the data points, making it suitable for a wide range of problems in machine learning such as regression or classification. Similarly, the RM kernel also has the virtue that it can catch both linear and special nonlinearity of data, and the Minkowski distance is a generalization of both the Euclidean distance and the Manhattan distance. Thus, the reason for us to use the Minkowski distance instead of the Euclidean distance is that the Minkowski distance is a more general metric for normed vector space when it is used to determine the similarity between the sequences. It is believed that the Minkowski distance has the ability to capture long-range dependencies better than Euclidean distance since Euclidean distance is limited to L2, the Euclidean norm.

To demonstrate the characteristics of this proposed kernel and compare it with some existing kernels, a synthetic dataset was constructed from a linear equation of two standardized 322 variables with $x \in [-10, 10]$ as the input variable of kernel

https://doi.org/10.1021/acsomega.4c08978
ACS Omega 2025, 10, 2908−2918

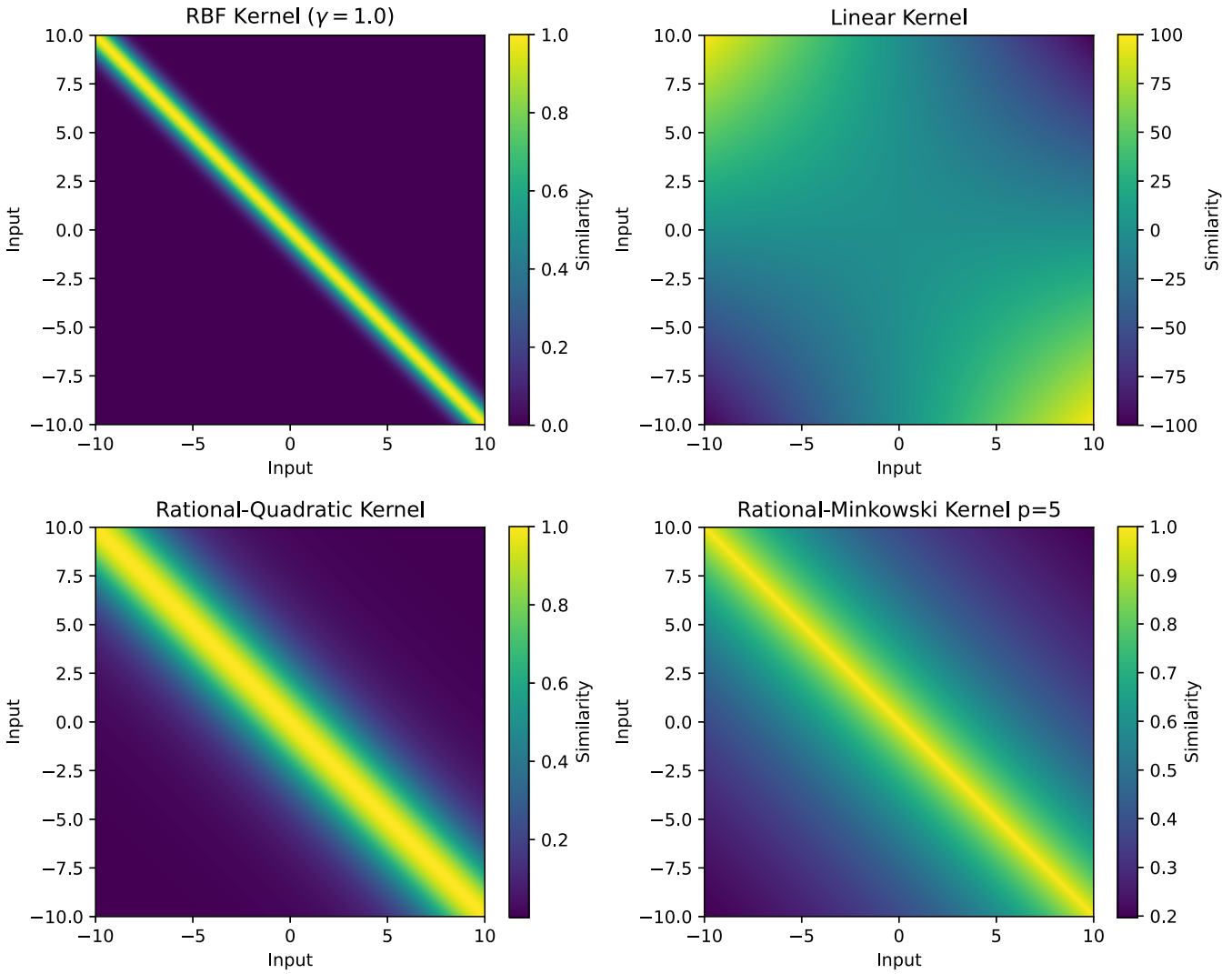


Figure 6. Kernel image of the visual representation.

functions. Figure 6 shows the visual representation of the kernel covariance matrix. Every kernel in these images has two input variables, $x_a \in \mathbf{x}$ and $x_b \in \mathbf{x}$. The maximum values of the covariance matrix of the kernels RBF, RQ, and RM are the same as we set the same length scale $l = 2$ and the overall variance $\sigma^2 = 1$. However, the overall spread of RBF is much smaller than RQ and RM. With norm order $p = 5$, the spread of RM is much wider than RBF and RQ. Figure 7 shows the distance plot in line with respect to $k(0, x)$, in which the similarity output of each kernel covariance decreases toward 0 as the input is farther away from the center. The center has the maximum similarity at $x_a = x_b$. From the center, the similarity decreases exponentially by the RBF kernel and RQ in an inverted U-shape. If the length scale value is higher, then, the zero roots (where the parabola intersects the x-axis) is further away from the center. With the same length scale value $l = 2$, the parabola of RM has the widest spread. In addition, the spread of the RM parabola does not change with the p value. Minkowski kernel curves nearly overlap each other, although p varies from 1 to 10 and the slope is rather flat compared to RBF and RQ. The shape of RM is in a sharp triangle in Figure 7, which is quite different from the U top of RBF and RQ. Generally speaking, the RM kernel covariance matrix has many different characteristics compared to RBF and RQ.

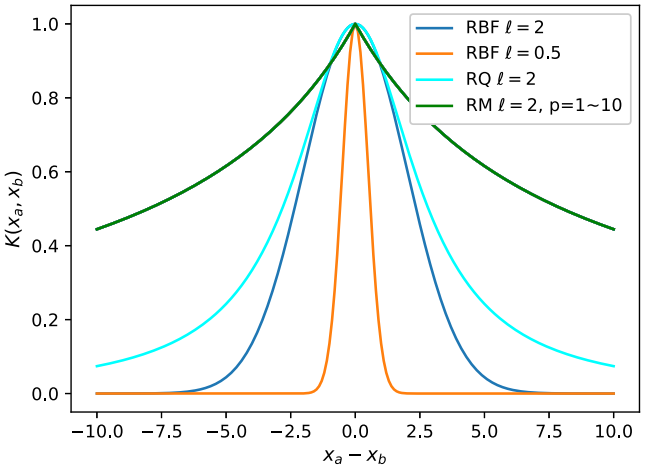


Figure 7. Kernel visualization.

To further compare the smoothness which is estimated by maximum absolute derivative between different kernels, we use a sine function with added random noise to generate a one-dimensional vector with 15 points for training and 100 points for the test as shown in Figure 8. Due to the nonlinearity of the

https://doi.org/10.1021/acsomega.4c08978
ACS Omega 2025, 10, 2908−2918

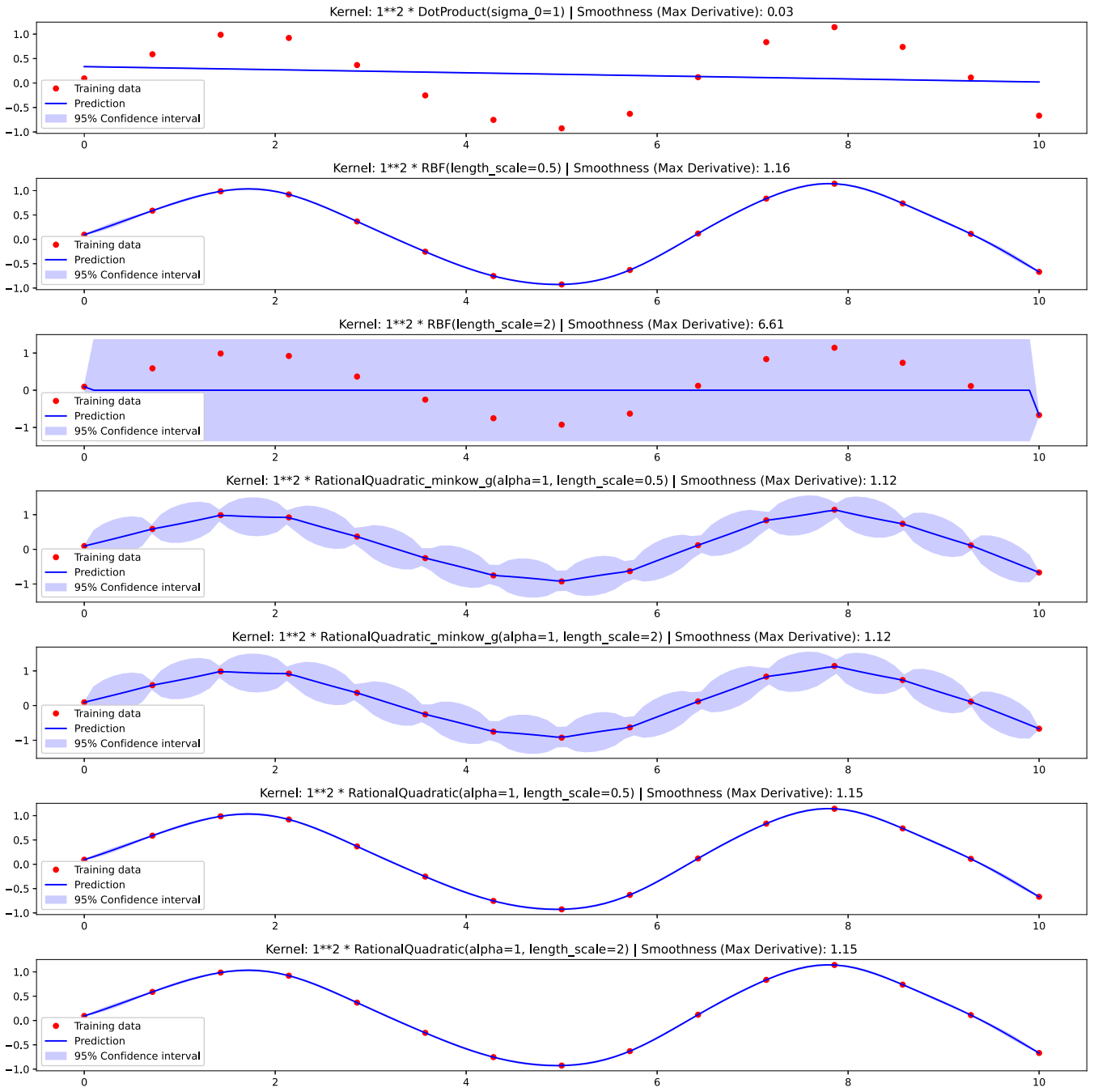


Figure 8. Kernel visualization.

sine function, Figure 8 shows that the RBF, RM, and RQ kernels in the Gaussian process can accurately predict the sine wave. Gaussian process regression with the RBF kernel shows more sensitivity to hyperparameters like length scale l. As the increased l will produce smoother predictions, when the length scale is increased from $l = 0.5$ to $l = 2$, the RBF kernel is underfit and the smoothness level is as high as 6.61. By contrast, the RM and RQ kernels are unaware of the length-scale parameter change and keep the same prediction accuracy and smoothness level. The RM kernel has a wider confidence interval area than RQ($l = \{0.5, 2\}$) and RBF($l = 0.5$). The confidence interval area of RM is in a proper size, and the size is stable for various length scale values, indicating that RM can better accommodate noise in the data and capture more complex relationship than RBF and RQ.

5. RESULTS OF LASER PERFORATION DATA ANALYSIS

All the above kernels were then implemented to our laser perforation data, and the regression performances of different learning models are compared. In the Sklearn python library used in this project, the hyperparameters of the Gaussian process are automatically optimized by the conjugate gradient algorithm by default. As shown in Figure 9, with all the laser perforation data as the training and test data, the RM kernel in both SVR and Gaussian process regression has better fitness than the other kernels.

https://doi.org/10.1021/acsomega.4c08978
ACS Omega 2025, 10, 2908−2918

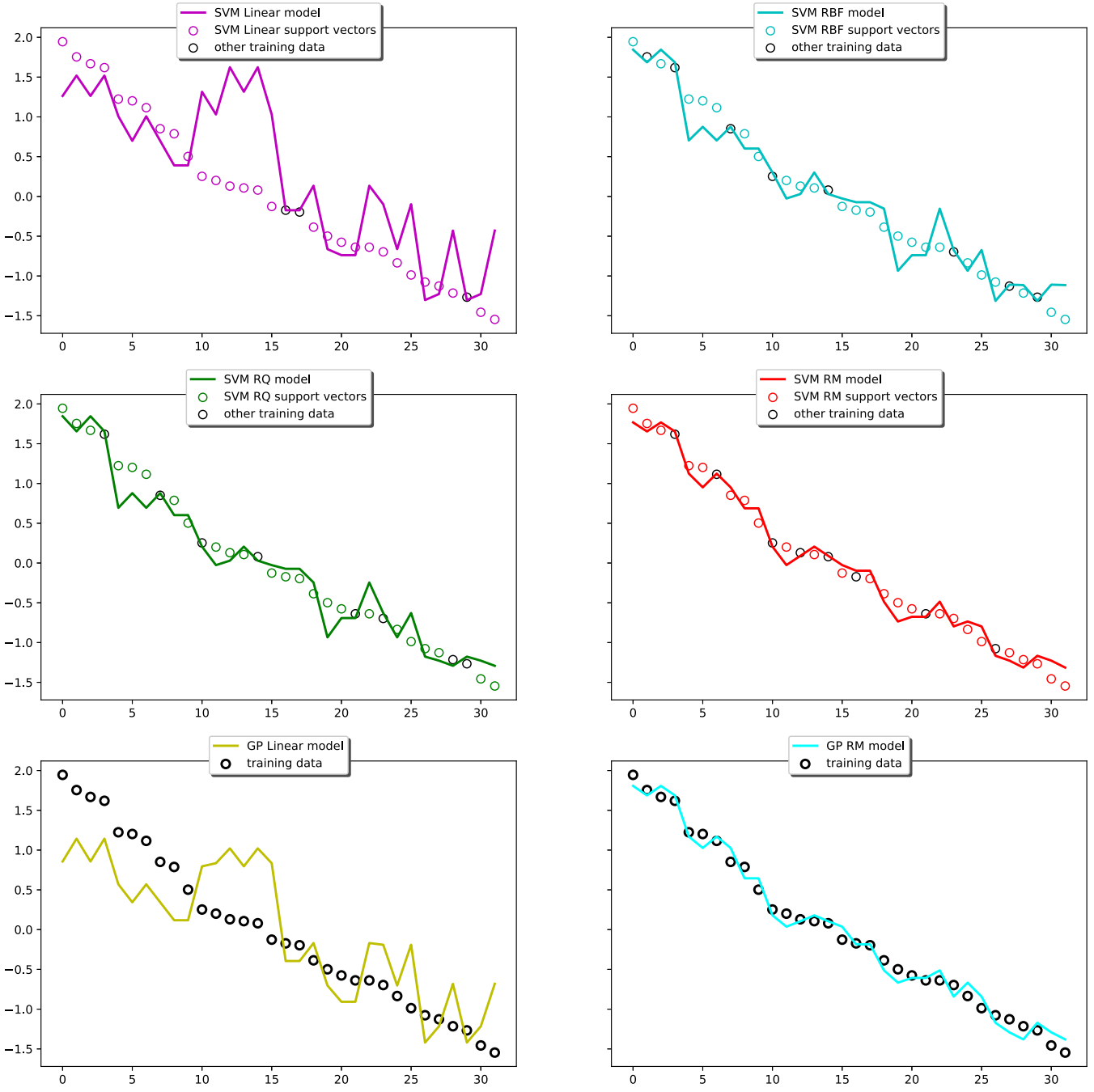


Figure 9. Tar reduction prediction results from different kernel-based regression algorithms.

To further estimate the performance numerically, Table 2 uses several measures to compare these models' performance by leave-one-out cross-validation. Several parameters derived from these validation results are used to estimate the learning model's performance as listed in Table 2. The first measure is NRMSE (normalized root mean squared error), and the second measure is RMSE (root mean squared error). Essentially, they all measure the prediction error, which is the difference between the prediction and observation in the absolute value. NRMSE normalizes the RMSE with the mean of observation values, which is helpful for comparing different datasets within their own scales. Here, the time cost is the total time cost of the computation duration of leave-one-out cross-validation. It should be noted that the soft margin parameters C and ϵ parameters of SVR have been optimized by a search grid.

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{I}}{(\hat{y}_i - y_i)^2}} \tag{14}$$

$$\mathrm{NRMSD} = \frac{\mathrm{RMSD}}{\overline{y}} \tag{15}$$

where $\overline{y}$ is the mean of the real tar ratio.

The third analytical measure used for result comparison is the coefficient of determination, namely R^2 or R-squared, which is the proportion of variance in the dependent variable $\hat{y}_t$ predicted from the independent variables y_t.

https://doi.org/10.1021/acsomega.4c08978
ACS Omega 2025, 10, 2908−2918

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{16}$$

in which

$$SS_{res} = \sum_i (y_i - \hat{y}_i)^2 = \sum_i e_i^2 \tag{17}$$

$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \tag{18}$$

R^2 measures how well real values are replicated by a model based on the proportion of the total variation of model outcomes. There are several different kinds of definitions for coefficient determination, and the formula varies accordingly. The higher the R^2 value, the less likely it is for the model to overfit. The fourth measure is the ratio of performance to deviation (RPD), which is the ratio between the standard deviation of a variable and the standard error of prediction. It is often used to report the quality of a model in the field of spectroscopy in particular. The higher the RPD, the more stable the prediction accuracy will be when the input varies.

In Table 2, the results show that the performance of nonlinear kernel methods surpasses the no-kernel methods in all parameters. The SVR and GP models with a linear kernel have worse performance than direct nonlinear regression. The SVR performance measurement results in Table 2 are concluded from massive SVR tests with SVR hyperparameters C, γ, and ϵ. The SVR parameters are grid-searched in the range $C \in [1, 200]$, $\gamma \in [10^{-7}, 10]$, and $\epsilon \in [0.01, 1]$. Table 2 shows that the SVR-RM regression model has the best accuracy in SVR kernel methods. However, when C is lower than the threshold, the SVR-RM NRMSE value can be as high as over 24.5% when $C = 1$ and $\epsilon = 1$ as shown in Figure 10 grid search of SVR-RM. The NRMSE error increases with ϵ and the C has a negligible effect when C is larger than 17. The best performance appears when $\epsilon < 0.01$ and $C > 20$. The most critical disadvantage of SVR is time cost, as SVR needs the grid search method to find the best C, γ, and ϵ parameters, which is an exhaustive search method that tries all possible combinations of the hyperparameters in a predefined grid.

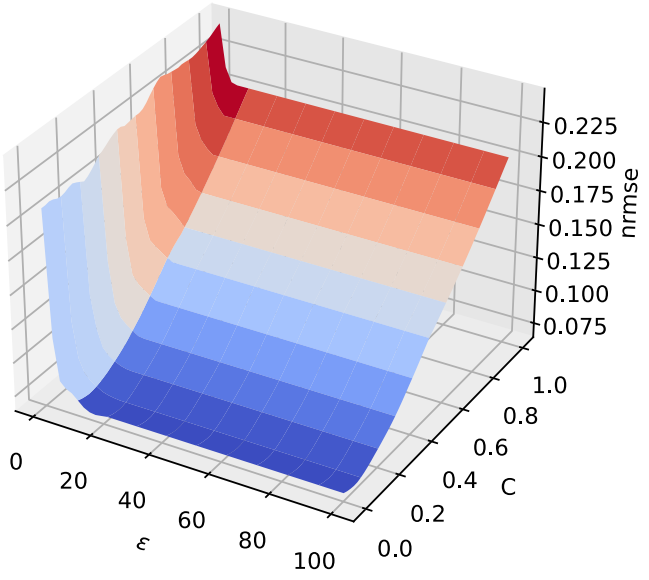


Figure 10. NRMSE color-warm map of SVR-RM regression cross-validation results with different C and ϵ values.

Compared to SVR-RM, the GP-RM model also has high accuracy but is more stable and costs much less time. This is because although SVR and GPR are nonparametric models, they all require hyperparameter tuning. While SVR most often uses a grid search method and GPR uses the L-BFGS-B, a quasi-Newton optimization algorithm, L-BFGS-B is generally faster than exhaustive search methods like grid search, especially when the objective function is smooth and differentiable. SVR's objective function (the combination of fitting the data and penalizing complexity) is often nonconvex, meaning that it has many local minima. This makes gradient-based optimization methods (such as L-BFGS) unreliable for finding the global optimum. Many SVR hyperparameters (such as the kernel type and regularization parameter C) are discrete or have a limited set of possible values. Grid search systematically explores these options, which are effective but can be computationally expensive. GPRs often have a smooth and well-defined objective function, which is the logarithmic marginal likelihood. This function is typically unimodal or has a few well-defined local minima. This makes GPRs more suitable to use faster L-BFGS to tune hyperparameters. Another advantage of the Gaussian process compared to SVR also lies in its ability to handle nonlinearly separable data and provides probabilistic predictions. Gaussian processes are able to model complex, nonlinear relationships between inputs and outputs by defining a covariance function that captures the underlying structure of the data. In addition, as the sample size is small, there is a correlation between the two input features. GPR uses kernel functions to define the covariance among the data points. These kernels can capture complex dependencies and correlations between features. GPR also provides uncertainty estimates for its predictions, which can be helpful in understanding the impact of correlated features. The choice of kernel function and its hyperparameters can often be learned from the data using maximum likelihood estimation or other Bayesian techniques. Gaussian processes regression thus are able to automatically learn the appropriate complexity of the model based on the data, resulting in more efficient and effective regression.

5.1. Model Application. For the tobacco industry, there must be a balance between the filter ventilation level and sensory satisfaction. At present, the tobacco sensory score was investigated by human investigators. Table 3 gives an example of the laser perforation data design results. According to the tobacco company's product standard, the tar value of each cigarette is expected to be around 7 mg/cig. The hole/row parameter is then set sequentially from 7 to 20. With the GP-RM model, the corresponding pulse time was then selected and is listed in Table 3, which can lead to the expected tar reduction ratio (the without-hole cigarette tar is 9.56 mg/cig). After the cigarettes with these parameters were laser-perforated, human investigators smoked and gave a sensory score to these cigarettes following the GB5606 4−2005 standard procedure. As shown in Table 3, the sensory scores are all close to 90, which is consistent with the same tar reduction ratio level. The most proper pairs of laser perforation holes/row and pulse time were then selected from this table, which are also kept in the online system in Figure 2.

https://doi.org/10.1021/acsomega.4c08978
ACS Omega 2025, 10, 2908−2918

Table 3. Perforation Parameter and Sensory Score when the target tar value is 6 mg

holes/row	time (s)	int time (s)	human score	predicted tar reduction ratio (%)	predicted tar (mg)
8	82.37	82	88	25.68	7.11
9	73.21	73	88.5	25.94	7.08
10	67.01	67	88.5	25.97	7.08
11	62.22	62	90.5	25.84	7.09
12	58.18	58	91	25.49	7.12
13	54.59	55	90.5	25.38	7.13
14	51.26	51	90	25.03	7.17
15	48.12	48	90.5	25.89	7.08
16	45.11	45	90	26.18	7.06
17	42.2	42	89.5	26.34	7.04
18	39.35	39	89	26.43	7.03
19	36.56	37	88	26.47	7.03
20	33.81	34	87.5	26.49	7.03

6. CONCLUSIONS

Laser perforation is a widely used technique in the tobacco industry to reduce the harmful content per cigarette in recent years. With the development of Industry 4.0, in this paper, an online platform was built to collect laser perforation data from lab and factory manufacturing. After data analysis, it is found that there is a special complex nonlinearity relationship between the laser perforation parameters (holes/row and pulse time) and tar reduction ratio. This relationship cannot be well explained by a nonlinear function directly. As the tobacco tar measurement process is complex, there can be only limited laser perforation parameters tested with the laboratory tar measurement. Therefore, it is required that an learning algorithm can deal with a small sample dataset and complex special nonlinear computation well to model this study. Hereby, a proposed kernel based on the rational-quadratic Minkowski distance, which is called the RM kernel, was built and embedded in kernel-based learning algorithms such as Gaussian process regression and SVR to model this special nonlinearity. The latter test results show that this new RM kernel can fit the laser perforation regression better than the other tested kernels in both SVR and GP learning with good performance. With more data tested in further investigation, the RM kernel shows less sensitivity to hyperparameter change, the greater ability to capture complex relationship, and more flexibility than the RBF model. The GP-RM model was implemented into the laser perforation parameter selection process by tar reduction ratio prediction for different paired holes per row and pulse time. With the selected laser perforation parameters for the expected tar reduction value, the selected cigarette samples were then tasted by human investigators and received a consistent sensory score, which is a reference for further industrial application.

7. DISCUSSION

Although initially developed for modeling nonlinear regressions of tar reduction levels, further investigation into the rational-Quadratic Minkowski (RM) kernel reveals its potential for broader applicability. This stems from the RM kernel's demonstrated efficacy in this study, suggesting its utility across diverse domains.

AUTHOR INFORMATION

Corresponding Authors

Juan Huo − *School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou 450001, China;* orcid.org/0000-0002-0399-7307; Email: juanhuo@126.com

Wenning Feng − *China Tobacco Hebei Industrial Co., Ltd, Shijiazhuang 050051, China*; Email: fengwn@126.com

Rong Ma − *China Tobacco Hebei Industrial Co., Ltd, Shijiazhuang 050051, China*; Email: ma.rong1979@163.com

Authors

Feng He − *China Tobacco Hebei Industrial Co., Ltd, Shijiazhuang 050051, China*

Changtong Lu − *China Tobacco Henan Industrial Co., Ltd, Zhengzhou 450001, China*

Meng Zhu − *China Tobacco Hebei Industrial Co., Ltd, Shijiazhuang 050051, China*

Yifan Bu − *China Tobacco Hebei Industrial Co., Ltd, Shijiazhuang 050051, China*

Di Kang − *China Tobacco Hebei Industrial Co., Ltd, Shijiazhuang 050051, China*

Rui Wang − *China Tobacco Hebei Industrial Co., Ltd, Shijiazhuang 050051, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c08978

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank the staff in Hebei Industrial co., Ltd. that have contributed to this experiment.

REFERENCES

(1) Lasi, H.; Fettke, P.; Kemper, H.-G.; Feld, T.; Hoffmann, M. Industry 4.0. *Bus. Inf. Syst. Eng.* **2014**, *6*, 239−242.

(2) Kokol, P.; Kokol, M.; Zagoranski, S. Machine learning on small size samples: A synthetic knowledge synthesis. *Sci. Prog.* **2022**, *105*, No. 368504211029777.

(3) Green, S. B. How many subjects does it take to do a regression analysis. *Multivar. Behav. Res.* **1991**, *26*, 499−510.

(4) Pauwels, C. G.; Klerx, W. N.; Pennings, J. L.; Boots, A. W.; van Schooten, F. J.; Opperhuizen, A.; Talhout, R. Cigarette filter ventilation and smoking protocol influence aldehyde smoke yields. *Chem. Res. Toxicol.* **2018**, *31*, 462−471.

(5) Coggins, C. R. E.; Merski, J. A.; Oldham, M. J. A comprehensive evaluation of the toxicology resulting from laser-generated ventilation holes in cigarette filters. *Inhalation Toxicol.* **2013**, *25*, 59−63.

(6) Salonitis, K.; Stournaras, A.; Tsoukantas, G.; Stavropoulos, P.; Chryssolouris, G. A theoretical and experimental investigation on limitations of pulsed laser drilling. *J. Mater. Process. Technol.* **2007**, *183*, 96−103.

(7) Parker, J. A.; Montgomery, R. T. Design criteria for ventilated filters. *Contrib. Tob. Nicotine Res.* **1979**, *10*, 1−6.

(8) Browne, C. L.; Keith, C. H.; Allen, R. E. The effect of filter ventilation on the yield and composition of mainstream and sidestream smokes. *Beitr. Tabakforsch. Int.* **1980**, *10*, 81−90.

(9) Jing, Y. Q.; Li, G. L.; Liu, J. J.; Huang, H. T.; Hu, H. M.; Zhang, H. L.; Xie, X. H.; Zhu, J. F. The Effects of Filter Ventilation on Flavor Constituents in Cigarette Smoke. *Appl. Mech. Mater.* **2011**, *79*, 35−42.

(10) Norman, V.; Ihrig, A. M.; Shoffner, R. A.; Ireland, M. S. The Effect of Tip Dilution on the Filtration Efficiency of Upstream and

https://doi.org/10.1021/acsomega.4c08978
ACS Omega 2025, 10, 2908−2918

Downstream Segments of Cigarette Filters. *Contrib. Tob. Nicotine Res.* **1984**, *12*, 178−185.
(11) King, B.; Borland, R. The "low-tar" strategy and the changing construction of Australian cigarettes. *Nicotine Tob. Res.* **2004**, *6*, 85−94.
(12) Schneller, L. M.; Zwierzchowski, B. A.; Caruso, R. V.; Li, Q.; Yuan, J.; Fong, G. T.; O'Connor, R. J. Changes in tar yields and cigarette design in samples of Chinese cigarettes, 2009 and 2012. *Tob. Control* **2015**, *24*, iv60−iv63.
(13) Caraway, J. W.; Ashley, M.; Bowman, S. A.; Chen, P.; Errington, G.; Prasad, K.; Nelson, P. R.; Shepperd, C. J.; Fearon, I. M. Influence of cigarette filter ventilation on smokers' mouth level exposure to tar and nicotine. *Regul. Toxicol. Pharmacol.* **2017**, *91*, 235−239.
(14) Cao, F.; Xie, X.; Wang, X. Effects of online laser perforation parameters on cigarette ventilation rate and deliveries of routine smoke constituents. *Tob. Sci. Technol.* **2014**, *11*, 45−49.
(15) Duan, Y.; Zhang, T.; Chen, J. Impact of perforated tipping paper by laser and static electricity on cigarette smoke. *Hubei Agric. Sci.* **2012**, *51*, 5403−5405.
(16) Feng, W.; Liao, Z.; Xu, S. The contrast affecting research of perforated tipping paper by laser or static electricity on cigarette smoke. *J. Yunnan Univ., Nat. Sci.* **2010**, *32*, 115−117.
(17) Han, Y.; Chen, P.; Zhou, Z. Y.; Chen, Z.; Zhang, X.; Deng, G. Laser perforating technology of tipping paper. *Laser Technol.* **2002**, *26*, 330−333.
(18) Jianfu, L.; Yong, J.; Ke, L. Influences of tipping paper on tar and 7 harmful components in cigarette smoke. *Tob. Sci. Technol.* **2013**, *8*, 67−70.
(19) Li, D.; Bao-ping, Q.; Ya, L.; Xiao-xu, L.; Qiang, L.; Lian-min, G.; Peng, G.; Jun-song, Z. Review of the effects of filter ventilation on delivery of aromatic component in mainstream smoke. *J. Zhengzhou Univ. Light Ind., Nat. Sci. Ed.* **2014**, *29*, 33 DOI: 10.3969/j.issn.2095-476X.2014.04.008.
(20) Song, Y.; Xiaodong, Z.; Haiying, T.; Xuehui, S.; Liang, G.; Xiaoyu, W.; Hongwei, W.; Mingzhe, L.; Yaqiong, Q.; Wenjuan, C. Differential analysis of puff-by-puff deliveries of routine chemical analyses and five key roasted sweet aroma components in mainstream smoke of slim cigarettes and normal cigarettes. *Acta Tab. Sin.* **2019**, *25*, 1−9.
(21) Xiaocui, X. The relationship between ventilation rate and physicochemical indicator of cigarette made by online laser punching. *J. Zhengzhou Univ. Light Ind., Nat. Sci.* **2015**, *30*, 52−56.
(22) Xianyue, Z.; Chi, C.; Yulong, X.; Xingliang, L.; Junlan, C.; Huaiqi, L.; Lei, H.; Tao, T.; Hongbo, W.; Xiaobing, Z.; Yujun, Z.; Guixin, P. Effects of filter ventilation on releases of major nitrogen-containing heterocyclic basicaroma components in mainstream cigarette smoke. *Tob. Sci. Technol.* **2019**, *52*, 52−59.
(23) Konstantinidis, E.; Matsouki, N.; Tsipa, C.; Gareiou, Z.; Drimili, E.; Vatikiotis, L.; Zervas, E. Development of a model linking TNCO emissions with filter ventilation in conventional cigarettes. *IOP Conf. Ser.: Earth and Environ. Sci.* **2021**, *899*, 012008.
(24) O'Connor, R. J.; Hammond, D.; McNeill, A.; King, B.; Kozlowski, L.; Giovino, G.; Cummings, K. How do different cigarette design features influence the standard tar yields of popular cigarette brands sold in different countries? *Tob. Control* **2008**, *17*, i1−i5.
(25) Stephens, W. E. Dependence of tar, nicotine and carbon monoxide yields on physical parameters: implications for exposure, emissions control and monitoring. *Tob. Control* **2007**, *16*, 170−176.
(26) Lindner, M.; Raunic Vadanjel, R. Plasma perforation of tipping paper-a novel method to generate ventilated filter cigarette. *Tob. Sci. Technol.* **2014**, *47*, 105−114.
(27) Xiaohang, Z.; Zhibin, Z.; Jifeng, W.; Ling, Z.; Wei, R.; Chungning, D.; Jingqiang, J.; Zhihua, H. Design and Application of On-line Laser Perforating Device for Tipping Paper. In *10th International Symposium on Advanced Optical Manufacturing and Testing Technologies: Advanced and Extreme Micro-nano Manufacturing Technologies*; Chengdu, Peoples R China, Jun 14−17, 2021.
(28) He, F.; Lin, Y.; Liu, W.; Ma, R.; Su, Q.; Zhao, H. Effect of On-line Laser Drilling on Smoke and Sensory Quality of Cigarette. *Food Ind.* **2020**, *41*, 86−88.
(29) Archontoulis, S. V.; Miguez, F. E. Nonlinear Regression Models and Applications in Agricultural Research. *Agron. J.* **2015**, *107*, 786−798.
(30) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer Science & Business Media, 1999.
(31) Cervantes, J.; Garcia-Lamont, F.; Rodríguez-Mazahua, L.; Lopez, A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing* **2020**, *408*, 189−215.
(32) Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1−27.
(33) Pedregosa, F.; Varoquaux, V.; Gramfort, A.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(34) Luo, H.; Nattino, G.; Pratola, M. Sparse Additive Gaussian Process Regression Luo, Nattino and Pratola. *J. Mach. Learn. Res.* **2022**, *23*, 1−34.
(35) Sigrist, F. Gaussian Process Boosting. *J. Mach. Learn. Res.* **2022**, *23*, 1−46.
(36) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; The MIT Press, 2006.

https://doi.org/10.1021/acsomega.4c08978
ACS Omega 2025, 10, 2908−2918