water

MDPI

Article

Application of Oversampling Techniques for Enhanced Transverse Dispersion Coefficient Estimation Performance Using Machine Learning Regression

Sunmi Lee and Inhwan Park *

Department of Civil Engineering, Seoul National University of Science and Technology, 232 Gongreung-ro, Nowon-Gu, Seoul 01811, Republic of Korea; sunmi619@seoultech.ac.kr
* Correspondence: ihpark@seoultech.ac.kr

Abstract: The advection–dispersion equation has been widely used to analyze the intermediate field mixing of pollutants in natural streams. The dispersion coefficient, manipulating the dispersion term of the advection–dispersion equation, is a crucial parameter in predicting the transport distance and contaminated area in the water body. In this study, the transverse dispersion coefficient was estimated using machine learning regression methods applied to oversampled datasets. Previous research datasets used for this estimation were biased toward width-to-depth ratio (W/H) values ≤ 50, potentially leading to inaccuracies in estimating the transverse dispersion coefficient for datasets with $W/H > 50$. To address this issue, four oversampling techniques were employed to augment the dataset with $W/H > 50$, thereby mitigating the dataset's imbalance. The estimation results obtained from data resampling with nonlinear regression method demonstrated improved prediction accuracy compared to the pre-oversampling results. Notably, the combination of adaptive synthetic sampling (ADASYN) and eXtreme Gradient Boosting regression (XGBoost) exhibited improved accuracy compared to other combinations of oversampling techniques and nonlinear regression methods. Through the combined ADASYN–XGBoost approach, it is possible to enhance the transverse dispersion coefficient estimation performance using only two variables, W/H and bed friction effects (U/U^*), without adding channel sinuosity; this represents the effects of secondary currents.

Keywords: transverse dispersion coefficient; imbalanced dataset; data oversampling; machine learning; nonlinear regression

Citation: Lee, S.; Park, I. Application of Oversampling Techniques for Enhanced Transverse Dispersion Coefficient Estimation Performance Using Machine Learning Regression. *Water* **2024**, *16*, 1359. https://doi.org/10.3390/w16101359

Academic Editor: Constantinos V. Chrysikopoulos

Received: 24 March 2024
Revised: 7 May 2024
Accepted: 9 May 2024
Published: 10 May 2024

1. Introduction

Water quality management is a significant task for public health and aquatic environments. The mixing stages of introduced polluted water in natural rivers are classified into three processes: near-, intermediate-, and far-field mixing. In near-field mixing, longitudinal, transverse, and vertical mixing simultaneously occur by turbulent and molecular diffusion. After finishing the vertical mixing, intermediate-field mixing begins with longitudinal and transverse dispersion. Intermediate-field mixing persists over significantly longer distances than near-field mixing due to the complex flow structures accompanying the delays in transverse mixing completion, which are caused by the irregular channel geometries [1]. In those mixing processes, the advection–dispersion equation has been used for the analysis of polluted water mixing in aquatic environments, such as rivers, lakes, and water conveyance channels. In particular, in intermediate-field mixing, following the completion of vertical mixing, the depth-averaged two-dimensional advection–dispersion equation (2D ADE) has been widely used [2–5]. The 2D ADE is defined as follows:

$$\frac{\partial C}{\partial t} + u\frac{\partial C}{\partial x} + v\frac{\partial C}{\partial y} = D_L\frac{\partial^2 C}{\partial x^2} + D_T\frac{\partial^2 C}{\partial y^2} \quad (1)$$

where C is the depth-averaged concentration; u, v are the depth-averaged longitudinal and transverse velocities, respectively; D_L, D_T represent the longitudinal and transverse dispersion coefficients, respectively. From the 2D ADE, mixing behaviors, such as the arrival time of polluted water, the concentration change, and the polluted area over longitudinal and transverse distances, can be predicted by the appropriate determinations of D_L and D_T. D_T is particularly significant in analyzing lateral mixing of polluted water caused by accidentally spilled pollutants, suspended solids, and continuous sources from tributaries and wastewater treatment effluents.

Tracer tests have been conducted to estimate D_T for laboratory channels [1,6–8] and natural rivers [9–12]. However, the tracer test is a labor-intensive and costly experiment, and tracer materials input is limited for natural streams, especially in large-scale rivers [13]. Thus, practically, empirical formulas have been used to estimate D_T using hydraulic (velocity magnitude, shear velocity, and Froude number) and geometrical (depth, width, radius of curvature, and sinuosity) parameters. Fischer et al. [14] defined that D_T is proportional to HU^*, where H is the flow depth and U^* is the shear velocity; they suggested a proportional constant of 0.15 for straight channels and one of 0.6 for meandering channels. Rutherford [15] presented the range of the proportional constant according to the geometrical properties of rivers, where 0.15–0.3 could be used in straight channels, and 0.3–0.9 could be used in meandering channels. For expanding applicability, the empirical formulas were developed by conducting multiple linear regression using tracer test results [10,16–20]. The accuracy of estimated D_T from the proposed empirical formulas depends on the diversity of data reflecting various hydraulic conditions that influence transverse mixing, which are used to develop the formula. In other words, the empirical formula is limited to specific river conditions used for the regression [10]. Furthermore, the unexplainable nonlinear relationship between D_T and complex flow structures of natural rivers raises uncertainty in the estimation of D_T using empirical formulas.

The data-driven approach can be a solution to unraveling complex relations between input and output data [21]. The soft computing technique has begun to be used for the estimation of the longitudinal dispersion coefficient for the far-field mixing due to sufficient tracer test datasets [22–28]. In recent studies, Sattar and Gharabaghi [24] compiled 150 datasets from natural streams for adopting the machine learning technique, and Ghiasi et al. [27] used 503 datasets from laboratory channels and natural streams. These researchers presented results of superior accuracy compared to the proposed empirical formulas derived by multiple linear regression. For intermediate mixing analysis, D_T has also been estimated using a machine learning model [19,29–33]. In these studies, 165–420 datasets, a significant portion of which included lab-scale results, were adopted to develop machine learning models, and the estimated D_T showed enhanced performance compared to the empirical formulae. However, the performance enhancement of machine learning models for D_T would be mitigated in natural rivers because the datasets used in previous studies are biased to lab-scale results. Therefore, it can be seen that the trained machine learning model has potential to overfit lab-scale data, resulting in errors when applied in natural rivers. To resolve the limitations of such imbalanced datasets, field-scale data need to be used in compensation.

Recent studies have introduced strategies for overcoming the disadvantages that are encountered due to imbalanced training datasets through data oversampling of minority class data. The Synthetic Minority Oversampling Technique (SMOTE) is an algorithm that beings balance between majority and minority data classes by generating new data samples for the minority data class [34]. The SMOTE is adopted for a data preprocessing technique and supports the enhancement of the performance of machine learning models by mitigating overfitting problems [35]. For imbalanced water quality and quantity data, the SMOTE has been used to improve data balance for the enhancement of prediction performance using machine learning techniques [36–40]. Furthermore, to improve SMOTE, an adaptive synthetic sampling (ADASYN) was proposed, introducing a density distribution to determine the number of synthetic samples [41]. Research has been

conducted on resolving water quality data imbalances and improving predictive performance using machine learning models with ADASYN [36]. Additional techniques, such as combining undersampling methods with oversampling techniques for the removal of samples from synthetically generated data, have been proposed. Hybrid approaches, like SMOTE-ENN and SMOTE-Tomek, incorporating Edited Nearest Neighbor (ENN) and Tomek-link techniques for noise and duplicate data removal from SMOTE-generated data, have been suggested [42]. Studies have also presented streamflow data prediction and flood forecasting using such hybrid techniques [43–45]. From the improvements shown in previous research, the imbalanced datasets of D_T can be improved through oversampling techniques, but such research has not been reported until now.

This study aims to enhance D_T estimation performance using two variables: width-to-depth ratio (W/H) and bed friction (U/U^*). Here, W is the channel width, and U is the cross-sectional averaged velocity. This aim will be achieved through data oversampling techniques by compensating for the imbalanced dataset comprising lab-scale data. In this study, four oversampling techniques, SMOTE, SMOTE-ENN, ADASYN, and SVM-SMOTE, were employed to reduce the data imbalance. Using the improved datasets, D_T was estimated using multiple linear regression (MLR), and three nonlinear regression methods were used: *k*-nearest neighbor's regression (KNR), support vector regression (SVR), and eXtreme Gradient Boosting regression (XGBoost). By comparing the accuracy of D_T estimation, a feasible combination of oversampling techniques and regression methods was proposed; the effectiveness of data oversampling in enhancing accuracy was discussed in comparison to the effectiveness of adding sinuosity for estimating D_T.

2. Materials and Methods

2.1. Dataset Explanations

The statistical properties of the collected dataset were analyzed to establish regression models for D_T. In total, 216 datasets were collected, consisting of 160 from laboratory channels and 56 from natural streams [1,6,8,9,11,46–71]. Laboratory experiments were predominantly conducted in straight channels, while 12 datasets [1,6] were obtained from meandering channels, exhibiting sinuosity ranging from 1.32 to 1.7. The corresponding Froude numbers for these experiments ranged from 0.032 to 0.972. Field experiments were conducted across streams in the USA, Canada, Europe, China, and South Korea; these were characterized by sinuosity ranging from 1.0 to 2.38, and the Froude number was in the range of 0.06–0.48. Fluorescent dye (specifically, Rhodamine B and Rhodamine WT) and neutrally buoyant solutions (such as nigrosine solution, gentian violet dye, carbon tetrachloride–benzine solution, etc.) were utilized for tracer tests in laboratory experiments; field experiments also employed fluorescent dye (Rhodamine B and Rhodamine WT). Table 1 presents the statistical properties for the laboratory channels and natural streams separately, focusing on W/H, U/U^* and D_T/HU^*. Both data groups exhibited similar ranges and average values of U/U^*. However, the average values of W/H and D_T/HU^* in natural streams were larger than those in the laboratory channels.

Table 1. Statistical properties of collected tracer test results.

	Laboratory Channels (No. of Datasets = 160)			**Natural Streams (No. of Datasets = 56)**		
	W/H	U/U^*	D_T/HU^*	W/H	U/U^*	D_T/HU^*
Max	65.1	24.6	0.70	169.5	25.7	1.21
Min	0.1	1.6	0.05	14.4	3.7	0.12
Average	17.7	12.5	0.16	67.9	12.8	0.51
Median	14.7	11.9	0.14	57.4	11.0	0.49
Standard Deviation	12.8	4.9	0.07	40.3	6.0	0.24

Figure 1 depicts histograms illustrating the distributions of the datasets. The statistical analysis revealed comparable distributions of U/U^* in both the laboratory and the natural

stream datasets. In contrast, W/H from the lab-scale experiments tended to accumulate in the range of $W/H < 50$, resulting in relatively small values of D_T/HU^* compared to datasets from natural streams. Significantly, the abundance of lab-scale datasets, approximately three times larger than those from the natural streams, raises concerns about the narrowing applicability of empirical formulas for D_T/HU^*, developed from imbalanced datasets. To enhance the applicability of empirical formulas, it is imperative to augment the dataset by acquiring more experimental data of the natural stream scale.

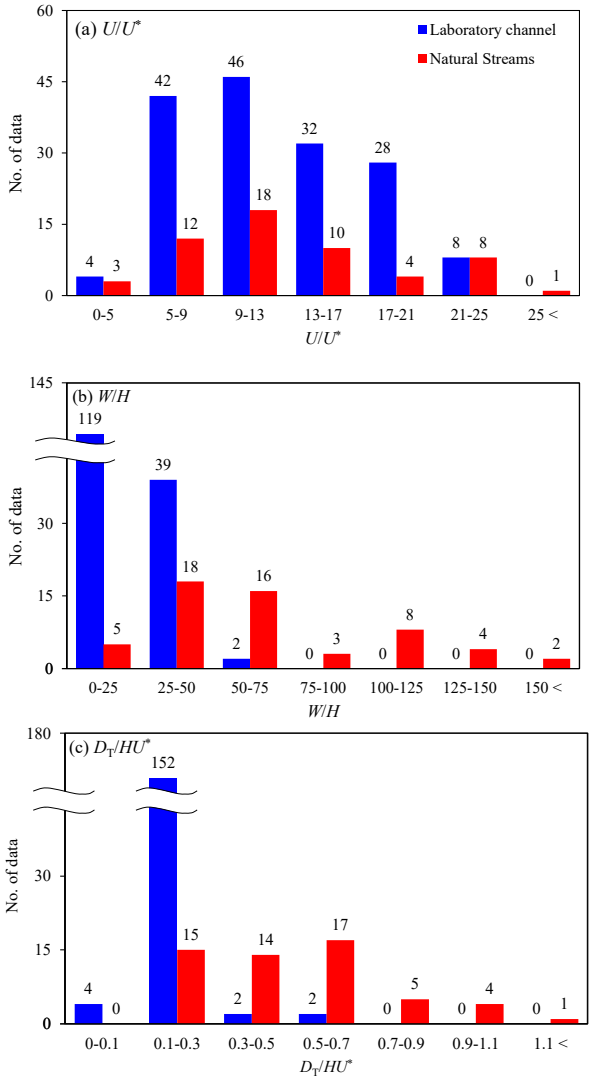

Figure 1. Distribution properties of the tracer test datasets according to the ranges of W/H.

2.2. Estimation of D_T

In this study, D_T was estimated using MLR, which is the traditional approach to obtain D_T, and the nonlinear regression algorithms, which are SVR, XGBoost, and KNR; these are known to be efficient for the regression of nonlinear datasets [72–74]. Through dimensional analysis and theoretical derivations, dimensionless hydraulic parameters were derived to formulate empirical expressions for the dimensionless transverse dispersion coefficient (D_T/HU^*), as follows:

$$\frac{D_T}{HU^*} = f\left(\frac{W}{H}, \frac{U}{U^*}, \frac{W}{R_c}, \frac{H}{R_c}, S_n\right) \quad (2)$$

where R_c is the radius of curvature; S_n is the channel sinuosity [10,18,71]. Table 2 presents the empirical formulas proposed using dimensionless hydraulic parameters suggested in previous studies. D_T is primarily influenced by the vertical profiles of transverse velocity,

as derived theoretically by Fischer et al. [14]. Therefore, hydraulic parameters such as bed friction (U/U^*) and geometrical configurations (W/H, H/R_c, W/R_c, and S_n), affecting vertical variations of transverse velocity, were incorporated into the formulas. The formulas, proposed by Jeon et al., Baek and Seo, and Yotsukura and Sayre [10,18,70], consider R_c or S_n to account for the effects of secondary currents. However, obtaining R_c and S_n is challenging due to the lack of information in the datasets. For instance, Jeon et al., and Baek and Seo [10,18] collected S_n data from 16 and 18 field datasets, respectively. Aghababaei et al. [19] gathered 230 datasets, but only 49, including 29 field cases and 20 flume experiments, were available for S_n. Consequently, establishing explainable relations between D_T and S_n from tracer test results, especially for natural streams with a large width-to-depth ratio, is challenging [12]. For these reasons, in this study, W/H and U/U^* were considered as the input variables for estimating D_T.

Table 2. Empirical formulas for estimating transverse dispersion coefficient.

References	Empirical Formulas	Method
Yotsukura and Sayre [70]	$\frac{D_T}{HU^*} = 0.4\left(\frac{U}{U^*}\right)^2\left(\frac{W}{R_c}\right)^2$	MLR
Bansal [50]	$\frac{D_T}{HU^*} = 0.002\left(\frac{W}{H}\right)^{1.498}$	
Deng et al. [17]	$\frac{D_T}{HU^*} = 0.145 + \left(\frac{1}{3530}\right)\left(\frac{U}{U^*}\right)\left(\frac{W}{H}\right)^{1.38}$	
Jeon et al. [10]	$\frac{D_T}{HU^*} = 0.03\left(\frac{U}{U^*}\right)^{0.46}\left(\frac{W}{H}\right)^{0.3}S_n^{0.73}$	
Baek and Seo [18]	$\frac{D_T}{HU^*} = (77.88P)^2\left\{1 - exp\left(-\frac{1}{77.88P}\right)\right\}$, $P = \frac{U}{U^*}\frac{H}{R_c}$	
Gond et al. [12]	$\frac{D_T}{HU^*} = f(\lambda) + (2.6\kappa^3)\left(\frac{U}{U^*}\right)\left(\frac{W}{H}\right)$, $f(\lambda) = 0.13$ ($\lambda = 8\left(\frac{U}{U^*}\right) > 0.08$), κ: flow nonuniformity parameter	
Aghababaei et al. [19]	$\frac{D_T}{HU^*} = 0.463 + \left(0.464^{U/U^*}\right) + \left[8.824 \times 10^{-9}(S_n)^{U/U^*}\right] + 0.149S_n^{\left(\frac{U}{U^*}+2.306(Fr)(S_n^2)-25.283\right)} - 0.474S_n^{[0.054\frac{W}{H}-20.371]}$	Genetic-programming-based symbolic regression (GP-SR)
Huai et al. [30]	$\frac{D_T}{HU^*} = \frac{0.693}{262+\left(\frac{U}{U^*}\right)^2-31.8\left(\frac{U}{U^*}\right)} + \frac{0.121\left(\frac{W}{H}\right)}{\frac{W}{H}+0.222\left(\frac{U}{U^*}\right)-1.99}$ (straight flume); $\frac{D_T}{HU^*} = \frac{0.693\left(\frac{U}{U^*}\right)^{0.47}}{262+\left(\frac{U}{U^*}\right)^2-31.8\left(\frac{U}{U^*}\right)} + \frac{0.121\left(\frac{W}{H}\right)^{1.07}\left(\frac{U}{U^*}\right)^{0.35}S_n{}^{0.395}}{\frac{W}{H}+0.222\left(\frac{U}{U^*}\right)-1.99}$ (natural streams)	Genetic programming (GP)

Figure 2 depicts the research procedure for obtaining D_T from 216 datasets, as listed in Table 1. The datasets were classified across three ranges based on W/H: $W/H < 50$ (Class 0), $50 \leq W/H < 100$ (Class 1), and $100 \leq W/H$ (Class 2), by the river scale, as proposed by Baek and Seo [71]. The original datasets consisted of 180 datasets in Class 0 (majority class), and 21 and 14 datasets in Class 1 and 2 (minority class), respectively. To generate new data, the training and validation datasets were split into 80% (172 datasets) and 20% (44 datasets), respectively, according to suggestions from previous studies [34,75,76]. Utilizing oversampling techniques, new datasets comprising W/H, U/U^*, and D_T/HU^* were resampled from the training datasets classified as the minority class. After data oversampling, the dataset was divided into 70% training and 30% validation sets. D_T was estimated using both the traditional MLR method and nonlinear regression methods, specifically SVR, XGBoost, and KNR. The Python Scikit-learn library (https://scikit-learn.org, accessed on 1 May 2024) [77] was utilized for conducting the aforementioned data regression in this study. From the MLR analysis, an empirical formula for D_T was derived:

$$\ln\left(\frac{D_T}{HU^*}\right) = \ln(a) + b\ln\left(\frac{W}{H}\right) + c\ln\left(\frac{U}{U^*}\right) \quad (3)$$

where a, b, and c are empirical coefficients. The derived empirical formula and the three nonlinear regression models were evaluated by comparing them with 30% of test datasets extracted from the original datasets. The comparison results addressed the feasibility of using oversampling techniques to estimate D_T in comparison to results obtained using the original datasets.

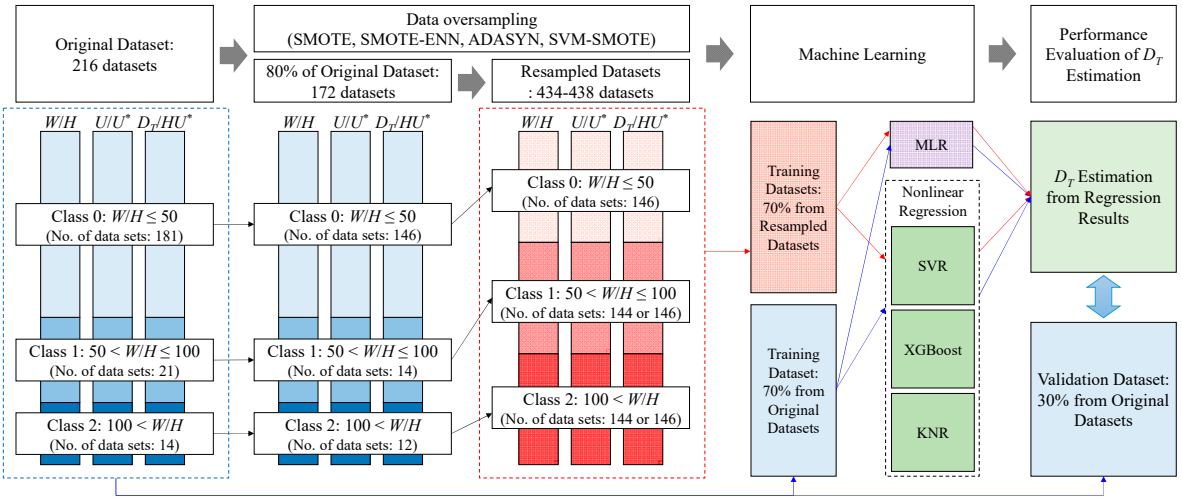

Figure 2. Research outlines to estimate transverse dispersion coefficient.

2.3. Data Oversampling

From the collected tracer test results, it is evident that there is an imbalance in the data concerning W/H, and this imbalance may lead to errors in the empirical formula for D_T. In this study, we aim to address this data imbalance by employing oversampling techniques. The oversampling techniques chosen for this study are summarized in Table 3, which includes data resampling properties, advancements, and limitations of each oversampling technique.

Table 3. Comparisons of oversampling techniques.

Technique	Data Resampling	Pros	Cons	Reference
SMOTE	Generates synthetic samples near minority instances	Mitigates class imbalance	Sensitive to noisy data	Chawla et al. [34]
SMOTE-ENN	Applies Edited Nearest Neighbor (ENN) for noise reduction	Effective in handling noisy data	Possible to discard informative instances during undersampling	Batista et al. [42]
ADASYN	Utilizes density distribution for minority class data synthesis	Adapts to data density variations	Possible to introduce noise due to adaptability	He et al. [41]
SVM-SMOTE	Integrates with support vector machine (SVM) for minority data synthesis	Generates samples in the feature space of minority class	Computationally expensive and sensitive to SVM parameters	Nguyen et al. [78]

SMOTE [34] generates synthetic data to increase the number of minority group instances, aiming to balance the overall dataset. SMOTE achieves this by resampling data from the k-nearest neighbors (KNN) within the minority group. The correct formula for generating synthetic data in SMOTE is as follows:

$$s_i = x_i + (x_{ni} - x_i) \cdot \lambda \quad (4)$$

where i is the sample number of a minority group, s_i is the new synthetic data, x_i is a sample from the minority group, x_{ni} is a randomly selected data from the k-nearest neighbors

within the minority group, and λ is a random number in the range of 0 and 1. The new data are created by interpolating among the minority group data, ensuring that the generated samples lie within the boundaries of the minority group.

The SMOTE-ENN algorithm [40] represents a hybrid approach, integrating SMOTE with ENN, an undersampling technique introduced by Wilson [79]. This method starts by generating synthetic data through SMOTE and subsequently employs ENN to eliminate instances identified as noisy and irrelevant. In ENN, synthetic data are classified as noisy if their class differs from the majority class among their k-nearest neighbors, with k set to 3. The incorporation of the ENN algorithm enhances the quality of the synthesized data group by effectively mitigating the introduction of misleading information or noise during the data synthesis process, as facilitated by SMOTE.

ADASYN [41] employs Equation (5) to generate new samples, and the number of resampled data (N_i) is determined from the density distribution ($\hat{r}_i$). N_i is calculated as:

$$N_i = \hat{r}_i \cdot (n_{mj} - n_{mn})\lambda \tag{5}$$

$$\hat{r}_i = r_i / \sum_{i=1}^{n_{mn}} r_i = r_i / \sum_{i=1}^{n_{mn}} (\Delta_i / k) \tag{6}$$

where n_{mj} and n_{mn} represent the number of majority and minority group data, respectively, $r_i = \Delta_i / k$, k is the number of the nearest neighbors, and Δ_i is the number of majority group data in the k-nearest neighbors of x_i. From the calculations of N_i and $\hat{r}_i$, ADASYN algorithm synthesizes data, accounting for the difficulties in learning levels by assigning weights to minority group data.

SVM-SMOTE [78] is a variation of SMOTE, integrated with the support vector machine (SVM). The primary objective is to generate synthetic samples specifically in the feature space of the minority class. This approach aims to enhance the representation of the minority class through a combination of SVM principles and SMOTE. SVM-SMOTE generates new synthetic data as:

$$s_i = sv_i + (sv_i - x_i) \cdot \lambda \tag{7}$$

where sv_i is the support vector by training SVM on x_i. SVM-SMOTE prioritizes the augmentation of minority class instances near the decision boundaries, which are critical areas for boundary establishment. Furthermore, the generation of new instances strategically expands the minority class domain, particularly in regions with sparse majority class representation.

The four oversampling techniques were employed using the imbalanced-learn library from Python (https://www.jmlr.org/papers/v18/16-365.html, accessed on 1 May 2024) [80]. Accuracy, precision, recall, and F1 score were used to evaluate the classification performance of the oversampling techniques. These indices are calculated to validate the resampled data, as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + TN + FP + FN} \tag{10}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision \times Recall} \tag{11}$$

where TP (true positive) represents the number of samples accurately predicted as positive, TN (true negative) indicates the number of samples accurately predicted as negative, FP (false positive) is the count of samples falsely predicted as positive, and FN (false negative)

denotes the number of samples falsely predicted as negative. In addition, the statistical similarity of the oversampled data to the original data was assessed using the Kolmogorov–Smirnov test (KS test) [81], which compared the cumulative distribution functions of the two datasets.

2.4. Machine Learning Regression Methods

2.4.1. Support Vector Machine Regression (SVR) Model

SVR is an extension of the support vector machine (SVM) algorithm, which is primarily used for data classification. SVM aims to determine a hyperplane that maximizes the margin around the given dataset, ensuring that each data point lies within the margin boundary [82]. SVR addresses regression problems by mapping nonlinear data to a higher-dimensional space using kernel functions, transforming low-dimensional nonlinear regression problems into high-dimensional linear regression problems [72]. Consequently, SVR solves the regression problem by maximizing the margin from the given dataset $(x_i,\ y_i)$ through the following optimization problem:

$$\begin{aligned} &\text{minimize } \tfrac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\left(\xi_i + \xi_i^*\right) \\ &\text{subject to } \begin{cases} y_i - \left(w^T x_i + b\right) \le \varepsilon + \xi_i^* \\ \left(w^T x_i + b\right) - y_i \le \varepsilon + \xi_i \\ \xi_i,\ \xi_i^* \ge 0,\ i = 1, ..l \end{cases} \end{aligned} \tag{12}$$

where w is the weighting vector, C is a positive constant, ξ_i and ξ_i^* are slack variables used to estimate the deviation between actual data and a predicted data, b is a bias term, and ε is the margin. The kernel function employed for this study is the radial basis function (RBF), as follows:

$$k\left(x_i, x_j\right) = \exp\left(-\gamma\|x_i - x_j\|^2\right) \tag{13}$$

where x_i and x_j are data points and γ is a parameter for the RBF kernel function.

2.4.2. eXtream Gradient Boosting Regression (XGBoost) Model

XGBoost is a machine learning regression model that applies the gradient boosting algorithm, known for its advantages in parallel processing and optimization in solving both classification and regression problems [83]. XGBoost is an ensemble technique that combines multiple decision trees to create an ensemble model for nonlinear regression. A decision tree is a method that classifies data by stacking multiple binary nodes with various conditions to predict the final value. For instance, in a scenario with three decision tree models, M_1, M_2, M_3, boosting adjusts the weights of poorly predicted samples x_i in M_1 to train M_2, and similarly adjusts weights for poorly predicted samples x_i in M_2 to train M_3, and so on. The final prediction is obtained by combining predictions from each model with their respective weights, W_n, as shown in Equation (14).

$$y_i = \sum_{i=1}^{K} W_i M_i(x_i) + \varepsilon_j(x_j) \tag{14}$$

where K is the number of decision trees. This boosting technique, implemented in XGBoost, differs from traditional gradient boosting as it incorporates weight assignments for regularization, which helps reduce overfitting. Furthermore, XGBoost allows users to define optimization goals and evaluation criteria, and it includes built-in routines to handle missing values, enabling various learning experiments [83].

2.4.3. k–Nearest Neighbors Regression (KNR) Model

KNR algorithm is a method used in machine learning regression models to predict results for new inputs by utilizing information from the k-nearest data points [84]. KNR offers a flexible approach by considering the local structure of the data. The algorithm

does not assume any specific functional form for the relationship between the predictors and the response variable, making it suitable for capturing complex nonlinear patterns in the data. When estimating the desired value, the algorithm calculates the distance from each of the *k*-nearest data points in the given dataset. For this purpose, Euclidean distance is employed to measure distances between the training data points. The Euclidean distance (d) between two points, $X(x_1,\ x_1,\ \cdots,\ x_n)$ and $Y(y_1,\ y_1,\ \cdots,\ y_n)$, is defined by the following equation:

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \tag{15}$$

In a regression model that outputs numerical values, the output is the mean value of the *k*-nearest neighbors, where weights inversely proportional to the distances of the neighboring points are applied and averaged. Given the *k*-nearest neighbor, when the input x is provided, the output, y, is computed using both the mean value, Equation (16), and the weighted mean value, Equation (17).

$$y = \frac{1}{k}\sum_{i=1}^{k} y_i \tag{16}$$

$$y = \frac{\sum_{i=1}^{k} w_i y_i}{\left(\sum_{i=1}^{k} w_i\right) w_i} = \frac{1}{d(X,\ X_i)} \tag{17}$$

In this study, to determine the nearest neighbor count K for the data samples, the value of K that minimizes the root mean square error (RMSE) was selected from the range of 1 to 20.

3. Results

3.1. Oversampling Results and Performance Evaluations

The transverse dispersion coefficients and accompanying hydraulic data were resampled using four oversampling techniques, SMOTE, SMOTE-ENN, ADASYN, and SVM-SMOTE. The data of W/H, U/U^*, and D_T/HU^* included in minority classes (classes 1 and 2 depicted in Figure 2) were resampled and plotted with original datasets in Figure 3. The number of resampled data increased from 216 to 438 using SMOTE and SVM-SMOTE and rose to 436 and 434 using SMOTE-ENN and ADASYN, respectively. Since oversampling is based on the classification according to the range of W/H, the classes of the resampled data are clearly distinguished in Figure 3a. However, the classes among the resampled data based on U/U^* are not as clearly distinguished (Figure 3b), and therefore are represented based on the classification according to W/H. SMOTE-ENN, being rooted in SMOTE, exhibited a similar distribution in the resampled data and generated data points between the original data using the KNN technique. In contrast, ADASYN adapts its sampling density according to the local distribution of minority class samples, resulting in increased sampling around the borderline instances, as depicted in the relatively higher density near the boundary of the Classes 1 and 2. SVM-SMOTE, leveraging the SVM algorithm, generates synthetic samples focusing on regions that are difficult to classify, thereby reinforcing the characteristics of the minority class by creating data points centered on specific instances.

The resampled datasets were evaluated based on two criteria: whether the newly generated dataset was accurately classified according to the original dataset's class distribution, and whether it exhibited statistically similar characteristics to the original dataset. For the assessments, the calculation results using the classification performance indicators (Equations (8)–(11)) and p-values from the KS test were included in Table 4 for comparing the performance of the resampled datasets. Both the classification performance indicators and the KS test results indicate that all tested oversampling techniques provide acceptable results. Specifically, the results obtained by SMOTE outperformed the other oversampling techniques, followed by SMOTE-ENN, SVM-SMOTE, and ADASYN. However, in line with the purpose of this study, it is required to test whether the resampled data are applicable to estimate D_T/HU^* beyond the statistical reproducibility of the datasets.

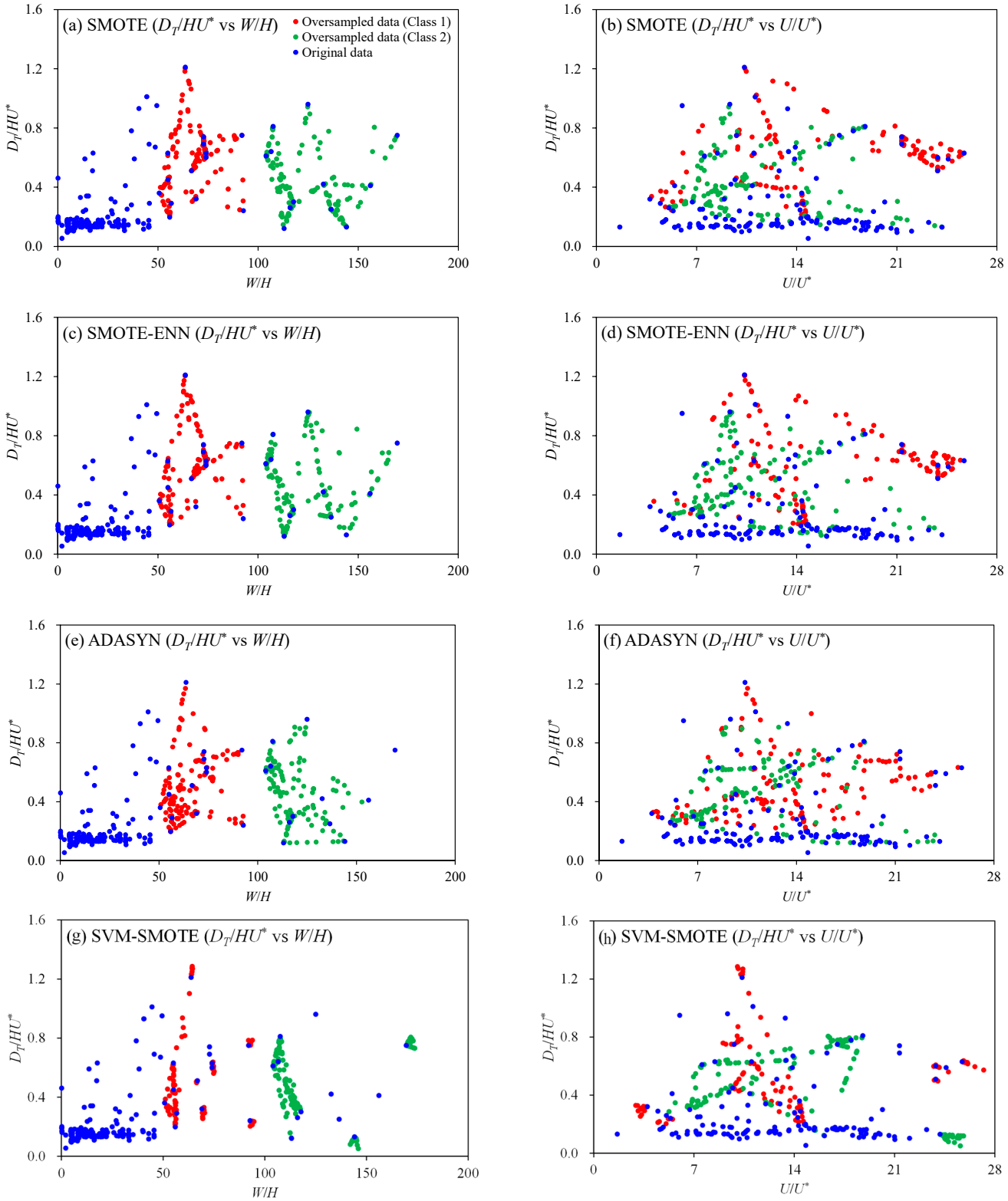

Figure 3. Resampling results using oversampling techniques.

Table 4. Performance evaluations of the oversampled samples.

Oversampling	Classification Performance Indicators					Kolmogorov–Smirnov Test: *p*-Value			
	Accuracy (Equation (8))	Precision (Equation (9))	Recall (Equation (10))	F1 (Equation (11))	AUC *	W/H	U/U^*	D_T/HU^*	Average
SMOTE	0.826	0.937	0.884	0.910	0.983	0.992	0.988	0.979	0.986
SMOTE-ENN	0.820	0.939	0.874	0.905	0.983	0.988	0.960	0.994	0.981
ADASYN	0.749	0.931	0.806	0.864	0.971	0.889	0.595	0.783	0.756
SVM-SMOTE	0.763	0.937	0.815	0.872	0.969	0.846	0.833	0.954	0.878

Note: * AUC = Area under the receiver operating characteristic (ROC) curve: this metric evaluates the performance of an oversampling model.

3.2. D_T Predictions Using MLR

From both the original and resampled datasets, empirical formulas were derived using the conventional method, multiple linear regression (MLR). The training dataset for obtaining empirical coefficients of Equation (3) using MLR was 70% of each oversampled dataset. The derived empirical coefficients are listed in Table 5. The results obtained using the original dataset indicate a larger value of b compared to c, suggesting that the effects of W/H are more dominant than those of U/U^* in determining the transverse dispersion coefficient. Except for the results in SMOTE, larger weighting in W/H appeared even though there are differences in degree. The results by SMOTE suggested more weighting in U/U^*.

Table 5. Empirical coefficients obtained from the multiple linear regression.

Data	Coefficients		
	a	***b***	***c***
Original	0.0443	0.4430	0.1228
SMOTE	0.0323	0.3648	0.4055
SMOTE-ENN	0.0408	0.3652	0.3118
ADASYN	0.0352	0.4437	0.2348
SVM-SMOTE	0.0558	0.4021	0.1273

Note: a, b, and c are empirical coefficients for an empirical formula, $D_T/HU^* = a\left(\frac{W}{H}\right)^b \left(\frac{U}{U^*}\right)^c$.

D_T/HU^* was estimated using derived empirical formulas and compared with measurements. This comparison was conducted using a validation dataset comprising 30% of the original dataset. Figure 4 shows the comparison results of D_T/HU^*, plotted alongside computation results using empirical formulas presented in Table 2. The accuracy of the estimated results was evaluated using the mean absolute percentage error (MAPE), as follows:

$$\text{MAPE} = \sum_{i=1}^{n} \frac{|O_i - P_i|}{O_i} \tag{18}$$

where O_i is the measurements, P_i is the estimated value, and n is the number of validation datasets. The calculation results of MAPE are presented in Table 6. MAPE was computed for both the entire validation set and for separated sets based on the range of W/H. For the entire validation set, MAPE calculation results from the oversampled dataset-derived empirical formulas demonstrated lower accuracy compared to the results using the original dataset. This lower accuracy in the oversampling results is attributed to the majority class dataset ($W/H \leq 50$), which resulted in large errors. Conversely, for the minority class dataset ($W/H > 50$), which incorporates the resampled data, MAPE calculations resulted in higher accuracy compared to the results obtained using the original dataset across all oversampling techniques. These results indicate that data oversampling improves the estimation accuracy, especially for the minority class ($W/H > 50$). However, the empirical formulas were strongly influenced by the resampled data, particularly as the number of resampled data points in the majority class ($W/H \leq 50$) was approximately doubled, resulting in decreased performance in the estimation of D_T/HU^* for the majority class.

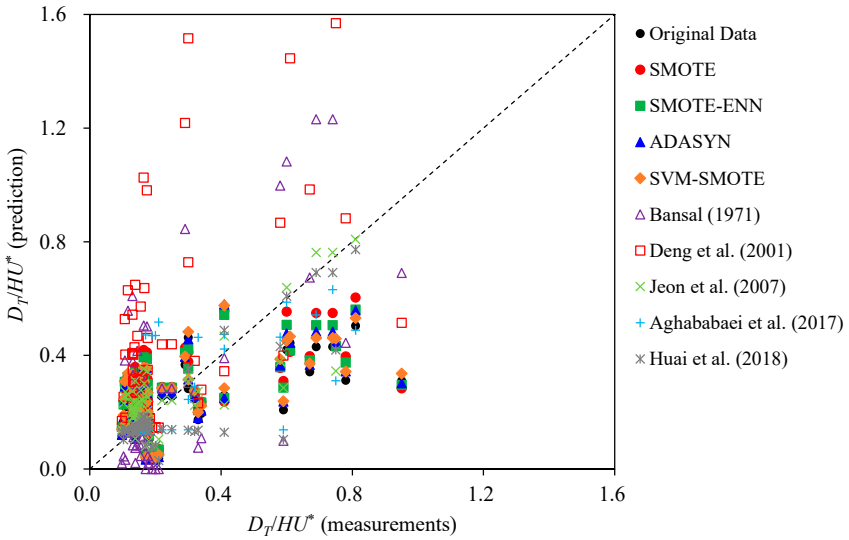

Figure 4. Comparisons of the prediction results of D_T/HU^* using empirical formulas.

Table 6. Comparisons of prediction errors resulted from empirical formulas.

	This Study					Previous Studies				
	Original Data	**SMOTE**	**SMOTE-ENN**	**ADASYN**	**SVM-SMOTE**	**Bansal [50]**	**Deng et al. [17]**	**Jeon et al. [10]**	**Aghababaei et al. [19]**	**Huai et al. [30]**
MAPE (%)	53.4	67.3	65.7	57.2	63.0	108.8	155.4	51.2	27.0	15.0
MAPE (%) ($W/H \leq 50$)	56.4	73.8	71.6	61.2	67.9	80.8	131.7	55.5	27.7	15.0
MAPE (%) ($W/H > 50$)	37.1	31.2	33.0	35.0	36.4	262.5	285.5	23.9	22.2	14.9

The MAPE calculation results from the formulas listed in Table 2 reveal that the results by Huai et al. [30] exhibited the highest accuracy, and among the formulas derived through MLR, the results by Jeon et al. [10] showed the highest accuracy. Despite using data from natural streams, Jeon et al. [10] achieved higher accuracy than the results presented in this study by utilizing three variables (W/H, U/U^*, and S_n) for the entire W/H ranges. Consequently, even with data imbalance alleviated through oversampling, there may be limitations in improving accuracy when using MLR to obtain estimates of D_T/HU^*. These results suggest that, in developing empirical formulas using MLR, the application of more variables may be more advantageous than increasing the number of data points. However, since empirical formulas derived from MLR are based on the linearity between independent and dependent variables, there are limitations to improving accuracy. Therefore, empirical formulas developed through GP suggested by Agababaei et al., and Huai et al. [19,30] outperformed compared to those obtained by Jeon et al. [10] using MLR. However, determining whether the results of Agababaei et al., and Huai et al. [19,30] were due to improved accuracy through nonlinear regression analysis or an increase in the number of data is challenging, as they utilized more data and three or more variables (W/H, U/U^*, S_n, and Fr) for empirical formula derivation compared to the study by Jeon et al. [10]. This suggests that there is a need for applying nonlinear regression analysis for transverse dispersion coefficient estimation and implies the necessity of verifying whether applying nonlinear regression analysis can overcome the limitations of using a limited number of regression variables.

3.3. D_T Predictions Using Nonlinear Regression Methods

D_T/HU^* was estimated using nonlinear regression methods, including SVR, XGBoost, and KNR, applied to both the original and resampled datasets. During the data learning phase, 70% of each dataset was utilized to train the nonlinear regression models. Subse-

quently, the trained regression models were applied to estimate D_T/HU^* for the validation dataset, which corresponded to the dataset depicted in Figure 4. The estimation results is presented in Figure 5, where the results by Aghababei et al., and Huai et al. [19,30] are also plotted for comparison. The results obtained using nonlinear regression models, as depicted in Figure 5, exhibit significant improvement compared to those in Figure 5, with a larger proportion of estimations closely aligned along the diagonal line. Computation results of MAPE, provided in Table 7, underscore this enhancement. Notably, the original data yielded a MAPE of 43.6%, representing an improvement over MLR (MAPE = 53.4%). Particularly noticeable improvements were observed in the results derived from the resampled datasets, with averaged MAPE ranging from 20.9% to 25.5% across different oversampling techniques. These values significantly outperformed those shown in Table 4 (MAPE = 57.2% to 67.3%).

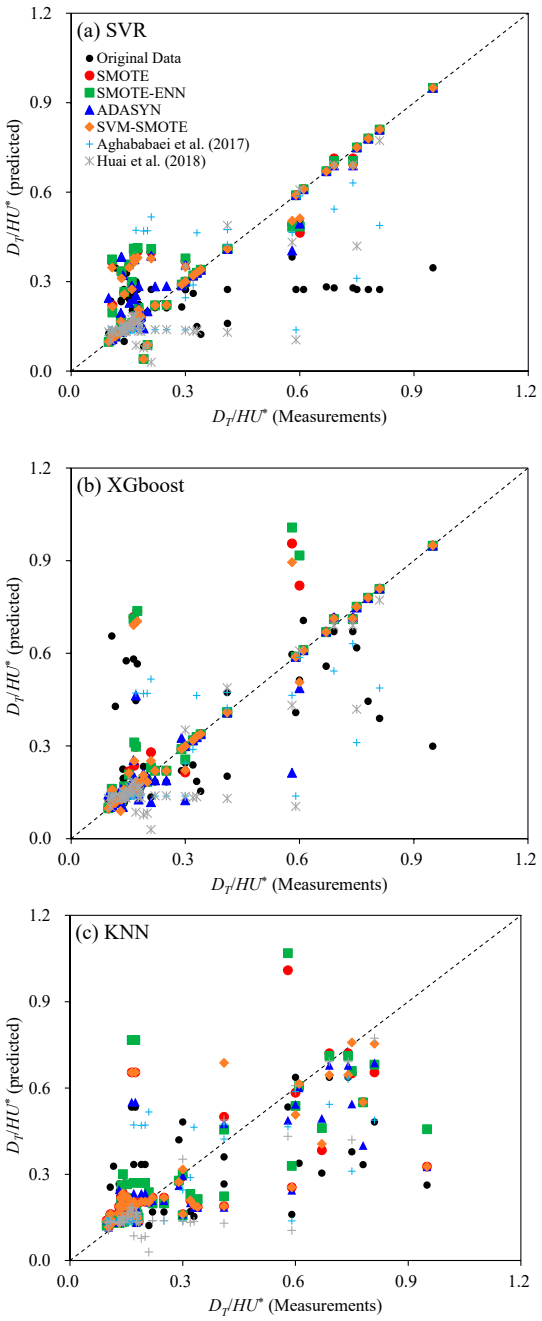

Figure 5. Comparisons of the prediction results of D_T/HU^* using nonlinear regression methods. The dashed line indicates the best-fit to the measurements.

Table 7. Comparisons of prediction errors resulted from nonlinear regression methods.

Data	Data Range	MAPE (%)			Average	Rank
		SVR	XGBoost	KNR		
Original Data	Total	44.1	44.3	42.2	43.6	5
	$W/H \leq 50$	44.4	49.2	44.2	46.0	
	$50 < W/H$	42.5	17.3	31.3	30.4	
SMOTE	Total	24.3	18.0	31.6	24.6	3
	$W/H \leq 50$	27.9	19.3	32.7	26.7	
	$50 < W/H$	4.5	10.9	25.0	13.5	
SMOTE-ENN	Total	24.8	18.2	33.4	25.5	4
	$W/H \leq 50$	28.5	19.1	34.4	27.3	
	$50 < W/H$	4.3	13.4	28.0	15.2	
ADASYN	Total	21.5	10.9	30.2	20.9	1
	$W/H \leq 50$	24.4	11.0	31.6	22.4	
	$50 <W/H$	5.6	10.2	22.4	12.7	
SVM-SMOTE	Total	22.5	16.5	32.9	23.9	2
	$W/H \leq 50$	25.9	18.1	34.1	26.1	
	$50 < W/H$	3.5	7.7	25.9	12.4	

The results obtained using the resampled data demonstrate comparable accuracy to those obtained by Aghababaei et al., and Huai et al. [19,30]. An encouraging aspect is that the results presented in this study improved accuracy using only two variables (W/H and U/U^*), while Aghababaei et al., and Huai et al. [19,30] utilized four and three variables, respectively. For the resampled dataset range ($W/H > 50$), the combination of SVM-SMOTE and SVR yielded the most accurate results. However, it is worth noting that the results of SVM-SMOTE were deemed to be overfitted for the data belonging to $W/H > 50$, leading to a decrease in accuracy in the range of $W/H \leq 50$. Thus, the best combination of an oversampling technique and a nonlinear regression model was found to be ADASYN and XGboost. Specifically, the datasets resampled by ADASYN provided the best results in every case using the three nonlinear regression methods. These results suggest that the combination of ADASYN and the nonlinear regression methods offers comparatively improved results, especially in the non-oversampled data range ($W/H \leq 50$).

4. Discussion

4.1. D_T Estimation Performance Using MLR through Data Augmentation

The efficacy of empirical formulas in estimating D_T relies on both accuracy and expansibility. The acquisition of data plays a crucial role in enhancing the performance of D_T estimation through empirical formulas derived by MLR. However, the limited availability of datasets for $W/H > 50$ poses constraints on improving the performance of empirical formulas. Among the 216 datasets collected in this study, 181 correspond to $W/H \leq 50$, while 35 correspond to $W/H > 50$. Therefore, there is a risk of deriving overfitted equations for $W/H > 50$ datasets when estimating D_T using datasets biased towards $W/H \leq 50$. To improve the accuracy of the estimation, it is necessary to either increase the number of variables or acquire new datasets including $W/H > 50$. However, there are limitations due to the increase in the complexity of the estimation and the need to obtain results from field tracer tests.

To address the limitation of insufficient measured data, this study employed oversampling techniques to generate new datasets within the $W/H > 50$ range, ensuring that the generated data reflects the statistical properties of the original data, as validated by the KS test (Table 4). The empirical formulas derived with W/H and U/U^* from the oversampled data demonstrated improved accuracy within the $W/H > 50$ range compared to the pre-oversampling results (Table 5). Nevertheless, when considering the accuracy for the $W/H \leq 50$ range, the formula utilizing the original dataset exhibited higher accuracy than

those using oversampled data. These findings underscore the idea that data point increase is insufficient for enhancing the performance of an empirical formula derived through MLR. Therefore, as proposed by Baek and Seo [71], it is imperative to apply distinct empirical formulas based on W/H. Alternatively, increasing the complexity of empirical formulas by incorporating additional variables or employing nonlinear regression methods would offer a viable solution.

The range of reproducible D_T/HU^* through empirical formulas including two variables (W/H and U/U^*) demonstrates the extendibility of the developed formula. Figure 6 shows the range of D_T/HU^* derived from empirical formulas developed using original and oversampled data by ADASYN. To depict the possible range of D_T/HU^* concerning the variation in W/H, the upper and lower boundaries were determined by applying the maximum and minimum values of U/U^* from the original dataset (Figure 6a). Similarly, the calculable range concerning the variation in U/U^* was determined by applying the maximum and minimum values of W/H (Figure 6b). These results indicate that the range of D_T/HU^* widens when utilizing empirical formulas derived from an oversampled dataset. This suggests that empirical formulas derived from resampled datasets may possess higher applicability than those utilizing only original data. However, compared to the reproducible range proposed by Jeon et al. [10] using three variables (W/H, U/U^*, and S_n), these findings reveal a considerably limited reproducible range. Additionally, as indicated in Table 6, despite using only 32 field datasets, the results by Jeon et al. [10] exhibited the highest accuracy among those derived using MLR. Hence, these results indicate that adding variables that account for secondary currents' effects on transverse dispersion is more effective in improving D_T estimation performance than increasing the number of datasets. Consequently, data oversampling may offer limited enhancements to D_T estimation performance when employing MLR.

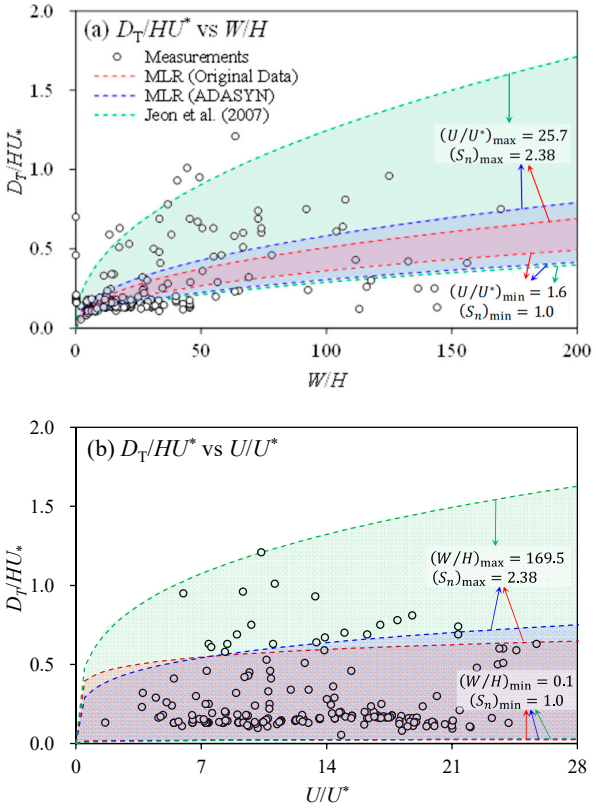

Figure 6. Reproducible range of empirical formulas for estimating the transverse dispersion coefficient: red, blue, and green areas represent the available ranges using empirical formulas derived from original data, resampled data by ADASYN, and Jeon et al. [10], respectively.

4.2. Comparisons of D_T Estimation Results Using MLR and Nonlinear Regression Methods

In this study, original tracer test datasets (216 datasets) and resampled datasets (434–438 datasets) were used for estimating D_T (see Figure 2). Of these datasets, 70% were classified as training datasets to derive empirical formulas for D_T estimation using MLR and three nonlinear regression models: SVR, XGBoost, and KNR. The accuracy of the D_T estimations obtained through each method was compared using MAPE based on a range of W/H values, using a validation dataset comprising 30% of the data (Tables 6 and 7). Results derived from MLR using two variables (W/H and U/U^*) from the original dataset (Table 5) showed errors of 53.5% for $W/H \leq 50$ and 37.1% for $W/H > 50$ (Table 6). These results exhibited similar performance to those derived using formulas developed by Jeon et al. [10] based on three parameters, W/H, U/U^*, and S_n, with errors of 55.5% for $W/H \leq 50$ and 23.9% for $W/H > 50$. Jeon et al. [10] developed D_T estimation formulas using 32 tracer test datasets from natural streams, demonstrating that increasing the number of datasets, instead of incorporating S_n into the estimation formulas, could enhance the performance of MLR-based estimations. However, increasing the number of datasets through oversampling had limitations in improving MLR-based estimations, and resulted in larger errors than results derived solely from original data. These limitations using MLR were reduced by employing machine-learning-based nonlinear regression methods, as demonstrated in the studies by Aghababaei et al., and and Huai et al. [19,30], through the utilization of additional variables (Table 6).

Adding variables for D_T estimation may increase the complexity of the estimation formulas and may have limited applicability. Therefore, instead of adding variables, we investigated the performance-improvement effect through D_T estimation using nonlinear regression methods (SVR, XGBoost, and KNR) (Table 7). The results showed that estimations using W/H and U/U^* from the original dataset had average errors of 46.0% and 30.4% for $W/H \leq 50$ and $W/H > 50$, respectively, indicating improved performance compared to MLR-based estimations. The performance-improvement effect through nonlinear regression methods was further enhanced when using resampled datasets, particularly when estimating D_T using XGBoost from datasets resampled using ADASYN; this showed errors of 11.0% and 10.7% for $W/H \leq 50$ and $W/H > 50$, respectively, outperforming MLR-based D_T estimations. These results demonstrated improved performance compared to the utilization of estimation formulas by Huai et al. [30] based on three variables, W/H, U/U^*, and S_n, which showed errors of 15.0% and 14.9% for $W/H \leq 50$ and $W/H > 50$, respectively. In conclusion, considering these results, machine-learning-based nonlinear regression methods were found to be more effective than MLR for D_T estimation; additionally, using data oversampling to alleviate dataset imbalance yielded superior performance in D_T estimation compared to the effects of increasing variables.

4.3. The Feasibility of Two Variables for D_T Estimation

The results presented in Table 7 demonstrate that the combination of data oversampling and a nonlinear regression method improves the accuracy of D_T estimation using only W/H and U/U^*. The previous discussion in Section 4.1 discussed the potential of incorporating S_n to extend the predictability range of D_T. While incorporating S_n could potentially improve the accuracy of estimations, its availability to consider the effects of secondary currents on transverse dispersion is considerably restricted for the present datasets. Although W/H and U/U^* were commonly available in all datasets collected in this study, the utilization of hydraulic parameters such as R_c or S_n to reflect flow structures such as secondary currents are highly limited. Furthermore, Gond et al. [12] presented that even in rivers with S_n close to 1, D_T can be significantly increased due to longitudinal flow nonuniformity; however, the lack of sufficient data hinders the utilization of this information in D_T estimation. Hence, until sufficient research data are secured to enhance the regression accuracy of D_T, the application of methodologies to improve D_T estimation accuracy using common variables such as W/H and U/U^*, obtainable across previously published papers, is required.

As demonstrated in Table 7, the combination of the ADASYN–XGBoost method improved D_T estimation accuracy using only W/H and U/U^*. To assess the improvement effect of including S_n in MLR results, an empirical formula including S_n, similar to that proposed by Jeon et al. [10], was derived from the datasets collected in this study, yielding Equation (19):

$$\frac{D_T}{HU^*} = 0.051\left(\frac{U}{U^*}\right)^{0.17}\left(\frac{W}{H}\right)^{0.32} S_n^{1.1} \quad (19)$$

Additionally, utilizing the XGBoost regression combined with the oversampling technique of ADASYN, data resampling was performed on 195 datasets where S_n could be applied out of the 216 original datasets. S_n for the straight channel was set to 1 (152 datasets). Through the resampling, the total dataset increased to 406, with the average S_n value increasing from 1.09 to 1.22 and the median increasing from 1.0 to 1.12. D_T was estimated through XGBoost, using the resampled data including S_n. Figure 7 compares the D_T estimation results obtained through MLR and XGBoost, excluding field-scale data without S_n information among validation datasets shown in Figure 5. Table 8 presents the MAPE calculation results for D_T estimation and the accuracy improvement effect of each D_T estimation result compared to the results from MLR with two variables (W/H and U/U^*). Both MLR and XGBoost yielded higher accuracy when incorporating three variables (W/H, U/U^*, and S_n) compared to utilizing only two variables (W/H and U/U^*) for both the original and oversampled datasets, respectively. Moreover, the XGBoost regression results using oversampled data consisting of only two variables demonstrated higher accuracy than the XGBoost regression results from the original dataset including S_n. Thus, it can be concluded that data oversampling resolves the issue of reduced D_T estimation performance due to variable scarcity, and significant improvements in accuracy comparable to those from increasing variables can be achieved using only W/H and U/U^*.

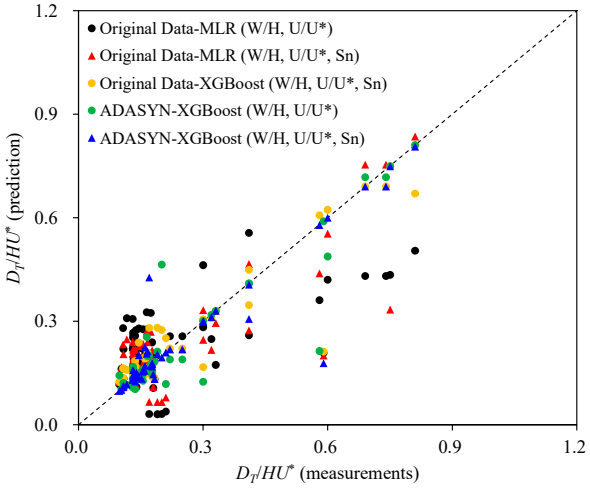

Figure 7. Comparisons of D_T/HU^* estimation results according to selection of the number of variables.

Table 8. Comparisons of prediction errors according to the number of variables.

	Original Data—MLR		Original Data—XGBoost	ADASYN—XGBoost	
	W/H, U/U^*	W/H, U/U^*, S_n	W/H, U/U^*, S_n	W/H, U/U^*	W/H, U/U^*, S_n
MAPE (%)	54.2	38.0	15.7	12.4	9.5
Performance Improvement (%)	-	29.8	71.1	77.1	82.4

5. Conclusions

In this study, we addressed the issue of reduced D_T estimation performance due to the data imbalance, in which the datasets are accumulated in $W/H \leq 50$ from the data resampling employing four oversampling techniques, SMOTE, SMOTE-ENN, ADASYN, and SVM-SMOTE. From the resampled datasets, D_T was estimated using both MLR and three machine learning regression algorithms, SVR, XGBoost, and KNR. The estimated D_T was compared to the empirical formulas proposed by previous research to evaluate performance of D_T estimation by reducing the data imbalance. The results revealed that there was no significant improvement in accuracy with MLR using the oversampled datasets. However, when employing nonlinear regression methods, the effectiveness of accuracy improvement due to data oversampling increased substantially. Notably, when estimating D_T using only two variables, W/H and U/U^*, through the ADASYN–XGBoost method, a higher improvement in performance was observed compared to XGBoost regression results from the original dataset, including S_n. These findings suggest that data oversampling is more effective than increasing the number of variables for employing nonlinear regression. While data oversampling cannot replace field data acquisition, it provides benefits such as improving the accuracy of imbalanced data and enhancing D_T estimation accuracy using minimal variables and restricted tracer test data.

Author Contributions: Conceptualization, S.L. and I.P.; methodology, S.L. and I.P.; software, S.L. and I.P.; validation, S.L. and I.P.; formal analysis, S.L.; investigation, S.L.; resources, S.L. and I.P.; data curation, S.L.; writing—original draft preparation, S.L. and I.P.; writing—review and editing, I.P.; visualization, S.L. and I.P.; supervision, I.P.; project administration, I.P.; funding acquisition, I.P. All authors have read and agreed to the published version of the manuscript.

Funding: This study was financially supported by Seoul National University of Science and Technology.

Data Availability Statement: The data presented in this study will be shared by the authors if requested.

Acknowledgments: This study was conducted by the financial support by Seoul National University of Science and Technology.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Shin, J.; Seo, I.W.; Baek, D. Longitudinal and transverse dispersion coefficients of 2D contaminant transport model for mixing analysis in open channels. *J. Hydrol.* **2020**, *583*, 124302. [CrossRef]
2. Piasecki, M.; Katopodes, N.D. Identification of stream dispersion coefficients by adjoint sensitivity method. *J. Hydraul. Eng.* **1999**, *125*, 714–724. [CrossRef]
3. King, I.; Letter, J.V.; Donnel, B.P. *RMA4 Users Guide 4.5x*; US Army, Engineer Research and Development Center, WES, CHL: Vicksburg, MI, USA, 2008.
4. Lee, M.E.; Seo, I.W. Analysis of pollutant transport in the Han River with tidal current using a 2D finite element model. *J. Hydro-environ. Res.* **2007**, *1*, 30–42. [CrossRef]
5. Park, I.; Seo, I.W.; Shin, J.; Song, C.G. Experimental and numerical investigations of spatially-varying dispersion tensors based on vertical velocity profile and depth-averaged flow field. *Adv. Water Res.* **2020**, *142*, 103606. [CrossRef]
6. Baek, K.O.; Seo, I.W.; Jeong, S.J. Evaluation of dispersion coefficients in meandering channels from transient tracer tests. *J. Hydraul. Eng.* **2006**, *132*, 1003–1119. [CrossRef]
7. Seo, I.W.; Lee, M.E.; Baek, K.O. 2D modeling of heterogeneous dispersion in meandering channels. *J. Hydraul. Res.* **2008**, *44*, 350–362. [CrossRef]
8. Tabatabaei, S.H.; Heidarpour, M.; Ghasemi, M.; Hoseinipour, E.Z. Transverse mixing coefficient on dunes with vegetation on a channel wall. In Proceedings of the World Environmental and Water Resources Congress 2013: Showcasing the Future, Cincinnati, OH, USA, 19–23 May 2013; pp. 1903–1911.
9. Beltaos, S. Transverse mixing tests in natural streams. *J. Hydraul. Div.* **1980**, *106*, 1607–1625. [CrossRef]
10. Jeon, T.M.; Baek, K.O.; Seo, I.W. Development of an empirical equation for the transverse dispersion coefficient in natural streams. *Environ. Fluid Mech.* **2007**, *7*, 317–329. [CrossRef]
11. Seo, I.W.; Choi, H.J.; Kim, Y.D.; Han, E.J. Analysis of two-dimensional mixing in natural streams based on transient tracer tests. *J. Hydraul. Eng.* **2016**, *142*, 04016020. [CrossRef]

12. Gond, L.; Mignot, E.; Le Coz, J.; Kateb, L. Transverse mixing in rivers with longitudinally varied morphology. *Water Resour. Res.* **2020**, *57*, e2020WR029478. [CrossRef]
13. Jung, S.H.; Seo, I.W.; Kim, Y.D.; Park, I. Feasibility of velocity-based method for transverse mixing coefficients in river mixing analysis. *J. Hydraul. Eng.* **2019**, *145*, 04019040. [CrossRef]
14. Fischer, H.B.; List, J.E.; Koh, R.C.Y.; Imberger, J.; Brooks, N.H. *Mixing in Inland and Coastal Waters*, 2nd ed.; Academic Press: San Diego, CA, USA, 1979; pp. 80–147.
15. Rutherford, J.C. *River Mixing*; John Wiley and Sons: London, UK, 1994; pp. 62–63.
16. Gharbi, S.; Verrette, J.L. Relation between longitudinal and transversal mixing coefficients in natural streams. *J. Hydraul. Res.* **1998**, *36*, 43–54. [CrossRef]
17. Deng, Z.Q.; Singh, V.P.; Bengtsson, L. Longitudinal dispersion coefficient in straight rivers. *J. Hydraul. Eng.* **2001**, *127*, 919–927. [CrossRef]
18. Baek, K.O.; Seo, I.W. Empirical equation for transverse dispersion coefficient based on theoretical background in river bends. *Environ. Fluid Mech.* **2013**, *13*, 465–477. [CrossRef]
19. Aghababaei, M.; Etemad-Shahidi, A.; Jabbari, E.; Taghipour, M. Estimation of transverse mixing coefficient in straight and meandering streams. *Water Resour. Manag.* **2017**, *31*, 3809–3827. [CrossRef]
20. Baek, K.O.; Lee, D.Y. Development of simple formula for transverse dispersion coefficient in meandering rivers. *Water* **2023**, *15*, 3120. [CrossRef]
21. Tao, H.; Al-Khafaji, Z.S.; Qi, C.; Yassen, Z.M. Artificial intelligence models for suspended river sediment prediction: State-of-the art, modeling framework appraisal, and proposed future research directions. *Eng. Appl. Comput. Fluid Mech.* **2021**, *15*, 1585–1612. [CrossRef]
22. Tayfur, G.; Singh, V.P. Predicting longitudinal dispersion coefficient in natural streams by artificial neural network. *J. Hydraul. Eng.* **2005**, *131*, 991–1000. [CrossRef]
23. Noori, R.; Karbassi, A.; Farokhnia, A.; Dehghani, M. Predicting the longitudinal dispersion coefficient using support vector machine and adaptive neuro-fuzzy inference system techniques. *Environ. Eng. Sci.* **2009**, *26*, 1503–1510. [CrossRef]
24. Sattar, A.M.A.; Gharabaghi, B. Gene expression models for prediction of longitudinal dispersion coefficient in streams. *J. Hydrol.* **2015**, *524*, 587–596. [CrossRef]
25. Seifi, A.; Riahi-Madvar, H. Improving one-dimensional pollution dispersion modeling in rivers using ANFIS and ANN-based GA optimized models. *Environ. Sci. Pollut. Res.* **2019**, *26*, 867–885. [CrossRef]
26. Azar, N.A.; Milan, S.G.; Kayhomayoon, Z. The prediction of longitudinal dispersion coefficient in natural streams using LS-SVM and ANFIS optimized by Harris hawk optimization algorithm. *J. Contam. Hydrol.* **2021**, *240*, 103781. [CrossRef] [PubMed]
27. Ghiasi, B.; Noori, R.; Sheikhian, H.; Zeynolabedin, A.; Sun, Y.; Jun, C.; Hamouda, M.; Bateni, S.M.; Abolfathi, S. Uncertainty quantification of granular computing-neural network model for prediction of pollutant longitudinal dispersion coefficient in aquatic streams. *Sci. Rep.* **2022**, *12*, 4610. [CrossRef] [PubMed]
28. Ohadi, S.; Monfared, S.A.H.; Moghaddam, M.A.; Givehchi, M. Feasibility of a novel predictive model based on multilayer perceptron optimized with Harris hawk optimization for estimating of the longitudinal dispersion coefficient in rivers. *Neural Comp. Appl.* **2023**, *35*, 7081–7105. [CrossRef]
29. Azamathulla, H.M.; Ahmad, Z. Gene-expression programming for transverse mixing coefficient. *J. Hydrol.* **2012**, *434–435*, 142–148. [CrossRef]
30. Huai, W.; Shi, H.; Yang, Z.; Zeng, Y. Estimating the transverse mixing coefficient in laboratory flumes and natural rivers. *Water Air Soil Pollut.* **2018**, *229*, 252. [CrossRef]
31. Zahiri, J.; Nezaratian, H. Estimation of transverse mixing coefficient in streams using M5, MARS, GA, and PSO approaches. *Environ. Sci. Pollut. Res.* **2020**, *27*, 14553–14566. [CrossRef] [PubMed]
32. Nezaratian, H.; Zahiri, J.; Peykani, M.F.; Haghiabi, A.; Parsaie, A. A genetic algorithm-based support vector machine to estimate the transverse mixing coefficient in streams. *Water Qual. Res. J.* **2021**, *56*, 128. [CrossRef]
33. Najafzadeh, M.; Noori, R.; Afroozi, D.; Ghiasi, B.; Hosseini-Moghari, S.M.; Mirchi, A.; Haghighi, A.T.; Kløve, B. A comprehensive uncertainty analysis of model-estimated longitudinal and lateral dispersion coefficients in open channels. *J. Hydrol.* **2021**, *603*, 126850. [CrossRef]
34. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
35. Huang, R.; Ma, C.; Ma, J.; Huangfu, X.; He, Q. Machine learning in natural and engineered water systems. *Water Res.* **2021**, *205*, 117666. [CrossRef]
36. Xu, T.; Coco, G.; Neale, M. A predictive model of recreational water quality based on adaptive synthetic sampling algorithms and machine learning. *Water Res.* **2020**, *177*, 115788. [CrossRef]
37. Bourel, M.; Segura, A.M.; Crisci, C.; López, G.; Sampognaro, L.; Vidal, V.; Kruk, C.; Piccini, C.; Perera, G. Machine learning methods for imbalanced data set for prediction of faecal contamination in beach waters. *Water Res.* **2021**, *202*, 117450. [CrossRef]
38. Prasad, D.V.V.; Kumar, P.S.; Venkataramana, L.Y.; Prasannamedha, G.; Harshana, S.; Srividya, S.J.; Harrinei, K.; Indraganti, S. Automating water quality analysis using ML and auto ML techniques. *Environ. Res.* **2021**, *202*, 111720. [CrossRef]
39. Snieder, E.; Abogadil, K.; Khan, U.T. Resampling and ensemble techniques for improving ANN-based high-flow forecast accuracy. *Hydrol. Earth Syst. Sci.* **2021**, *25*, 2543–2566. [CrossRef]

40. Nasir, N.; Kansal, A.; Alshaltone, O.; Barneih, F.; Sameer, M.; Shanableh, A.; Al-Shamma'a, A. Water quality classification using machine learning algorithms. *J. Water Proc. Eng.* **2022**, *48*, 102920. [CrossRef]
41. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1 June 2008.
42. Batista, G.E.; Prati, R.C.; Monard, M.C. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explor. Newsl.* **2004**, *6*, 20–29. [CrossRef]
43. Zhou, H.; Dong, X.; Xia, S.; Wang, G. Weighted oversampling algorithms for imbalanced problems and application in prediction of streamflow. *Knowl.-Based Syst.* **2021**, *229*, 107306. [CrossRef]
44. Rahman, M.A.; Akter, A.; Richi, F.S.; Shoud, A.; Ahmed, T. A comparative study of undersampling and oversampling methods for flood forecasting in Bangladesh using machine learning. In Proceedings of the 2023 IEEE 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 6–8 July 2023.
45. Hasan, M.A.; Rouf, N.T.; Hossain, M.S. A location-independent flood prediction model for Bangladesh's rivers. In Proceedings of the 2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI), Atlanta, GA, USA, 6–8 November 2023.
46. Kalinske, A.A.; Pien, C.L. Eddy diffusion. *Ind. Eng. Chem.* **1944**, *36*, 220–223. [CrossRef]
47. Elder, J.W. The dispersion of marked fluid in turbulent shear flow. *J. Fluid Mech.* **1959**, *5*, 544–560. [CrossRef]
48. Sayre, W.W.; Chang, F.M. *A Laboratory Investigation of Open-Channel Dispersion Processes for Dissolved, Suspended, and Floating Dispersants*; Professional Paper, No. 433-E; U.S. Geological Survey: Washington, DC, USA, 1968; pp. 37–71.
49. Sullivan, P.J. Dispersion in a Turbulent Shear Flow. Ph.D. Thesis, University of Cambridge, Cambridge, UK, 1968.
50. Bansal, M.K. Dispersion and Reaeration in Natural Stream. Ph.D. Thesis, Universite de Kansas Laurence, Lawrence, KS, USA, 1970.
51. Okoye, J.K. Characteristics of Transverse Mixing in Open-Channel Flows. Ph.D. Thesis, California Institute of Technology, Pasadena, CA, USA, 1971.
52. Prych, E.A. Effects of Density Differences on Lateral Mixing in Open-Channel Flows. Ph.D. Thesis, California Institute of Technology, Pasadena, CA, USA, 1970.
53. Yotsukura, N.; Fischer, H.B.; Sayre, W.W. *Measurement of Mixing Characteristics of the Missouri River between Sioux City, Iowa, and Plattsmouth, Nebraska*; Water Supply Paper. No. 1899-G; U.S. Geological Survey: Washington, DC, USA, 1970; pp. 11–26.
54. Holly, E.R. *Transverse Mixing in Rivers*; Report No. S132; Delft Hydraulics Laboratory: Delft, The Netherlands, 1971; pp. 34–84.
55. Yotsukura, N.; Cobb, E.D. *Transverse Diffusion of Solutes in Natural Streams*; U.S. Geological Survey: Washington, DC, USA, 1972; pp. 2–19.
56. Fischer, H.B. Longitudinal dispersion and turbulent mixing in open-channel flow. *Annu. Rev. Fluid Mech.* **1973**, *5*, 59–78. [CrossRef]
57. Holley, E.R.; Abraham, G. Laboratory studies on transverse mixing in rivers. *J. Hydraul. Res.* **1973**, *11*, 219–253. [CrossRef]
58. Sayre, W.W.; Yeh, T. *Transverse Mixing Characteristics of the Missouri River Downstream from the Cooper Nuclear Station*; Rep. No.145; Iowa Institute of Hydraulic Research: Iowa City, IA, USA, 1973; pp. 1–46.
59. Engmann, J.E.O. Transverse Mixing Characteristics of Open and Ice-Covered Channel Flows. Ph.D. Thesis, University of Alberta, Edmonton, AB, Canada, 1974.
60. Miller, A.C.; Richardson, E.V. Diffusion and dispersion in open channel flow. *J. Hydraul. Div.* **1974**, *100*, 159–171. [CrossRef]
61. Lau, Y.L.; Krishnappan, B.G. Transverse dispersion in rectangular channels. *J. Hydraul. Div.* **1977**, *103*, 1173–1189. [CrossRef]
62. Beltaos, S.; Day, T.J. A field study of longitudinal dispersion. *Can. J. Civ. Eng.* **1978**, *5*, 572–585. [CrossRef]
63. Sayre, W.W.; Caro-Cordero, R. *Shore-Attached Thermal Plumes in Rivers. Modelling in Rivers*; Wiley-Interscience: London, UK, 1979; pp. 15.1–15.44.
64. Lau, Y.L.; Krishnappan, B.G. Modelling transverse mixing in natural streams. *J. Hydraul. Div.* **1981**, *107*, 209–226. [CrossRef]
65. Holly, F.M.; Nerat, G. Field calibration of stream-tube dispersion model. *J. Hydraul. Eng.* **1983**, *109*, 1455–1470. [CrossRef]
66. Webel, G.; Schatzmann, M. Transverse mixing in open channel flow. *J. Hydraul. Eng.* **1984**, *110*, 423–435. [CrossRef]
67. Long, T.; Guo, J.; Feng, Y.; Huo, G. Modulus of transverse diffuse simulation based on artificial neural network. *Chongqing Environ. Sci.* **2002**, *24*, 25–28. (In Chinese)
68. Seo, I.W.; Baek, K.O.; Jeon, T.M. Analysis of transverse mixing in natural streams under slug tests. *J. Hydraul. Res.* **2006**, *44*, 350–362. [CrossRef]
69. Fischer, H.B. The effect of bends on dispersion in streams. *Water Resour. Res.* **1969**, *5*, 496–506. [CrossRef]
70. Yotsukura, N.; Sayre, W.W. Transverse mixing in natural channels. *Water Resour. Res.* **1976**, *12*, 695–704. [CrossRef]
71. Baek, K.O.; Seo, I.W. Estimation of transverse dispersion coefficient for two-dimensional mixing in natural streams. *J. Hydro-environ. Res.* **2017**, *15*, 67–74. [CrossRef]
72. Smola, A.J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222. [CrossRef]
73. Zhou, W.; Yan, Z.; Zhang, L. A comparative study of 11 non-linear regression models highlighting autoencoder, DBN, and SVR, enhanced by SHAP importance analysis in soybean branching prediction. *Sci. Rep.* **2024**, *14*, 5905. [CrossRef] [PubMed]
74. Taunk, K.; De, S.; Verma, S.; Swetapadma, A. A brief review of nearest neighbor algorithm for learning and classification. In Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2019), Madurai, India, 15–17 May 2019.

75. Jeatrakul, P.; Wong, K.; Fung, C. Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm. In Proceedings of the Neural Information Processing Models and Applications: 17th International Conference, ICONIP 2010, Sydney, Australia, 22–25 November 2010.
76. Rastogi, A.K.; Narang, N.; Siddiqui, Z.A. Imbalanced big bata classification: A distributed implementation of SMOTE. In Proceedings of the Workshop Program of the 19th International Conference on Distributed Computing and Networking, ACM, 14, Varanasi, India, 4–7 January 2018.
77. Pedregosa, F.; Grise, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
78. Nguyen, H.M.; Cooper, E.W.; Kamei, K. Borderline oversampling for imbalanced data classification. *Int. J. Knowl. Eng. Soft Data Paradig.* **2011**, *3*, 4–21. [CrossRef]
79. Winson, D.L. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Trans. Syst. Man Cybern.* **1972**, *SMC-2*, 408–421.
80. Lemaitre, G.; Nogueira, F.; Aridas, C.K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *J. Mach. Learn. Res.* **2017**, *18*, 1–5.
81. Hodges, J.L. The significance probability of the Smirnov two-sample test. *Ark. Mat.* **1958**, *3*, 469–486. [CrossRef]
82. Drucker, H.; Burges, C.J.; Kaufman, L.; Smola, A.; Vapnik, V. Support vector regression machines. *Adv. Neural Inform. Process. Syst.* **1996**, *9*, 155–161.
83. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016.
84. Altman, N.S. An introduction to kernel and nearest neighbor nonparametric regression. *Am. Stat.* **1992**, *46*, 175–185.