A hybrid evolutionary machine learning pipeline for longitudinal dispersion coefficient modeling

April 28, 2020

Abstract

1 Introduction

Various types of contaminants are discharged into water bodies across the world both point and non-point forms and these contaminants adversely affects the water quality [1]. Hence, providing an accurate pollution dispersion analysis in rivers a tedious issue for authorities responsible for monitoring and managing water resources. Even though pollution can disseminate in all directions of rivers [2], the longitudinal dispersion is normally dominant at a considerable distance from the pollution source after discharge into the water body [3]. As per environmental and hydrological researchers, the most important parameter when studying the longitudinal transport of pollution in the rivers is the longitudinal dispersion coefficient (K_x) [4, 5, 6, 7, 8, 9, 10, 11]. K_x accurate estimation is essential for numerous practical applications, such as environmental engineering, river engineering, intake designs, assessment of dangerous contaminants discharge into water bodies, etc. [12]. Several models exist which relies on the advection-dispersion equation (Eq. 1); however, the modelling results are highly affected by the nature of the K_x [13].

——— eq (1)

where K_x = the longitudinal dispersion coefficient, t = time, x and u = longitudinal coordinate and longitudinal velocity, respectively, and C = averaged cross-sectional concentration [14]. Regarding K_x, it can be determined experimentally, empirically, and theoretically [15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27]. The direct experimental determination of the dispersion coefficient demands cost and time inefficient tracer studies and could be done only with rectangular flumes data [5]. The determination of K_x through theoretical means is also difficult as there are no knowledge of the transverse profiles of both the flow depth and flow velocity [19]. The earlier predictive equations differ in their findings and are laden with certain levels of uncertainty; therefore, several studies have focused on the development of empirical models for K_x estimation [8, 28, 29]. Several studies have been devoted on discovering the

relationship between K_x and the hydrodynamic and hydraulic parameters [30, 31, 32] and some equations that model this relationship have been developed from these studies.

Over the past two decades, the implementation of computer aid models reported an optimistic results in modeling hydraulic engineering problems [33]. The importance of artificial intelligence (AI) techniques in environmental modelling has increased over the years. One among several studies conducted on the K_x modeling, an artificial neural network (ANN) model developed by Tayfur and Singh [2] for K_x prediction in rivers and natural streams. The evaluation of the model proved its capability in predicting the K_x compared to the earlier suggested empirical approaches. Another study by Tayfur [34] presented fuzzy, ANN, and regression-based models for the K_x prediction in natural streams. The prediction results showed that the developed models outperformed the existing empirical equations as they are satisfactorily predicted the measured data with minimum errors. A fuzzy model developed by Fuat Toprak and Savci [35] for K_x prediction in natural channels. The study compared the performance of the developed model with measured data and the existing equations and from the results, the fuzzy model performed better than the other techniques in terms of results reliability. Toprak and Cigizoglu [36] applied three ANN models based on different learning algorithms (i.e., the radial basis function neural network, feed forward back propagation, and the generalized regression neural network) for K_x estimation when evaluating its behavior in dispersion characteristics prediction in natural streams. The outcome of the study showed that the accuracy of the developed model was higher compared to the accuracy of the other existing empirical equations. The capacity of support vector machine (SVM) and adaptive neuro fuzzy inference system (ANFIS) models have been inspected by Noori et al. [37] for K_x prediction in natural streams. The study found that the SVM model performed better than the ANFIS model in terms of achieving better threshold statistical analysis. In another study [3], the authors developed ANFIS model for K_x simulation in rivers and natural streams. The research finding showed satisfactory K_x values prediction using the proposed model in comparison to the measured data. According to the authors conclusion, it is a good way of K_x prediction in streams but can be combined with other mathematical pollutant transfer models for real-time updating of such models.

A back-propagation neural network (BPNN) model with a 2D convergent flow tracer transport model developed by Shieh et al. [38] for improved evaluation of transverse and longitudinal dispersivities from a convergent flow tracer test. From the results, the developed model required less computational time and offered more accurate values of the transport parameter. The study also found the developed method as an effective way of achieving fast and accurate transverse and longitudinal dispersivities evaluation for a field convergent flow tracer test. An ANN model presented by Sahay [39] for K_x prediction in natural rivers. The performance of the developed model was compared with that of earlier reported models and found to be more accurate and precise. The study by Noori et al. [14], authors employed the ANN technique for K_x prediction in natural streams. The outcome of the study showed that the developed approach is applicable in river water quality management studies. An SVM approach constructed by Azamathulla and Wu [40] for K_x prediction in natural rivers. The study developed the proposed model based on published data on dispersion coefficient for a range of flow conditions. Research findings showed that the developed SVM model is applicable for accurate K_x prediction. A genetic programming (GP) model developed by [41] for modeling K_x.

According to the studies, practicing engineers can rely on the modern data driven approaches to improve their designs and evaluations when using GP for LDC prediction in natural rivers. The use of M5 model tree for LDC prediction has been reported by Etemad-Shahidi and Taghipour [5]. The model was compared to the other existing equations in terms of performance based on error measures and from the results, the developed model performed better than the existing formulas and could be a valuable tool for LDC prediction. The development of an empirical formula for LDC prediction in pipe flow based on the use of the evolutionary gene expression programming (GEP) has been presented by Sattar [42]. The GEP was used for the establishment of the empirical relationships between the LDC and some of the control parameters, such as the Reynolds number, the pipe friction coefficient, the average velocity, and the pipe diameter. The results showed that the proposed relations are simple and effective in LDC evaluation in pipe flow.

The development of an empirical formulae for LDC prediction in pipe flow based on the adaptive Neuro fuzzy group method of data handling has been presented by Najafzadeh and Sattar [43]. The evaluation of the proposed method showed that the proposed relations are simpler compared to the existing numerical solutions; it was also found as an effective way of evaluating the LDC in pipe flow. The study by Sattar and Gharabaghi [44] presented two GEP models for LDC prediction based on 150 published data sets of hydraulic and geometric parameters in natural streams. The analysis showed that the proposed relations were accurate, simple, and effective in LDC prediction in natural streams. The suitability of empirical formulas, RBF, and MLP neural network for LDC prediction in rivers has been evaluated by Parsaie and Haghiabi [45]. The outcome of the analysis showed that MLP performed best in LDC prediction while the RBF model performed a bit better than the empirical formulas in terms of accuracy. The PSO has been introduced by Najafzadeh and Tafarojnoruz [46] for improving the performance of neuro-fuzzy-based group method of data handling (NF-GMDH) in LDC prediction in rivers. The NF-GMDH-PSO model was compared with DE, MT, GA, ANN, and traditional empirical equations in terms of performance. From the evaluation, DE and GA methods outperformed the other among ANN-based equations. The reliability of ANN, ANFIS, and SVM in LDC prediction in natural rivers has been studied by Noori et al. [13]. The study performed forward selection (FS) and gamma test (GT) in order to sort the input variables in the order of their effects and relevance on LDC prediction. The outcome of the study revealed less uncertainty in the SVM model compared to the ANN and ANFIS models for LDC estimation in natural rivers; furthermore, the performance of the ANFIS model was found better than that of the ANN model. A multivariate adaptive regression splines (MARS) has been developed by Haghiabi [47] for LDC prediction in rivers. The performance of the MARS model was compared to that of multi-layer neural network model and empirical formulas and the outcome showed that the MARS model performed better than the multi-layer neural network model and empirical formulas in terms of accuracy. The study by Alizadeh et al. [48] relied on the GA, imperialist competitive algorithm (ICA), bee algorithm (BA), cuckoo search (CS), and Levenberg–Marquardt (LM) algorithm for the training of ANN models for LDC prediction in rivers. The evaluation of the models showed that they can be successfully used to improve the performance of ANN models. However, the performance of the CS, ICA and BA algorithms was better than that of GA and LM in training the ANN model. The use of the Bayesian network (BN) for LDC prediction in natural

rivers has been presented by Mohamad Javad Alizadeh et al. [49]. First, the study applied the clustering technique as a data preprocessing technique to cluster the data in separate groups of similar characteristics. The study showed that the developed model is a suitable way of pollutant transport prediction in natural rivers.

The use of evolutionary polynomial regression (EPR) for accurate prediction of Kx in rivers has been presented by Balf et al. [50]; the prediction was based on the flow depth, channel width, and average and shear velocities. The EPR model-predicted Kx was compared with those estimated using other conventional Kx estimation formulas and from the results, the introduced EPR model for Kx estimation was found suitable to be incorporated in one-dimensional water quality models for better solute concentration prediction in natural rivers. An ANFIS-based PCA method of LDC prediction has been developed by Parsaie et al.,[51] . The evaluation of the model showed better accuracy of the ANFIS model compared to the experimental formulas. The study by Riahi-Madvar et al. [52] presented a Pareto-Optimal-Multigene Genetic Programming (POMGGP) equation for LDC prediction kX. The study analyzed 503 data sets of channel geometry and flow conditions in natural streams in order to develop a hybrid model. The developed hybrid model is a combination of the Subset Selection of Maximum Dissimilarity Method (SSMD) with Multigene Genetic Programming (MGP) and Pareto-front optimization. The aim of the combined approach is to establish a set of selected dimensionless equations of kX and the best equation with the widest applicability. As per the authors, the proposed equation provided accurate prediction of kX compared to the other published equations; hence, it can be applied in LDC prediction in natural river flows.

A hybrid approach called GRC-ANN has been presented by Ghiasi et al. [53] for LDC estimation in natural rivers. The hybrid method is a combination of the granular computing (GRC) method with an ANN model. The performance of the hybrid model was compared with those of ANFIS, ANN, and other empirical models in terms of accuracy and performance in different LDC values. The outcome of the evaluation showed that the hybrid GRC-ANN approach performed better than the other LDC prediction methods. Saberi-Movahed et al., [54] relied on the ELM to develop a novel group method of data handling (GMDH) called GMDH-ELM for LDC prediction in water pipelines. PSO and GSA were employed to improve the feed forward structure of the GMDH-ELM model for LDC prediction. The analysis of the GMDH-ELM model showed that it achieved a good level of precision in the training and testing phases. The results further showed that the proposed GMDH-ELM performed better than other soft computing and conventional predictive models.

The study by Kargar et al. [55] examined the performance of SVR, Gaussian process regression, random forest, M5 model tree (M5P), and MLR in LDC prediction in natural streams. The study found the M5P model with simple formulations to exhibit better performance compared to the other machine learning and empirical models and was recommended as a suitable tool for LDC prediction in rivers. An integrated model has been introduced by Memarzadeh et al. [56] based on the Subset Selection of Maximum Dissimilarity (SSMD) method and the Whale Optimization Algorithm (a simple optimization approach). The study presented a high accuracy formula for LDC prediction which was proven to be superior in terms of LDC prediction compared to the existing LDC prediction formulas. The Whale optimization algorithm was also found applicable in improving the predictive performance of the equations in

other related fields by establishing the optimum coefficient values. The study by Riahi-Madvar et al.[57] relied on the SSMD and ANFIS hybridized with the firefly algorithm (FFA) to develop a hybrid system for LDC prediction. The FFA was used for the derivation of the optimum parameters of the ANFIS model. The analysis of the proposed ANFIS-FFA model showed that it was significantly improved compared to the normal ANFIS, suggesting that the parameters optimization by the nature-inspired optimization algorithms contributed significantly towards improving the generality of the ANFIS estimations.

2 Materials and Methods

2.1 Longitudinal Dispersion Data

The dataset consists of 71 samples collected from 29 rivers on the United States and available in [2]. From this dataset, fifty-one samples (72%) are used to compose the training set, while the remaining 20 samples (28%) comprise the test set. Table 1 presents a short description of the variables and also displays the statistical information of the variables in the dataset. The relative shear velocity, also known as friction term, can be associated with as the roughnesses and hydrodynamic characteristics of the river bed [5]. The shape channel shape parameter, given $\ln B/H$, reflects the vertical irregularities of the river bed [19]. The channel sinuosity is defined as the ratio of the channel length to the valley length [62].

Table 1: Statistical information of the dataset and short description of he variables.

Variable	Short description	min	mean	std	max
B (m)	channel width	11.90	82.96	122.50	711.20
H (m)	cross-sectional average flow depth	0.22	1.71	2.60	19.94
U (m/s)	cross-sectional average flow velocity	0.034	0.54	0.39	1.74
u^* (m/s)	shear velocity	0.002	0.088	0.089	0.553
U/u^*	relative shear velocity	13.80	51.68	31.72	156.50
Q (m^3/s)	flow discharge	1.29	7.62	4.56	19.63
β	channel shape parameter	2.62	3.79	0.57	5.05
σ	channel sinuosity	1.08	1.51	0.42	2.54
K_x (m^2/s)	longitudinal dispersion coefficient	1.90	107.71	170.25	892.00

2.2 Proposed Hybrid Method

The focus of this work is developing a machine learning pipeline for the prediction of the dispersion coefficient of natural channels. A machine learning pipeline comprises a set of chained steps that involves processing the data, selected the features, build the machine learning

estimator, and finally releasing the outputs. Also, machine learning pipelines reduce the human effort when building a reliable and accurate model [63].

Figure 1 depicts the machine learning pipeline proposed in this paper. The pipeline consists of three steps: data processing, feature selection, and model building. In the the data processing step, the samples in the data set are scaled according to

$$v_i = \frac{v_i}{v_{max}}, \quad v_i \in \{B, H, Q, U, u^*, U/u^*, \beta, \sigma, \}$$

where v_i is the standardized value, and v_{max} is the maximum values for the samples of the considered variable.

A simple procedure is used in the feature selection step. A feature vector, composed by 8 entries,

$$\mathbf{x}^{FS} = [x_B, x_H, x_Q, x_U, x_{u^*}, x_{U/u^*}, x_\beta, x_\sigma]$$

represent the eight variables. As entry equals to 1 represent a selected feature and 0 a removed feature. For example, the feature vector $\mathbf{x}^{FS} = 10100011$ means that the selected set $\{B, Q, \beta, \sigma\}$.

In the model building step

$\mathbf{x}^{MB} = [x_{k_0}, x_\nu, x_l, x_\alpha]$

$\mathbf{x}^{MB} = [x_C, x_\varepsilon, x_\gamma]$

The parametrization of the pipeline comprises the adjutment of two vectors $\mathbf{x}^{FS}$ and $\mathbf{x}^{MB}$. These two vector can be concatenated in a parametric vector

$$\mathbf{x} = [\mathbf{x}^{FS}, \mathbf{x}^{MB}]$$

where $\mathbf{x}^{FS}$ is the feature vector and $\mathbf{x}^{MB}$ is the vector of the internal parameters of the macine learning method.

$\mathbf{x} = [x_B, x_H, x_Q, x_U, x_{u^*}, x_{U/u^*}, x_\beta, x_\sigma, x_{k_0}, x_\nu, x_l, x_\alpha]$

$\mathbf{x} = [x_B, x_H, x_Q, x_U, x_{u^*}, x_{U/u^*}, x_\beta, x_\sigma, x_C, x_\varepsilon, x_\gamma]$

The objective function to be minimized is the Root Mean Discrepancy Ratio. The Discrepancy Ratio DR is used as error measure in the literature and, for each sample, it is written as

$$\mathrm{DR} = \log \frac{K_{xp}}{K_{xm}} \tag{1}$$

where K_{xm} is measured dispersion coefficient, and K_{xp} is predicted dispersion coefficient. When $K_{xp} = K_{xm}$, DR is equals to zero and there is an exact prediction. Otherwise, there is either an overprediction when $\mathrm{DR} > 0$, that means $K_{xp} > K_{xm}$, or an underprediction which leads to $\mathrm{DR} < 0$ and consequently $K_{xp} < K_{xm}$.

The Root Mean Discrepancy is averaged over all N number of observations

$$\mathrm{RMDR} = \frac{1}{N} \sum_{i=1}^{N} \left(\log \frac{K_{xp_i}}{K_{xm_i}} \right)^2 \tag{2}$$

where K_{xp_i} and K_{xm_i} are the prediction and observed dispersions.

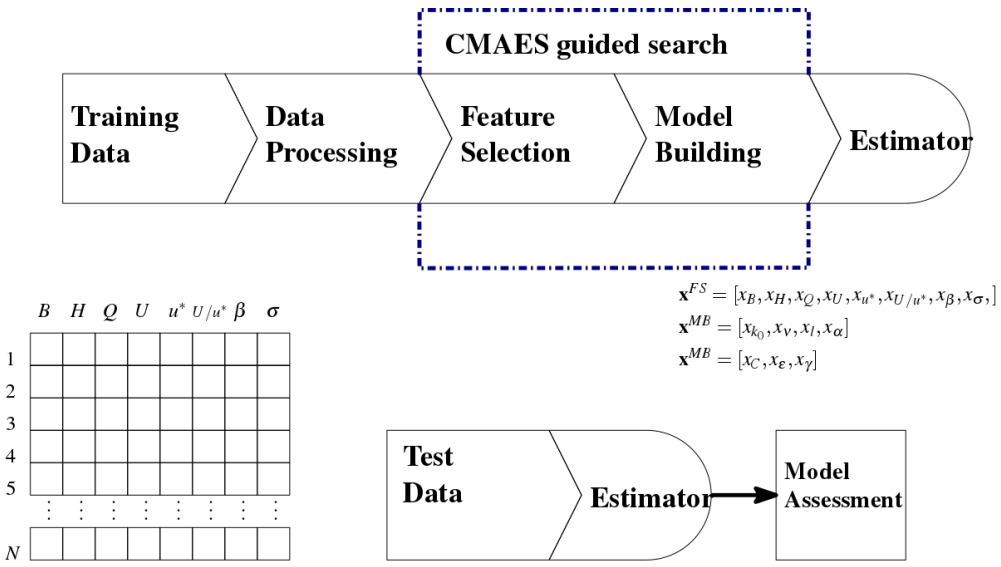

Figure 1: Proposed hybrid framework

2.3 Feature Selection Approach

Feature selection, also known as variable selection or attribute selection, is an independent process generally combined with the learning procedure of machine learning estimators, which aims to reduce the number of attributes to improve the performance of an estimator. The central premise of feature selection is to remove from the dataset the irrelevant or redundant features, so the quality of the final predictions is not deteriorated.

Depending on their interaction with the estimator method, feature selection techniques can be classified into filter models, wrapper models, and embedded models. Filter models work independently of the estimator using information form the dataset relying on a metric constructed upon several statistical tests, which can be a correlation coefficient, a distance function, or an importance measure. Wrapper models use an estimation model and select a subset of features based on the estimator performance. Although wapper methods produce better performance, they are computationally intensive and time-consuming than the filter methods [64]. In the embed methods, the selection of relevant features is integrated into the learning of the estimation model [65]. The selected features are those that best contribute to the accuracy of the estimation model [66].

2.4 Covariance Matrix Evolution Strategies

2.5 Gaussian Processes Regression

In Gaussian process regression (GPR), it is assumed the output $\hat{y}$ of a function $f(\mathbf{x})$ can be written as

$$\hat{y} = f(\mathbf{x}) + \varepsilon$$

with the noise term is considered as $\varepsilon \sim N(0,\sigma^2)$. The function $f(\mathbf{x})$ is distributed as a Gaussian Process. A Gaussian Process (GP) is completely specified by its mean $m(\mathbf{x})$ function and covariance function $k(\mathbf{x},\mathbf{x}')$[67]

$$f(\mathbf{x}) \approx GPR(m(\mathbf{x}),k(\mathbf{x},\mathbf{x}'))$$

where mean function reflects the expected function value, and for simplicity, it is assumed to be zero. The covariance function below models the dependence between the function values at different input points $\mathbf{x}$ and $\mathbf{x}'$ [68]:

$$k_\nu(||\mathbf{x}-\mathbf{x}'||) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(||\mathbf{x}-\mathbf{x}'||\sqrt{\frac{2\nu}{\theta}}\right)^{\nu} K_\nu\left(||\mathbf{x}-\mathbf{x}'||\sqrt{\frac{2\nu}{\theta}}\right) \tag{3}$$

where ν and θ are positive parameters and K_ν is the modified Bessel function [69].

2.5.1 Support Vector Regression – SVR

Support Vector Regression (SVR) is a version of the Support Vector Machine developed for regression analysis. SVR maps the input vectors $\mathbf{x} = [x_1,\ldots,x_N]$ into a high dimensional space where a linear machine build an optimal function $f(\mathbf{x})$ that minimizes the functional [70]

$$J = \frac{1}{N}\sum_{i=1}^{N} L_\varepsilon(y_i - f(\mathbf{x}_i))$$

where L_ε is the loss function, which penalizes the model in case of differences between the training data and model predictions (errors) and y_i is output data associated with $\mathbf{x}_i$. The ε-insensitive loss function [71]

$$L_\varepsilon(y-f(\mathbf{x})) = \begin{cases} 0 \text{ if } |y-f(\mathbf{x})| \le \varepsilon \\ |y-f(\mathbf{x})| \text{ otherwise} \end{cases}$$

where ε is a SVR parameter.

This optimization problem can be transformed into a dual problem

$$\begin{aligned} \max_{\alpha,\alpha*} \quad & -\tfrac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)K(\mathbf{x}_i,\mathbf{x}_j) \\ & -\varepsilon\sum_{i=1}^{N}(\alpha_i+\alpha_i^*) - \varepsilon\sum_{i=1}^{N} y_i(\alpha_i+\alpha_i^*) \\ \text{subject to} \quad & \left\{\sum_{i=1}^{N} y_i(\alpha_i+\alpha_i^*) = 0 0 \ge \alpha_i \ge C,\ 0 \ge \alpha_i^* \ge C,\ i = 1,...,N \right. \end{aligned} \tag{4}$$

where α_i and α_i^* are the weights which determine the influence of each data point on the model (support vectors were the data with non-zero weights), $K(x_i,x_j)$ is the kernel function, and C is the regularization parameter, which determined the trade-off between the training errors and model complexity.

The optimization problem is decomposed into sub-problems, which were solved step by step. At each step, the algorithm selects two Lagrange multipliers, found their optimal values analytically, and updated the SVR function [70]

$$f(\mathbf{x}) = \sum_{i=1}^{N} (\alpha_i + \alpha_i^*) K(x_i, x_j) + b$$

where b is a constant threshold. The process was repeated until the Lagrange multipliers converged. The radial basis kernel function of the form

$$K(x_i, x) = \sum_{i=1}^{m} \exp(-\gamma \|x_i - x_j\|^2)) \tag{5}$$

where γ is the bandwidth parameter.

3 Computational Experiments and Discussion

The machine learning pipeline proposed in this paper were applied to predict the coefficient of longitudinal dispersion, and their performance was compared with the models proposed in [2]. A robust procedure was proposed to report the results: (1) perform 100 independent runs establishing a unique random seed for each run; (2) for each run, the hybrid pipeline models built on the training set are used to estimate the dispersion coefficients in the test set (3) calculate the metrics RMSE and Accuracy to asses the performance of each model. The computational experiments were divided into three parts:

1. Modeling LDC without feature selection. Perform experiments using the machine learning approach proposed considering all features (Case 0) and also reproduce the models presented in [2] (Cases 1-7) shown in Table 2. For these experiments, the vector $\mathbf{x}^{FS}$ is kept constant, depending on the LDC model. In this first part, only the machine learning parametric vector $\mathbf{x}^{MB}$ is adjusted to fine-tunning the machine learning model.

2. Modeling LDC using evolutionary feature selection. Employ the hybrid pipeline shown in Figure 1 to simultaneously search for the most suitable feature set and the model internal parameters. The feature selection allows for exploring combinations of features that are not presented in Table 2, and can potentially produce better results when used as inputs for GPR and SVR.

3. Modeling LDC with the most frequent features. Build machine learning models using the most frequent features found in the previous experiment, exploiting the specific knowledge on the features to propose a more accurate model.

Considering the Table 2, Case 0 employs the original features to model the dispersion coefficient of natural streams. Cases 1-7 were proposed in [2], where a neural network was used to to predict K_x. Case 1 considers three features involving flow and geometric characteristics: the

Table 2: Input features associated with models tested in this paper. Case 0 involves all input variables and Cases 1-7 were proposed by [2].

Case	$\mathbf{x}^{FS}$ (fixed mask)	Active features
Case 0	11111111	$U, H, B, u^*, Q, U/u^*, \beta, \sigma$
Case 1	11100000	U, H, B
Case 2	00001000	Q
Case 3	10000000	U
Case 4	10000010	U, β
Case 5	10000011	U, β, σ
Case 6	00000100	U/u^*
Case 7	00000111	$U/u^*, \beta, \sigma$

flow velocity U, the flow depth H, and the channel width B. Case 2 employs only flow discharge Q as the input feature. As the velocity plays an important role in the forecast of the effects of a pollutant spill, Case 3 considered only flow velocity U. Case 4 considered flow velocity U and channel shape parameter β as input variables, while Case 5 considered channel sinuosity σ in the feature vector $\mathbf{x}^{FS}$ along with the channel shape parameter β and flow velocity U. Case 6 considered only the relative shear velocity U/u^*. Finally, Case 7 considered the channel shaper parameter β and channel sinuosity σ along with the relative shear velocity $U/u*$ as input features.

3.1 Modeling LDC without feature selection

Figure 2 compares the results for LDC modeling without feature selection. The variable setups are shown in Table 2. According to Figure 2, we can observe for Case 0, the average value of RMSE is considerably lower than those produced by the reference model for both GPR and SVR models. The averaged accuracy obtained with GPR model was competitive with those produced by the reference model, but there was a decrease in performance for the accuracy of the SVR model. Notably, for the case where the width-to-depth ratios are small than 50, the RMSE ($B/H < 50$) values were remarkably better than the reference model, representing a decrease of 64.9% for the GPR and 67.8% for SVR. This finding shows that the hybrid method can provide accurate predictions for narrow channels. On the other hand, the hybrid model did not produce better results than the reference model when the extreme values were not considered, as can be observed for root mean squared errors less than 100 m^2/s, RMSE ($K_x < 100$).

The results for Case 1 are interesting since the features used as input variables in this model are a subset of the collected ones. The features use in this model, U, H and B, are those exhibit higher correlations with the dispersion coefficient as can be seen in Figure 3. Figure 4 show the strength of the linear relationships for U, H and B. It is important to notice that, although informative, the linear correlation may not be able to represent the complex and nonlinear nature of LDC [13].

Considering the results in Figure 2, although GPR performed better in accuracy, SVR produced smaller RMSE when compared to GPR and the reference model. In addition, GPR

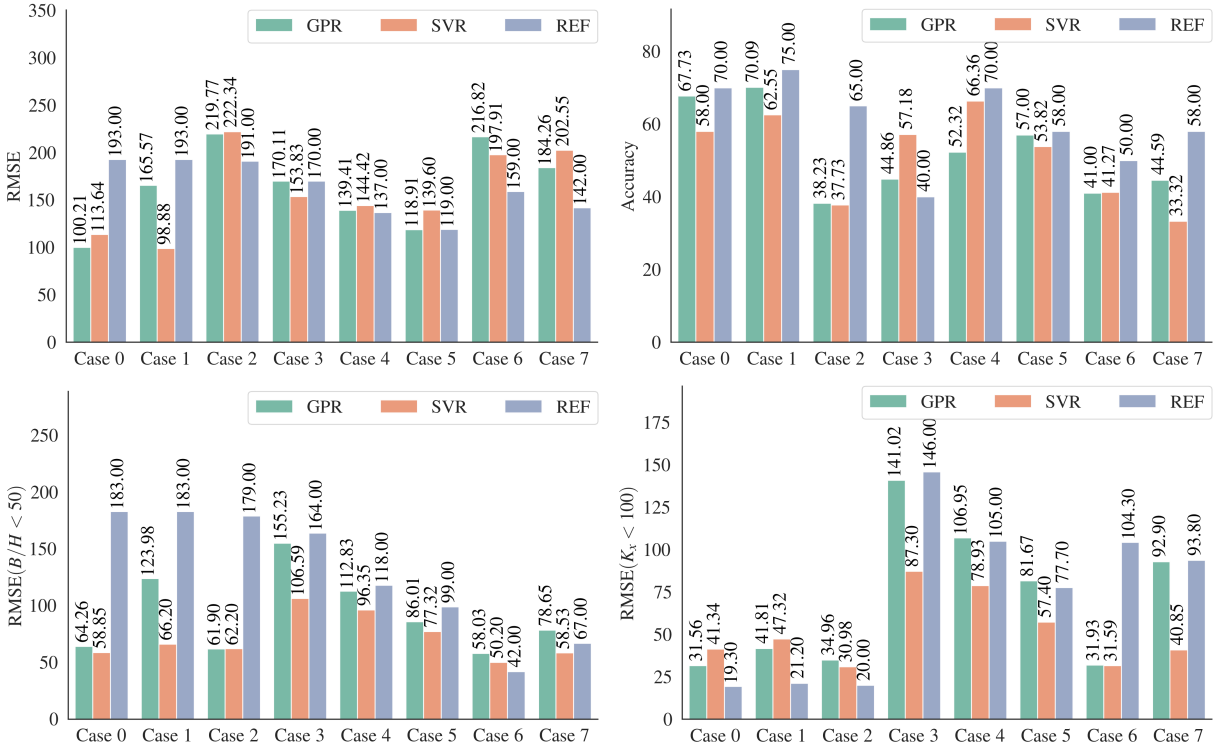

Figure 2: Longitudinal Dispersion Coefficient modeling using GPR and SVM. The results indicated with label REF were collected from [2].

predicts poorly the dispersion coefficient for narrow channels, as can be observed in the higher values for RMSE if $(B/H < 50)$ compared to SVR. However, considerint the RMSE $(B/H < 50)$, both models produced better predictions then the reference model.

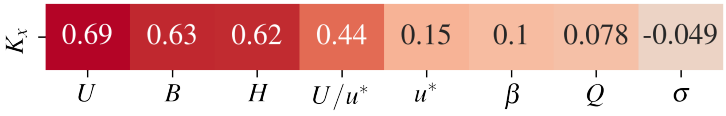

Figure 3: Correlation coefficients using Spearman rank correlations between the eight input variables and the longitudinal dispersion coefficient.

For Cases 2, 3, and 6, the hybrid approach performed poorly. It is interesting to notice that Case 2 involves the features implicitly in Case 1 since the flow discharge is the product of flow depth, velocity, and channel width, $Q = HUB$. In addition, the linear correlation between the flow discharge and the dispersion can be discarded due to the small correlation as shown in Figure 3. The models for Cases 2, 3, and 6 are built upon one variable, which is not enough to represent the nonlinear relationship between the single feature and the dispersion coefficient. The finding corroborates the results described in [2], that solely the flow discharge (Q), flow velocity (U), and the relative shear velocity (u^*) are not sufficient to predict the nonlinear behavior of the

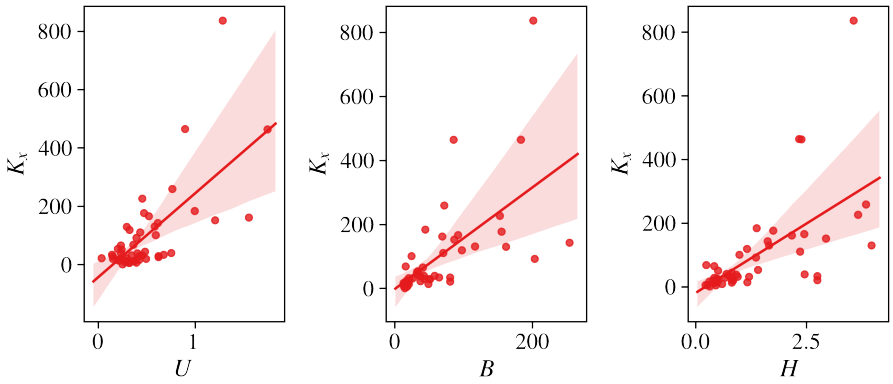

Figure 4: Scatter plots demonstrating visually the relationship among U, B, and H features and the dispersion coefficient.

dispersion coefficient.

Cases 4 is built upon Case 3, with the inclusion of the channel shape parameter (β) for Case 4, while in Case 5, the channel sinuosity is added to Case 4. The hybrid model performed similarly to the artificial neural network developed in the reference paper. However, it was not able to improve the results when compared to the previous cases.

Case 7 considered the relative shear velocity (U/u^*) as in Case 6, and channel shape parameter (β) along with the channel sinuosity (σ) to predict the dispersion coefficient. The metrics shown in Figure 2 show the inclusion of the channel shape parameter and the channel sinuosity did not improve the K_x estimations.

The results depicted in Figure 2 obtained after several independent runs allows for draw the following conclusions on the test Cases presented in Table 2. From Cases 0-7, GPR and SVR have produced the overall best results for Case 0 and Case 1, as can be seen in Figure 2. Given the performance of the GPR and SVR models in Cases 0 and 1, the comparative analysis will be focused on these two cases. The Root Mean Squared Error of both models was inferior to the artificial neural network developed in [2]. In Case 0, where the eight variables are taken into account, the GPR and SVR models showed an average reduction of 58.9% and 58.8% for the RMSE, respectively. Similar behavior was observed for narrow channels where $B/H < 50$: the reduction was 35.1% and 32.2% for GPR and SVR, respectively. Also, the averaged Accuracy for GPR is only 3.24% small them the reference model. We emphasize that the results are the average of 100 independent runs.

In Case 1 the averaged RMSE values obtained by GPR and SVR were smaller than the RMSE attained by the reference model. However, the RMSE produced by GPR was higher than that provided by SVR. Similar behavior is observed for Root Mean Squared Error of narrow channels ($B/H < 50$). For the same Case 1, GPR produced an averaged Accuracy of 70.09% against 75% for the reference model, resulting in a percentage difference of 6.55% on average.

For Cases 0 and 1, Figure 5 depicts the scatterplots of the predicted and estimated dispersion coefficient for the best models in the test set (according to RMSE). Besides the RMSE, the accuracy and the coefficient of determination associated with each model are also shown.

Although a few predicted K_x present a visible error, the predictions for GPR and SVR models agree well with the observations. This can be verified by the R^2 values, above 0.90 for for GPR in Cases 0 and 1. For SVR, $R^2 = 0.93$ for Case 0 and $R^2 = 0.89$ for Case 1.

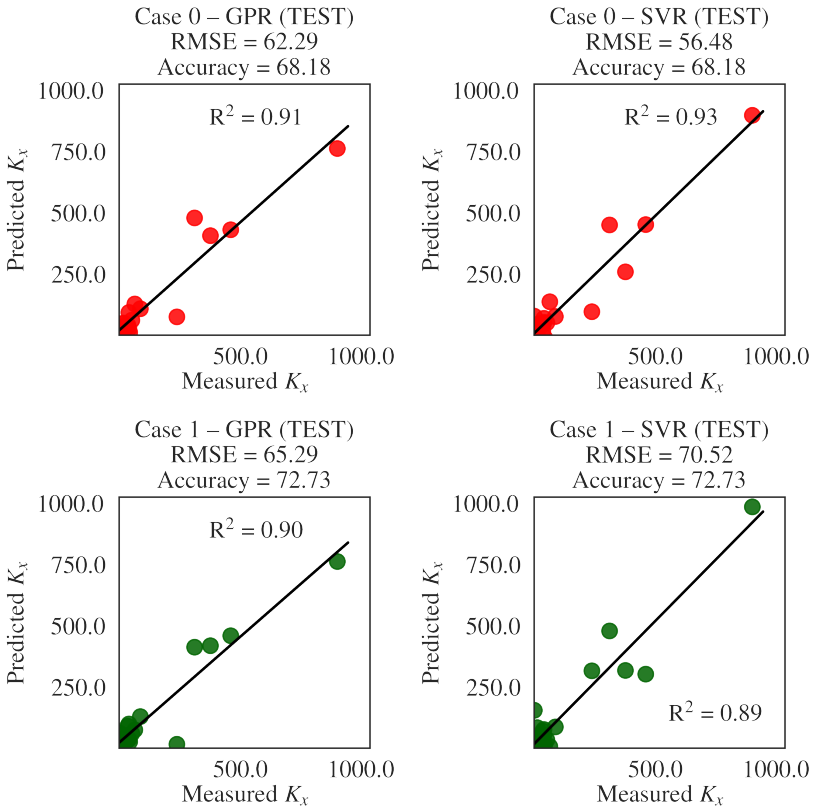

Figure 5: Scatter plots for the best models according to Root Mean Squared Error (RMSE).

In the computational experiments, a total of 100 independent runs were performed. As a consequence, it is interesting to analyze the distribution of the internal parameters in all executions. As Cases 0 and 1 produced better results compared to the other ones, the discussion is focused on these cases. Figure 6 shows for Cases 0-7 the distribution of the GPR parameters, namely, k_0, ν, l, and α. The first three parameters control the shape of the kernel shown in Equation 3 used to compute the estimations, while α, which is added to the diagonal of the kernel matrix during fitting, rules the level of noise in the observations. The parameter k_0 controls the magnitude of the GPR approximation and the final solutions showed similar distributions in cases 0 and 1. Similar behavior was noticed for the parameter α which is associated with noise in the observations. The dispersion is a measure associated with a natural process, so that noise becomes an important factor in the estimates. The parameter ν is a critical paramater in the kernel function and controls the GPR smoothness [72]: the smaller ν the less smooth the approximated solution is. The distribution of ν in Figure 6 shows the GPR estimation functions produced for case for Case 1 are smoother than Case 0. In addition, the length scales l for Case 1 are higher than Case 0 as well. These combined circumstances potentially led to the highest RMSE values

for case 1, as can be seen in Figure 2.

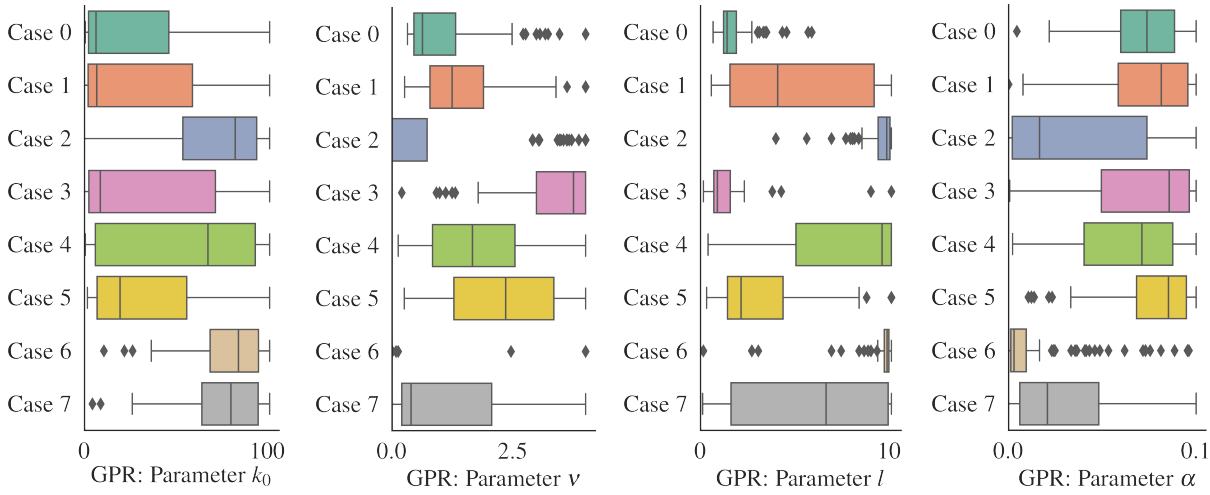

Figure 6: Distribution of the internal parameters for the GPR models over 100 independent runs.

Figure 7 displays the distribution of the SVR internal parameters C, γ and ε. C is the penalization parameter, ε is the specifies the penalization associated in the training loss function, and γ is kernel coefficient in Equation (5). Analyzing the boxplots, one can observe the distribution of the parameters γ and ε showed similar distributions in both cases. On the other hand, the values of parameter C for Case 1 are higher than those found for Case 0. In support vector machines, C plays an important role as a regularization parameter. The regularization strength is inversely proportional to C. As a result, the SVR estimations for Case 0 are smoother than those produced for Case 1. The smaller RMSE values for Case 1 shown in the barplots in Figure 7 (averaged in 100 runs) supports this interpretation.

When compared to the reference model, it can be observed that GPR and SVM presented a lower performance in predicting the dispersion coefficient when the extreme values ($K_x > 100$) are discarded for Cases 0 and 1. One possible explanation to the deprecated performance of RMSE for small K_x is that those models do not explore in a suitable manner the relationship among the features for small dispersion values. For case 0, the interaction among the eight features might be not beneficial to estimate small K_x values while for Case1 the features U, B, and H, that conserve linear relationship with the dispersion coefficient might not be enough to represent the nonlinearities of the dispersion coefficient. Considering that the evolutionary strategy appropriately determines the parameters $\mathbf{x}^{MB}$ of the models, an alternative is to choose a set of input variables that allow balancing the predictive resources with extreme values. Likewise, this set of variables should produce estimates with low RMSE values at high precision.

3.2 Modeling LDC with evolutionary feature selection

Considering the experiments in the last section, it is not clear whether other combinations of variables can lead to better results for RMSE and accuracy. Despite the explanation provided by

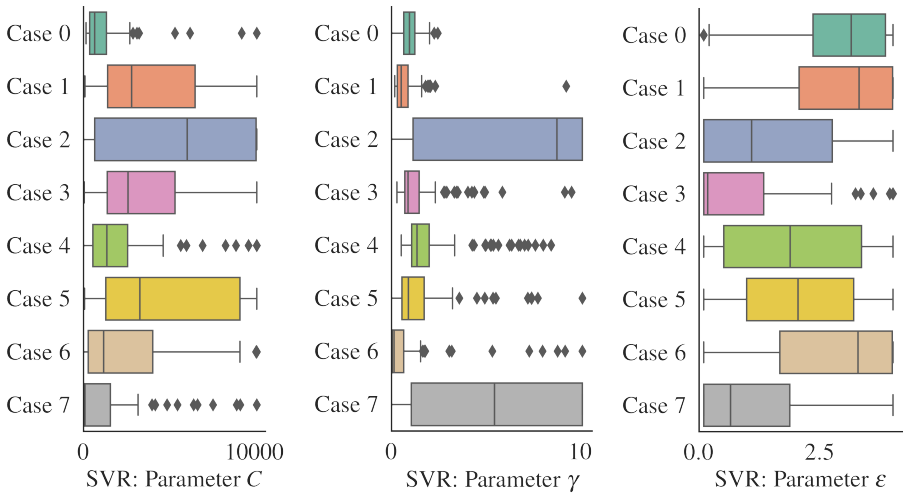

Figure 7: Distribution of the internal parameters for the SVR models over 100 independent runs.

the authors in their study [2], it can be argued that the variables associated with the cases shown in Table 2 were chosen, in their major extent, based on the correlation with the longitudinal dispersion coefficient. Figures 3 and 4 show that the variables with the highest correlations are those that comprise the sets in Table 2. For example, the U variable appears in 5 out of 8 cases. Besides, the variables U, B, and H are part of Cases 1 and 2, which produced the best results.

The hybrid approach developed here can search for the most suitable variables along with the estimator to explore combinations that not previously considered. It can be noticed that due to the nonlinear natures of the estimators GPR and SVR, a careful choice of variables can produce an improved performance of the estimators. Moreover, variables with smalls correlations with LDC may positively impact the final predictions. The next experiments aim at analyzing the impact of the evolutionary feature selection on the assessment of the performance metrics on the LDC prediction.

Table 3 summarized the results obtained for GPR and SVR assitsed by feature selection in the pipeline shown in Figure 1. The first column shows the cases, and the second column displays the estimator. The Root Mean Squared Error is shown in the third column, while the Accuracy appears in the fourth column. The fifth two columns show the Root Mean Squared Error when extreme values are discarded and the last one the RMSE for narrow channels. Analyzing the results in Table 3, it is possible to observe that the feature selection step on the pipeline has contributed considerably to improve the metrics produced by the machine learning methods. Table 3 also shows the results for Cases 1 and 2 to compare the performance of the models when the feature selection was not applied. Overall, the RMSE values decreased for both GPR and SVR models. As observed for the models without implementing feature selection, GPR-FS and SVR-FS produced smaller Root Mean Squared Errors than the reference model. GPR-FS model achieved, on average, an improvement of 16.5% when compared to Case 0 and 50.2% concerning Case 1. The improvement achieved by the model SVR-FS to Cases 0 and 1 were 21.0% and 21.6%, respectively.

Table 3: Hybrid evolutionary approach with feature selection. The best results are shown in boldface with standard deviation in parentheses. Values indicated with – were not available at the source.

Case	Estimator	RMSE	Accuracy	RMSE($K_x < 100$)	RMSE($B/H < 50$)
Case 0	GPR	99.70 (29.05)	67.91 (5.05)	31.80 (3.97)	63.13 (10.63)
	SVR	112.34 (26.40)	57.73 (7.06)	41.23 (4.28)	58.80 (7.58)
	REF	193.00 (–)	70.00 (–)	**19.30** (–)	183.00 (–)
Case 1	GPR	167.34 (115.45)	71.05 (6.19)	41.28 (8.16)	114.68 (98.36)
	SVR	113.22 (35.35)	62.59 (5.96)	45.18 (8.84)	61.73 (14.43)
	REF	193.00 (–)	75.00 (–)	21.20 (–)	183.00 (–)
–	GPR-FS	**83.26** (26.41)	**76.73** (6.60)	31.25 (5.18)	52.41 (14.61)
–	SVR-FS	88.78 (13.66)	65.77 (8.99)	39.51 (8.42)	**46.10** (7.21)

Regarding Accuracy, the SVR-FS model achieved a slight improvement over the SVR model. On the other hand, the GPR-FS model achieved a consistent improvement when compared to the model without feature selection. In particular, the averaged Accuracy is the highest among all models. When the extreme dispersion coefficients ($K_x > 100$) were excluded, there was no improvement in the RMSE for both methods. However, the evolutionary pipeline implementing GPR and SVR models allowed us to reach the lowest RMSE values for narrow channels. In particular, the average values of RMSE achieved by the SVF-FS model can be highlighted, as can be seen in the last column of Table 3. Figure 8 depicts the scatter plots for the best GPR-FS and SVR-FS models, that were chosen according to the smaller RMSE. Considering GPR-FS, it can be observed that RMSE deacreased while Accuracy increased when comparing the scatter plots from Figures 5 and 8.

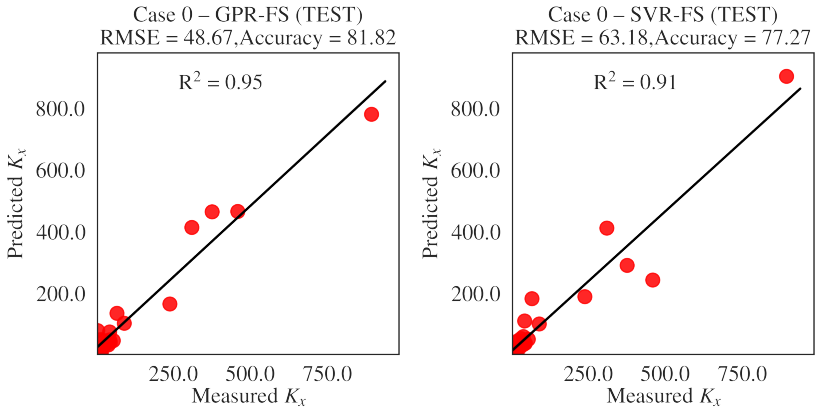

Figure 8: Scatter plots for the best models implementing evolutionary feature selection according to Root Mean Squared Error (RMSE).

The analysis of these results allows us to conclude that the feature selection step in the pipeline leads to a considerable improvement in the prediction of the dispersion coefficient. The proper selection of input variables is beneficial for the pipeline model as this interaction increases the accuracy while decreases the averaged errors.

The results reported in Table 3 allow us to assess the improvement in performance, but they do not show which variables led to these gains. Figure 9 shows the distribution of the selected features after training the evolutionary pipeline. From this figure, one can observe that GPF-FS and SVR-FS chose a total of 13 sets of features at the end of the evolutionary search procedure. We observed the variable selection process is heuristic and there is no guarantee of choosing the same set of variables in all runs. The barplot on the left side shows that for GPR-FS model, the set of variables B, U, σ, Q has been selected 24 times, the B, U, σ was selected 18 times and B, U, σ, β 13 times. These three sets were selected 55 out of 100 times, and they share the variables B, U and σ. A similar analysis for the SVR-FS model in the bar graph of Figure 9 shows that B, U, σ, β was chosen 33 times, B, U, σ 20 times, and B, U, σ 12 times. These three sets were selected 74 times over the course of 100 independent executions, approximately two thirds of the total runs.

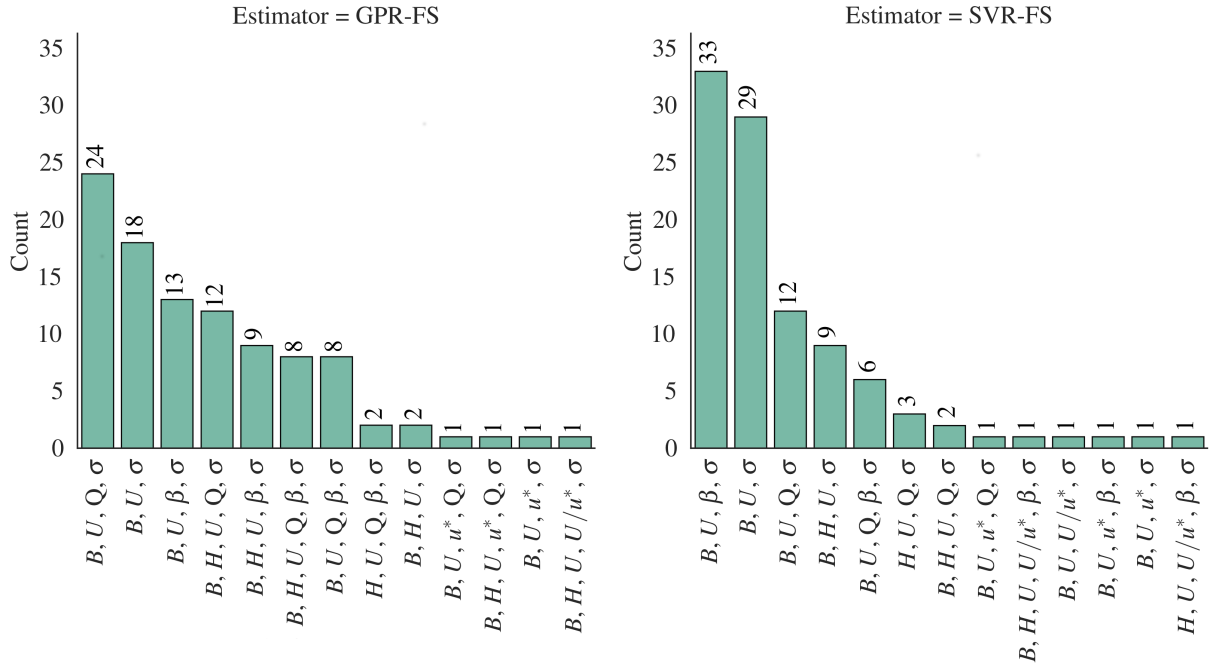

Figure 9: Distribution of sets of active features for evolutionary feature selection approach over 100 runs.

Figure 10 shows that U, σ, B, Q were the most frequent variables for the GPR-FS model, in that order. For the SVR-FS model, the most frequent ones were U, σ, B, β, in this order as well. Moreover, for both models, the variables U and σ were select in all runs, while B appears 98 times for the GPR-FS model and 96 times for the SVR-FS model. From most frequent variables, U and B have a strong correlation with the dispersion coefficient, whereas river sinuosity σ

has the lowest correlation, as can be seen in Figure 10. The feature selection has shown the sinuosity plays an important rule in the context of the machine learning pipeline. In addition, the correlation analysis in Figure 3 suggests that σ holds a strong non-linear relationship with K_x.

Some studies discuss that sinuosity is related to river meandering [20] other exogenous factors such as vegetation [73, 74] and topography [75, 76], that also affect the dispersion coefficient. As shown in the results, the incorporation of sinuosity is beneficial for the performance of the gaussian processes and support vector machines.

3.3 Modeling LDC using the most frequent features found by evolutionary selection

The computational experiments carried out in this study allow us to determine the importance of variables indirectly. This importance takes into account their impact on the machine learning pipeline. The results shown in Table 3 and Figure 9 suggest that some variables constitute the core that improves the learning of the machine learning models. A detailed analysis of the frequency that the variables appear in the selected sets is presented in Figure 10. Observing this figure, it is possible to verify that the variable σ appears in all selected sets. The same occurs with the variable U. Based on this analysis, we can see that the most frequent features are U, B, and σ. These three variables are used to propose the Case 8 are shown in Table 3.

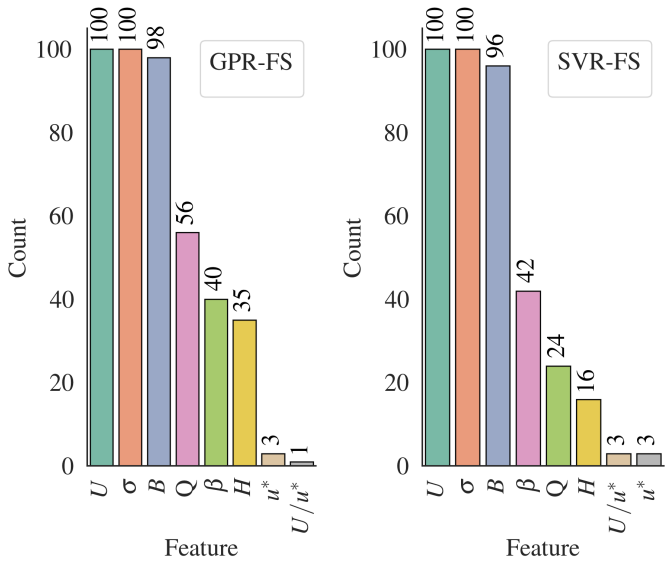

Figure 10: Hybrid model using evolutionary feature selection approach. Number of times each variable appears in the active sets over 100 runs.

A set of computational experiments were conducted on the Case 8, and the machine learning methods were trained using the features U, B, and σ. The results in Table 5 shows the improvement of all metrics. The averaged values show the improvement for GPR concerning the Root Mean Squared Error, Accuracy and RMSE for dispersion coefficients below 100 m^2/s.

Table 4: Feature sets suggested by the analysis of evolutionary feature selection.

Case	$\mathbf{x}^{FS}$	Active variables
Case 8	10100001	U, B, σ

This expected because the carefully choosing process of the most frequent variables represents the indirect introduction of specific knowledge about the features. In addition, this specific knowledge could only be obtained after analyzing the various simulations carried out with the evolutionary selection model. The most frequent variables allow capturing the essential behavior of the dispersion coefficient leadinf to an effective model to longitudinal dispersion coefficient. However, when this knowledge is not available before carrying out the estimations, the feature selection step in the pipeline is indispensable.

Table 5: LDC modeling with the most frequent features found by the evolutionary selection. The best results are shown in boldface with standard deviation in parentheses.

Case	Estimator	RMSE	Accuracy	RMSE($K_x < 100$)	RMSE($B/H < 50$)
–	GPR-FS	83.26 (26.41)	76.73 (6.60)	31.25 (5.18)	52.41 (14.61)
–	SVR-FS	88.78 (13.66)	65.77 (8.99)	39.51 (8.42)	46.10 (7.21)
Case 8	GPR	**82.86** (24.41)	**77.50** (5.43)	**28.85** (4.83)	53.12 (15.56)
Case 8	SVR	94.07 (15.45)	73.27 (4.33)	32.97 (7.97)	**42.06** (7.97)

Figure 11 shows the scatter plots for the best models in Case 8 according to RMSE. The plots of the predicted and observed dispersion coefficients show the models yelds accurate resutls.

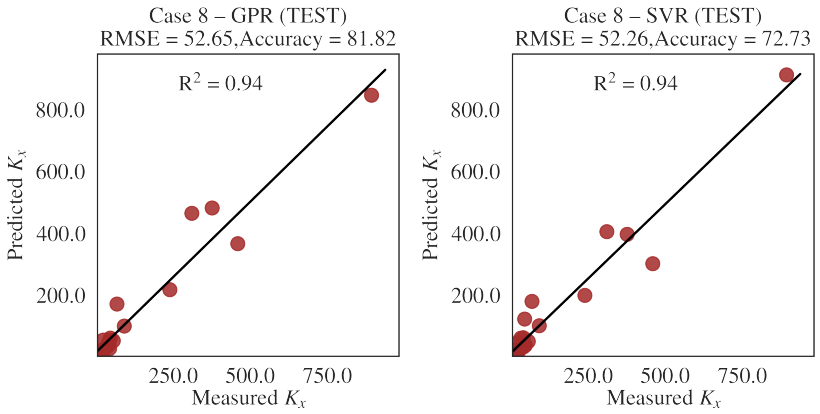

Figure 11: Scatter plots for the best models in Case 8 according to Root Mean Squared Error.

Further research includes multiobjective feature selection [77].

4 Conclusions

The dispersion of pollutants in rivers involves complex phenomena, requiring simplifications and assumptions, leading to drawbacks to accurate predictions. In this context, machine learning methods arise as an interesting alternative to dispersion coefficient modeling. However, the performance of these methods depends on the proper choice of the input variables.

In this paper, we have applied a machine learning pipeline that embeds a simple approach to the task of feature selection of dispersion coefficient of natural channels. The task of adjusting selected features and adjusting the parameters is assisted by an evolution strategy. Based on several studies on the literature, there is no attempt to integrate the feature selection to the model building step for predicting the dispersion coefficient. In general the feature selection task is performed in a separated step, and doe not considers the estimation model.

The model proposed in this paper automatically searches for the features and the internal parameters of the machine learning models. The performance of the models was measured based on the root mean squared error and accuracy. The results show that by the selection of the most suitable features, more efficient models for dispersion coefficient can be constructed.

References

[1] Rooholah Noori, Reza Kerachian, Ahmad Khodadadi Darban, and Ahmad Shakibaienia. Assessment of importance of water quality monitoring stations using principal components analysis and factor analysis: a case study of the karoon river. *J. of Water and Wastewater*, 63:60–69, 2007.

[2] Gokmen Tayfur and Vijay P Singh. Predicting longitudinal dispersion coefficient in natural streams by artificial neural network. *Journal of Hydraulic Engineering*, 131(11):991–1000, 2005.

[3] Hossien Riahi-Madvar, Seyed Ali Ayyoubzadeh, Ehsan Khadangi, and Mohammad Mehdi Ebadzadeh. An expert system for predicting longitudinal dispersion coefficient in natural streams by using anfis. *Expert Systems with Applications*, 36(4):8589–8596, 2009.

[4] Seyed M Kashefipour and Roger A Falconer. Longitudinal dispersion coefficients in natural channels. *Water Research*, 36(6):1596–1608, 2002.

[5] Amir Etemad-Shahidi and Milad Taghipour. Predicting longitudinal dispersion coefficient in natural streams using m5 model tree. *Journal of hydraulic engineering*, 138(6):542–554, 2012.

[6] Hossein Hamidifar, Mohammad Hossein Omid, and Alireza Keshavarzi. Longitudinal dispersion in waterways with vegetated floodplain. *Ecological engineering*, 84:398–407, 2015.

[7] Majid Dehghani, Mansoor Zargar, Hossien Riahi-Madvar, and Rasoul Memarzadeh. A novel approach for longitudinal dispersion coefficient estimation via tri-variate archimedean copulas. *Journal of Hydrology*, 584:124662, 2020.

[8] Yu-Fei Wang, Wen-Xin Huai, and Wei-Jie Wang. Physically sound formula for longitudinal dispersion coefficients of natural rivers. *Journal of Hydrology*, 544:511–523, 2017.

[9] Maryam Farzadkhoo, Alireza Keshavarzi, Hossein Hamidifar, and Mahmood Javan. A comparative study of longitudinal dispersion models in rigid vegetated compound meandering channels. *Journal of environmental management*, 217:78–89, 2018.

[10] Kyong Oh Baek. Deriving longitudinal dispersion coefficient based on shiono and knight model in open channel. *Journal of Hydraulic Engineering*, 145(3):06019001, 2019.

[11] Maryam Farzadkhoo, Alireza Keshavarzi, Hossein Hamidifar, and James Ball. Flow and longitudinal dispersion in channel with partly rigid floodplain vegetation. In *Proceedings of the Institution of Civil Engineers-Water Management*, volume 172, pages 229–240. Thomas Telford Ltd, 2019.

[12] Mohamad Javad Alizadeh, Davoud Ahmadyar, and Ali Afghantoloee. Improvement on the existing equations for predicting longitudinal dispersion coefficient. *Water resources management*, 31(6):1777–1794, 2017.

[13] Roohollah Noori, Zhiqiang Deng, Amin Kiaghadi, and Fatemeh Torabi Kachoosangi. How reliable are ann, anfis, and svm techniques for predicting longitudinal dispersion coefficient in natural rivers? *Journal of Hydraulic Engineering*, 142(1):04015039, 2016.

[14] R Noori, AR Karbassi, H Mehdizadeh, M Vesali-Naseh, and MS Sabahi. A framework development for predicting the longitudinal dispersion coefficient in natural streams using an artificial neural network. *Environmental Progress & Sustainable Energy*, 30(3):439–449, 2011.

[15] Eliana Perucca, Carlo Camporeale, and Luca Ridolfi. Estimation of the dispersion coefficient in rivers with riparian vegetation. *Advances in Water Resources*, 32(1):78–87, 2009.

[16] Yufei Wang and Wenxin Huai. Estimating the longitudinal dispersion coefficient in straight natural rivers. *Journal of Hydraulic Engineering*, 142(11):04016048, 2016.

[17] Coefficient Using Field Experimental Data. Estimation of longitudinal dispersion coefficient using field experimental data and 1d numerical model of solute transport. In *Advanced Technologies, Systems, and Applications IV-Proceedings of the International Symposium on Innovative and Interdisciplinary Applications of Advanced Technologies (IAT 2019)*, volume 83, page 305. Springer, 2019.

[18] Hata Milišić, Emina Hadžić, and Suvada Jusić. Estimation of longitudinal dispersion coefficient using field experimental data and 1d numerical model of solute transport. In *International Symposium on Innovative and Interdisciplinary Applications of Advanced Technologies*, pages 305–323. Springer, 2019.

[19] Zhi-Qiang Deng, Vijay P Singh, and Lars Bengtsson. Longitudinal dispersion coefficient in straight rivers. *Journal of hydraulic engineering*, 127(11):919–927, 2001.

[20] Z-Q Deng, Lars Bengtsson, VP Singh, and DD Adrian. Longitudinal dispersion coefficient in single-channel streams. *Journal of Hydraulic Engineering*, 128(10):901–916, 2002.

[21] Il Won Seo and Kyong Oh Baek. Estimation of the longitudinal dispersion coefficient using the velocity profile in natural streams. *Journal of hydraulic engineering*, 130(3):227–236, 2004.

[22] Jaehyun Shin, Il Won Seo, and Donghae Baek. Longitudinal and transverse dispersion coefficients of 2d contaminant transport model for mixing analysis in open channels. *Journal of Hydrology*, page 124302, 2019.

[23] Il Won Seo and Tae Sung Cheong. Predicting longitudinal dispersion coefficient in natural streams. *Journal of hydraulic engineering*, 124(1):25–32, 1998.

[24] Prabhata K Swamee, Santosh K Pathak, and Mohammad Sohrab. Empirical relations for longitudinal dispersion in streams. *Journal of Environmental Engineering*, 126(11):1056–1062, 2000.

[25] Yuhong Zeng and Wenxin Huai. Estimation of longitudinal dispersion coefficient in rivers. *Journal of hydro-environment research*, 8(1):2–8, 2014.

[26] T Disley, B Gharabaghi, AA Mahboubi, and EA McBean. Predictive equation for longitudinal dispersion coefficient. *Hydrological processes*, 29(2):161–172, 2015.

[27] VV Camacho Suarez, ANA Schellart, W Brevis, and JD Shucksmith. Quantifying the impact of uncertainty within the longitudinal dispersion coefficient on concentration dynamics and regulatory compliance in rivers. *Water Resources Research*, 55(5):4393–4409, 2019.

[28] Abdusselam Altunkaynak. Prediction of longitudinal dispersion coefficient in natural streams by prediction map. *Journal of hydro-environment research*, 12:105–116, 2016.

[29] Hosein Nezaratian, Javad Zahiri, and Seyed Mahmood Kashefipour. Sensitivity analysis of empirical and data-driven models on longitudinal dispersion coefficient in streams. *Environmental Processes*, 5(4):833–858, 2018.

[30] Sinan Sahin. An empirical approach for determining longitudinal dispersion coefficients in rivers. *Environmental Processes*, 1(3):277–285, 2014.

[31] Bulent Tutmez and Mehmet Yuceer. Regression kriging analysis for longitudinal dispersion coefficient. *Water resources management*, 27(9):3307–3318, 2013.

[32] Chaopeng Shen, Jie Niu, Eric J Anderson, and Mantha S Phanikumar. Estimating longitudinal dispersion in rivers using acoustic doppler current profilers. *Advances in Water Resources*, 33(6):615–623, 2010.

[33] Ahmad Sharafati, Masoud Haghbin, Davide Motta, and Zaher Mundher Yaseen. The application of soft computing models and empirical formulations for hydraulic structure scouring depth simulation: A comprehensive review, assessment and possible future research direction. *Archives of Computational Methods in Engineering*, pages 1–25, 2019.

[34] Gokmen Tayfur. Fuzzy, ann, and regression models to predict longitudinal dispersion coefficient in natural streams. *Hydrology Research*, 37(2):143–164, 2006.

[35] Z Fuat Toprak and M Emin Savci. Longitudinal dispersion coefficient modeling in natural channels using fuzzy logic. *CLEAN–Soil, Air, Water*, 35(6):626–637, 2007.

[36] Z Fuat Toprak and Hikmet Kerem Cigizoglu. Predicting longitudinal dispersion coefficient in natural streams by artificial intelligence methods. *Hydrological Processes: An International Journal*, 22(20):4106–4129, 2008.

[37] Roohollah Noori, Abdulreza Karbassi, Ashkan Farokhnia, and Majid Dehghani. Predicting the longitudinal dispersion coefficient using support vector machine and adaptive neuro-fuzzy inference system techniques. *Environmental Engineering Science*, 26(10):1503–1510, 2009.

[38] Hung-Yu Shieh, Jui-Sheng Chen, Chun-Nan Lin, Wei-Kuang Wang, and Chen-Wuing Liu. Development of an artificial neural network model for determination of longitudinal and transverse dispersivities in a convergent flow tracer test. *Journal of hydrology*, 391(3-4):367–376, 2010.

[39] Rajeev Ranjan Sahay. Prediction of longitudinal dispersion coefficients in natural rivers using artificial neural network. *Environmental Fluid Mechanics*, 11(3):247–261, 2011.

[40] H Md Azamathulla and Fu-Chun Wu. Support vector machine approach for longitudinal dispersion coefficients in natural streams. *Applied Soft Computing*, 11(2):2902–2905, 2011.

[41] Qin Tu, Hong Li, Xinkun Wang, and Chao Chen. Ant colony optimization for the design of small-scale irrigation systems. *Water Resources Management*, 29(7):2323–2339, 2015.

[42] Ahmed MA Sattar. Gene expression models for the prediction of longitudinal dispersion coefficients in transitional and turbulent pipe flow. *Journal of Pipeline Systems Engineering and Practice*, 5(1):04013011, 2014.

[43] Mohammad Najafzadeh and Ahmed MA Sattar. Neuro-fuzzy gmdh approach to predict longitudinal dispersion in water networks. *Water Resources Management*, 29(7):2205–2219, 2015.

[44] Ahmed MA Sattar and Bahram Gharabaghi. Gene expression models for prediction of longitudinal dispersion coefficient in streams. *Journal of Hydrology*, 524:587–596, 2015.

[45] Abbas Parsaie and Amir Hamzeh Haghiabi. Predicting the longitudinal dispersion coefficient by radial basis function neural network. *Modeling earth systems and environment*, 1(4):34, 2015.

[46] Mohammad Najafzadeh and Ali Tafarojnoruz. Evaluation of neuro-fuzzy gmdh-based particle swarm optimization to predict longitudinal dispersion coefficient in rivers. *Environmental Earth Sciences*, 75(2):157, 2016.

[47] Amir Hamzeh Haghiabi. Prediction of longitudinal dispersion coefficient using multivariate adaptive regression splines. *Journal of Earth System Science*, 125(5):985–995, 2016.

[48] MJ Alizadeh, A Shabani, and MR Kavianpour. Predicting longitudinal dispersion coefficient using ann with metaheuristic training algorithms. *International journal of environmental science and technology*, 14(11):2399–2410, 2017.

[49] Mohamad Javad Alizadeh, Hosein Shahheydari, Mohammad Reza Kavianpour, Hamid Shamloo, and Reza Barati. Prediction of longitudinal dispersion coefficient in natural rivers using a cluster-based bayesian network. *Environmental Earth Sciences*, 76(2):86, 2017.

[50] Mohammad Rezaie Balf, Roohollah Noori, Ronny Berndtsson, Alireza Ghaemi, and Behzad Ghiasi. Evolutionary polynomial regression approach to predict longitudinal dispersion coefficient in rivers. *Journal of Water Supply: Research and Technology-Aqua*, 67(5):447–457, 2018.

[51] Abbas Parsaie, Samad Emamgholizadeh, Hazi Mohammad Azamathulla, and Amir Hamzeh Haghiabi. Anfis-based pca to predict the longitudinal dispersion coefficient in rivers. *International Journal of Hydrology Science and Technology*, 8(4):410–424, 2018.

[52] Hossien Riahi-Madvar, Majid Dehghani, Akram Seifi, and Vijay P Singh. Pareto optimal multigene genetic programming for prediction of longitudinal dispersion coefficient. *Water resources management*, 33(3):905–921, 2019.

[53] Behzad Ghiasi, Hossein Sheikhian, Amin Zeynolabedin, and Mohammad Hossein Niksokhan. Granular computing–neural network model for prediction of longitudinal dispersion coefficients in rivers. *Water Science and Technology*, 80(10):1880–1892, 2019.

[54] Farid Saberi-Movahed, Mohammad Najafzadeh, and Adel Mehrpooya. Receiving more accurate predictions for longitudinal dispersion coefficients in water pipelines: Training group method of data handling using extreme learning machine conceptions. *Water Resources Management*, 34(2):529–561, 2020.

[55] Katayoun Kargar, Saeed Samadianfard, Javad Parsa, Narjes Nabipour, Shahaboddin Shamshirband, Amir Mosavi, and Kwok-wing Chau. Estimating longitudinal dispersion coefficient in natural streams using empirical models and machine learning algorithms. *Engineering Applications of Computational Fluid Mechanics*, 14(1):311–322, 2020.

[56] Rasoul Memarzadeh, Hossein Ghayoumi Zadeh, Majid Dehghani, Hossien Riahi-Madvar, Akram Seifi, and Seyed Mostafa Mortazavi. A novel equation for longitudinal dispersion coefficient prediction based on the hybrid of ssmd and whale optimization algorithm. *Science of The Total Environment*, 716:137007, 2020.

[57] Hossien Riahi-Madvar, Majid Dehghani, Kulwinder Singh Parmar, et al. Improvements in the explicit estimation of pollutant dispersion coefficient in rivers by subset selection of maximum dissimilarity hybridized with anfis-firefly algorithm (ffa). *IEEE Access*, 2020.

[58] Behzad Ghiasi, Hossein Sheikhian, Amin Zeynolabedin, and Mohammad Hossein Niksokhan. Granular computing–neural network model for prediction of longitudinal dispersion coefficients in rivers. *Water Science and Technology*, 80(10):1880–1892, 01 2020.

[59] Katayoun Kargar, Saeed Samadianfard, Javad Parsa, Narjes Nabipour, Shahaboddin Shamshirband, Amir Mosavi, and Kwok wing Chau. Estimating longitudinal dispersion coefficient in natural streams using empirical models and machine learning algorithms. *Engineering Applications of Computational Fluid Mechanics*, 14(1):311–322, 2020.

[60] Rasoul Memarzadeh, Hossein Ghayoumi Zadeh, Majid Dehghani, Hossien Riahi-Madvar, Akram Seifi, and Seyed Mostafa Mortazavi. A novel equation for longitudinal dispersion coefficient prediction based on the hybrid of SSMD and whale optimization algorithm. *Science of The Total Environment*, 716:137007, 2020.

[61] Adam P. Piotrowski, Pawel M. Rowinski, and Jaroslaw J. Napiorkowski. Comparison of evolutionary computation techniques for noise injected neural network training to estimate longitudinal dispersion coefficients in rivers. *Expert Systems with Applications*, 39(1):1354 – 1361, 2012.

[62] Rajeev Ranjan Sahay. Predicting longitudinal dispersion coefficients in sinuous rivers by genetic algorithm. *Journal of Hydrology and Hydromechanics*, 61(3):214 – 221, 2013.

[63] Shixia Liu, Xiting Wang, Mengchen Liu, and Jun Zhu. Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics*, 1(1):48 – 56, 2017.

[64] Ping Tan, Xin Wang, and Yong Wang. Dimensionality reduction in evolutionary algorithms-based feature selection for motor imagery brain-computer interface. *Swarm and Evolutionary Computation*, 52:100597, 2020.

[65] Jie Cai, Jiawei Luo, Shulin Wang, and Sheng Yang. Feature selection in machine learning: A new perspective. *Neurocomputing*, 300:70 – 79, 2018.

[66] Ali Mirzaei, Yalda Mohsenzadeh, and Hamid Sheikhzadeh. Variational relevant sample-feature machine: A fully bayesian approach for embedded feature selection. *Neurocomputing*, 241:181 – 190, 2017.

[67] CE. Rasmussen and CKI. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA, January 2006.

[68] Eric Schulz, Maarten Speekenbrink, and Andreas Krause. A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, 85:1 – 16, 2018.

[69] Rahul Kumar. The generalized modified bessel function and its connection with voigt line profile and humbert functions. *Advances in Applied Mathematics*, 114:101986, 2020.

[70] Kennedy Were, Dieu Tien Bui, Øystein B. Dick, and Bal Ram Singh. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an afromontane landscape. *Ecological Indicators*, 52:394 – 403, 2015.

[71] Steve R Gunn et al. Support vector machines for classification and regression. *ISIS technical report*, 14, 1998.

[72] Michael L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Series in Statistics. Springer-Verlag New York, 1 edition, 1999.

[73] Carlo Camporeale and Luca Ridolfi. Interplay among river meandering, discharge stochasticity and riparian vegetation. *Journal of Hydrology*, 382(1):138 – 144, 2010.

[74] J. Savickis, A. Bottacin-Busolin, M. Zaramella, N. Sabokrouhiyeh, and A. Marion. Effect of a meandering channel on wetland performance. *Journal of Hydrology*, 535:204 – 210, 2016.

[75] Gábor Timár. Controls on channel sinuosity changes: a case study of the tisza river, the great hungarian plain. *Quaternary Science Reviews*, 22(20):2199 – 2207, 2003. Fluvial response to rapid environmental change.

[76] W. M. van Dijk, R. Teske, W. I. van de Lageweg, and M. G. Kleinhans. Effects of vegetation distribution on experimental river channel dynamics. *Water Resources Research*, 49(11):7558–7574, 2013.

[77] F. Jiménez, G. Sánchez, J.M. García, G. Sciavicco, and L. Miralles. Multi-objective evolutionary feature selection for online sales forecasting. *Neurocomputing*, 234:75 – 92, 2017.