


Article

Cultivating Ensemble Diversity through Targeted Injection of Synthetic Data: Path Loss Prediction Examples

Sotirios P. Sotiroudis 

ELEDIA@AUTH, School of Physics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece; ssoti@physics.auth.gr

Abstract: Machine Learning (ML)-based models are steadily gaining popularity. Their performance is determined from the amount and the quality of data used at their inputs, as well as from the competence and proper tuning of the ML algorithm used. However, collecting high-quality real data is time-consuming and expensive. Synthetic Data Generation (SDG) is therefore employed in order to augment the limited real data. Moreover, Ensemble Learning (EL) provides the framework to optimally combine a set of standalone ML algorithms (base learners), capitalizing on their individual strengths. Base learner diversity is essential to build a strong ensemble. The proposed method of Targeted Injection of Synthetic Data (TloSD) combines the EL and SDG concepts in order to further diversify the base learners' predictions, thus giving rise to an even stronger ensemble model. We have applied TloSD in two different Path Loss (PL) datasets, using two well-established SDG methods (namely SMOGN and CTGAN). While the conventional ensemble model reached a Minimum Absolute Error (MAE) value of 3.25 dB, the TloSD-triggered ensemble provided a MAE value of 3.16 dB. It is therefore concluded that targeted synthetic data injection, due to its diversity-triggering characteristics, enhances the ensemble's performance. Moreover, the ratio between synthetic and real data has been investigated. The results showed that a proportion of 0.1 is optimal.

Keywords: ensemble learning; synthetic data; path loss prediction; base learners; diversity



Citation: Sotiroudis, S.P. Cultivating Ensemble Diversity through Targeted Injection of Synthetic Data: Path Loss Prediction Examples. *Electronics* **2024**, *13*, 613. <https://doi.org/10.3390/electronics13030613>

Academic Editor: Alberto Fernandez Hilario

Received: 30 December 2023

Revised: 28 January 2024

Accepted: 30 January 2024

Published: 1 February 2024



Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

While ML is steadily gaining attention in an increasing variety of domains, data shortage poses a significant challenge [1]. The resource-intensive nature of measurement campaigns in a variety of scientific fields, such as antenna design [2], magnetic materials [3] and path loss prediction [4] constitutes a considerable difficulty regarding the implementation of ML applications. Two important ML concepts that can be used to tackle problems regarding the quality and the amount of real data, are SDG and EL.

SDG [5,6] can be used in order to combat data deficiency, by creating synthetic data and augment a real dataset's size and variety. Both images and tabular data can be synthetically generated. With reference to the production of synthetic tabular data, many techniques have been deployed, including the Synthetic Minority Oversampling Technique (SMOTE) [7], Generative Adversarial Networks (GANs) [8] and Large Language Models (LLMs) [9].

EL techniques enable the integration of multiple machine learning algorithms, trained on the same dataset [10–12]. A variety of ensembling techniques, facilitating the most efficient base learner combination of either the same (homogeneous) or different (heterogeneous) type has been developed. The essence of ensembling lies at combining diverse individual learners [13,14]. That is, the base learners should demonstrate different strengths and weaknesses, so as to be integrated within a meta-model that profits from their proper combination.

The fusion of SDG and EL techniques is beginning to appear in the ML literature. Applications regarding classification problems are presented in [15], where various EL and SDG methods are combined. The authors conclude that traditional SDG methods, such

as SMOTE, outperform those that are based on GANs. The concept of producing diverse synthetic datasets through multiple SDG models and then ensemble the individual models, is introduced in [16], where the authors conclude that this approach performs better in comparison to the creation of a single synthetic dataset. In [17], the authors use noise as the source of diversity in differential privacy synthetic data generation mechanisms. In [18], Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise (SMOGR) [19] is used in order to synthetically augment an imbalanced Path Loss (PL) dataset. The results showed that the ensemble model that incorporated synthetic data led to better results as opposed to the ensemble model that was trained only with the initial data, due to its enhanced predictive capability at the edges of the prediction interval.

All the previously presented works perform SDG on the initial training dataset (which is the base learners' training dataset). The novelty of our work is that data augmentation is performed in the meta-learner's training dataset, as opposed to the base learners' training dataset. That way, the synthetic data generator is informed regarding the predictions of the base learners. As a result, synthetic instances containing the prediction differences between the base learners can be constructed. The addition of the most diversity-triggering synthetic instances to the meta-learner's training dataset is expected to expand base learner diversity and produce an even stronger meta-model.

We introduce the Targeted Injection of Synthetic Data (TioSD) method, in order to select a diversity-triggering subset of the synthetic dataset and infuse it to the meta-learner's training dataset. We have chosen to apply the proposed method for the topic of PL prediction, since both the concepts of EL [20–23] and SDG [24–26], have been extensively utilized.

For that matter, we have used two different PL datasets [27,28] and applied two distinct synthetic data generators, namely SMOGR [19] and Conditional Tabular Generative Adversarial Network (CTGAN) [8]. Moreover, we have investigated the ratio of synthetic to raw data, with regard to the ensemble's performance. Our contributions can be summarized as follows:

1. We propose the method of Targeted Injection of Synthetic Data (TioSD) for the purpose of diversifying an Ensemble's Base Learners;
2. We apply TioSD in two different PL datasets, using two state-of-the-art methods of generating tabular data;
3. We investigate the ratio of synthetic to raw data, with regard to the improvement of PL prediction.

The rest of the paper is organized as follows: Section 2 presents the Machine Learning concepts of Ensemble Learning, Diversity and Synthetic Data Generation, while Section 3 sheds light on the proposed method of Targeted Injection of Synthetic Data. Section 4 is devoted in providing the comparative results between the conventional and the TioSD-based stacked generalization approaches, while Section 5 discusses the results. The conclusions of the paper are presented in Section 6.

2. Machine Learning Concepts: Ensemble Learning, Diversity and Synthetic Data Generation

The current section describes the concepts, along with the related techniques, of Ensemble Learning, Diversity and Synthetic Data Generation, which are of profound importance within the context of our proposed method. The first subsection depicts the most prevalent ensembling techniques. The second subsection is devoted to explaining the fundamental role of Diversity in Ensemble Learning, while the third subsection presents two state-of-the-art techniques for generating tabular data.

2.1. Ensemble Learning

Ensemble Learning refers to the concept of combining various learners (usually referred to as weak or base learners), in order to assemble a stronger meta-learner (or ensemble learner). The intuition behind the ensembling concept stems from the fact that the base learners exhibit different strengths and weaknesses; by appropriately combining them,

the ensemble learner can benefit from their individual advantages and provide enhanced predictions. There would be no point in ensembling identical base learners; in that case, the ensemble's predictions would be the same to those obtained from a single base learner. It is the diversity among the base learners that adds value to the ensembling procedure [13].

Various ensembling techniques can be found in the literature, both for classification and regression tasks. Regarding regression, wherein the problem of PL prediction falls, the concepts of averaging, weighted averaging [29] and stacked generalization [30] are mostly being used [31]. Averaging is the procedure where the ensemble's prediction is equal to the mean of the predictions from the base learners. While being very easy to implement, the downside of averaging is that all first-level predictions contribute equally to the final prediction, regardless of their individual strength.

Weighted averaging takes care of the above-mentioned issue: each base learner influences the ensemble's prediction according to a predefined performance criterion. That is, the predictions made from strong base learners hold a larger percentage of the final prediction's value, as opposed to the predictions from the weaker base learners. Though better from averaging, this technique does not take into account the particular strengths of each base learner. In other words, the predictions from a strong base learner would always outweigh those made from a weaker base learner, even in the cases where the weaker base learner takes precedence over the stronger one.

Stacked generalization [30] is a technique that provides a framework to finely combine the weak learners. Their predictions are used as inputs from a second-level meta-learner. The meta-learner is trained according to this new set of predictions, ensuring that the final prediction would optimally be influenced from the base learners.

The above-described ensembling techniques can be used in combining heterogeneous, as well as homogeneous, base learners. In the specific case where only homogeneous base learners are combined, the concepts of bagging [32] and boosting emerge [33]; the first refers to the combination of learners (usually regression trees) that are grown in parallel on the basis of different views from the training dataset. The second performs serial tree growth, where each new tree tries to compensate for the errors of the previous one.

2.2. The Role of Diversity in Ensemble Learning

Ensemble Learning provides the framework to constitute a meta-learner with upgraded predictive capability in comparison with the performance of its base learners. However, the improvement brought by the ensembling procedure is depended on the level of diversity between the base learners [34]. A combination of identical first level individual regressors would be obviously pointless, while also an ensemble of diverse, yet strongly erroneous, base learners would also lead to poor final predictions. A group of strong, yet diverse, base learners is needed in order to produce a powerful ensemble model.

Diversity can be thought of as a hidden dimension in the bias-variance decomposition of an ensemble loss [13]. More particularly, diversity can be conceived as a measure of model fit, in the same way with bias and variance, keeping in mind, however, that diversity describes the correlation among the base learners.

Under this assumption, the concepts of bagging and boosting can be revisited: in both techniques, diversification among the base learners (usually decision trees) is encouraged. In the case of bagging, diversity stems from randomly resampling the training data for each base learner. In boosting, diversity is cultivated by training each new base learner according to the errors of its predecessor. In conclusion, it is straightforward to claim that the success of the models that are either based on bagging (Random Forest [35]) or on boosting (XGBoost [36], LightGBM [37]) is due to their diversity-triggering implementation [13].

Thereupon, the research question of whether other methods could be elaborated in order to cultivate ensemble diversity emerges. Our approach towards that direction is based on the exploitation of synthetic data.

2.3. Synthetic Generation of Tabular Data

The performance of most ML models is strongly influenced by the amount of data that are available for their training. Data shortage [1] is an issue of high importance in the ML-domain. Synthetic Data Generation [5] is gaining attention for the purpose of combating data shortage and providing the amounts of needed data to the ML models. Focusing on tabular SDG, a multitude of approaches can be found in the literature.

SMOTE [7], which was originally aimed at classification tasks for imbalanced datasets, is one of the first attempts for creating synthetic tabular data. SMOTE-made synthetic instances are produced by interpolating a randomly chosen instance of the minority class with one of its k nearest minority class neighbors. Its extension for imbalanced regression datasets is SMOTE for Regression (SMOTER) [38], which performs oversampling upon the infrequently occurring instances. The oversampling techniques of SMOTER and Gaussian Noise, form together the SMOGN method [19]. That is, with respect to the distance among the randomly chosen underrepresented instances, oversampling is either performed through SMOTER, or through the addition of Gaussian Noise.

When using SMOTER, synthetic instances are generated through interpolation. In each iteration, a pair of rare instances is used: one acts as a seed case while the other is randomly chosen from the k -nearest neighbors of the seed. Their features are interpolated, while the new target value is calculated as a weighted average of their corresponding target values. Sequentially, each rare instance functions as a seed example throughout the process. The default value of k is five [19].

When applying Gaussian Noise, its magnitude is determined from the perturbation parameter. Higher perturbation values allow the addition of more noise to the original samples when generating the synthetic data points. As a result, the synthetic dataset's diversity is proportional to the perturbation's value. The default choice for perturbation is 0.02 [19,39]. An example of using the SMOGN method in a PL prediction problem can be found in [18].

Generative Adversarial Networks (GANs) were introduced in 2014 [40] and are capable of producing both synthetic images and tabular data. Their inner architecture consists of two Neural Networks, namely the Generator and the Discriminator. The Generator learns to produce synthetic data (either images or tabular data), according to the characteristics of the authentic data, while the Discriminator is assigned with the task of distinguishing the authentic data samples from the synthetic ones, as depicted in Figure 1. Being a two-player dynamic system, the ultimate goal of GAN training is to reach Nash equilibrium. With regard to tabular SDG, a variety of GANs, such as the Conditional Tabular GAN (CTGAN) [8] and the TableGAN [41] can be found in the literature. While the TableGAN employs min-max normalization within the $[-1.1]$ range for continuous values, the CTGAN uses the variational Gaussian Mixture model for every individual column.

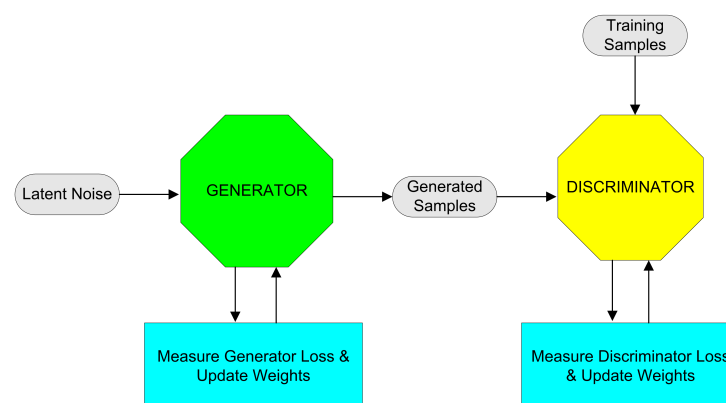


Figure 1. Block diagram of a Generative Adversarial Network (GAN).

3. Targeted Injection of Synthetic Data (TloSD) for Diversity Cultivation

The proposed method of Targeted Injection of Synthetic Data (TloSD), in order to further cultivate the diversity among an Ensemble's Base Learners and consequently enhance its predictive performance, is introduced in the first subsection. The second subsection presents the metrics that are used for the evaluation of the resulting ML models' performance.

3.1. Targeted Injection of Synthetic Data

An ensemble of regressors that performs stacked generalization, utilizes the outputs of its base learners as inputs for the meta-learner. The meta-learner is trained on this newly derived training set and is able to generalize his predictions on the corresponding testing set (which is also formed from the outputs of the base learners with regard to the initial testing set). We intent to augment the meta-learner's training set with synthetic data, aiming to enhance the diversity among the base learners, without degrading the ensemble's overall performance. In other words, our objective is to feed the meta-learner with additional, synthetically derived, diversity-triggering training instances and increase its performance (Figure 2).

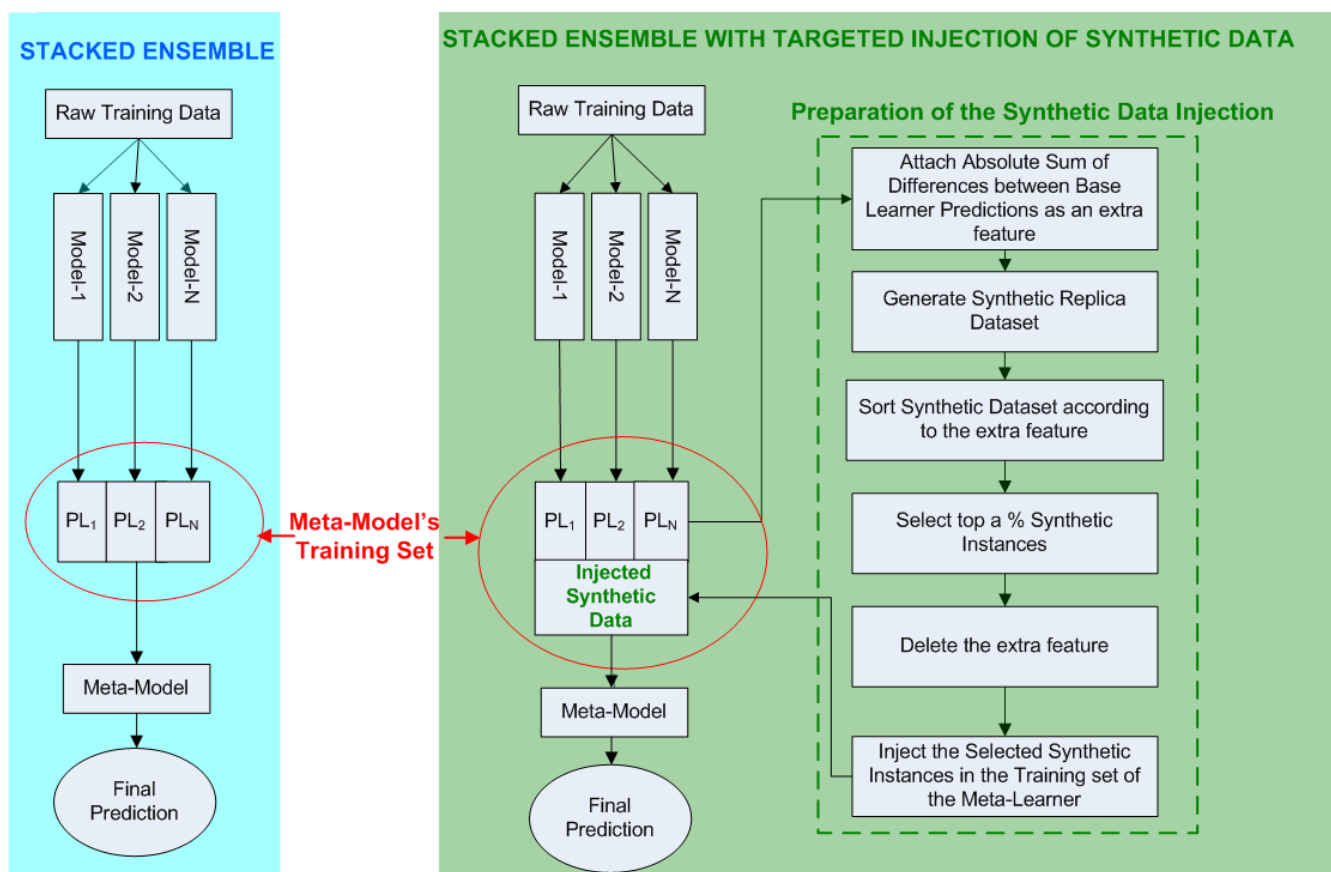


Figure 2. (Left) Conventional implementation of stacked generalization for N base learners. (Right) Implementation of TloSD-based stacked generalization for N base learners.

In order to better describe the proposed method, two algorithms are presented: the first illustrates the ensembling procedure, while the second details the injection of synthetic data. More precisely, lines 4–10 of Algorithm 1 are devoted in producing the predictions of all base learners for all instances of the training set X , using K -fold cross validation. These predictions are then used as inputs from the meta-learner (lines 11–12 of Algorithm 1).

At this point, Algorithm 2 is called in order to inject synthetic data to the *Training* set. As shown in lines 4–9 of Algorithm 2, the sum of absolute differences between base learner

predictions is calculated and added as an extra feature. This sum is an indicator of the overall diversity among the base learners. Then (lines 10–13 of Algorithm 2), a synthetic *Replica* dataset of the *Training* set is generated. Lines 13–18 of Algorithm 2 explain how the synthetic instances are sorted according to the *sum* column and how the top *a*% are chosen. After deleting the *sum* column, the chosen instances are injected as new rows into the *Training* set, transforming it to the *Training'* set.

Algorithm 1 continues using *Training'* to train the meta-learner (Linear Regressor). The performance metrics are derived using the *Testing* set.

Algorithm 1 TloSD-based Stacked Generalization Ensemble

- 1: Define number of folds $K = 5$ and number of base learners $N = 2$
 - 2: Split the training set X in K parts (folds)
 - 3: Define base learners: XGBoost, Random Forest
 - 4: **for** $n=1$ to N **do**
 - 5: **for** $k=1$ to K **do**
 - 6: Train the n -th base learner using all folds except the k -th one
 - 7: Obtain predictions PL_{nk} for the the k -th fold
 - 8: **end for**
 - 9: Create the prediction set $PL_n = PL_{n1} \cup PL_{n2} \cup PL_{n3} \cup PL_{n4} \cup PL_{n5}$ from the n -th base learner
 - 10: **end for**
 - 11: The input of the meta-learner is $X = \{PL_1, PL_2\}$, the output is the original PL
 - 12: Split the rows of the $\{PL_1, PL_2, PL\}$ set to *Training* and *Testing* (80/20 ratio)
 - 13: Call Algorithm 2 for the *Training* set:
 - 14: Train the Linear Regressor with (*Training'*) and test with *Testing*
 - 15: Calculate performance metrics in *Testing*
-

Algorithm 2 Synthetic Data Injection (SDG is SMOGN)

- 1: Define input set *Training*, which has $N + 1$ columns
 - 2: Define ratio *a* between synthetic and raw data
 - 3: Initialize $sum = 0$
 - 4: **for** $n = 1$ to N **do**
 - 5: **for** $i = n + 1$ to N **do**
 - 6: $sum = sum + abs(PL_n - PL_i)$
 - 7: **end for**
 - 8: **end for**
 - 9: Attach *sum* as an extra column in *Training*
 - 10: Call SMOGN [19] with the following arguments:
 - 11: The input set is *Training*, the target value is *sum*
 - 12: The number of nearest neighbors is 5 and the Gaussian noise perturbation is 0.02
 - 13: Obtain from SMOGN the synthetic *Replica* of *Training*
 - 14: Sort *Replica* according to descending value of *sum*
 - 15: Select top *a*% rows from *Replica*
 - 16: Delete column *sum* from the selected rows
 - 17: Add the selected rows to *Training*, as additional rows
 - 18: Synthetic Data has been injected, *Training* is transformed to *Training'*
-

3.2. Evaluation Metrics

Four well-known metrics of regression performance are used in order to evaluate the proposed method. These are the Mean Absolute Error (MAE), the Mean Absolute

Percentage Error (MAPE), the Root Mean Squared Error (RMSE) and the Coefficient of Determination (R^2). Their definitions are:

$$\begin{aligned}
 MAE &= \frac{1}{M_{test}} \sum_{m=1}^{M_{test}} |PL_m - y_{o,m}| \\
 RMSE &= \sqrt{\frac{1}{M_{test}} \sum_{m=1}^{M_{test}} [PL_m - y_{o,m}]^2} \\
 MAPE &= \frac{1}{M_{test}} \sum_{m=1}^{M_{test}} \left| \frac{PL_m - y_{o,m}}{PL_m} \right| \times 100\% \\
 R^2 &= 1 - \frac{\sum_{m=1}^{M_{test}} [PL_m - y_{o,m}]^2}{\sum_{m=1}^{M_{test}} [PL_m - PL_{mean}]^2}
 \end{aligned} \tag{1}$$

where M_{test} is the total number of test patterns, PL_m is the target PL value that corresponds to the m -th input pattern, $y_{o,m}$ is the ML model's output corresponding to the m -th input data pattern and PL_{mean} is the mean PL value.

4. Comparative Results

The proposed TIOsD-based ensembling method has been applied in two different datasets, using both the SMOGN and the CTGAN synthetic data generators. For both datasets, the tree-based algorithms XGBoost and Random Forest have been chosen as base learners, while the Linear Regression algorithm served as the meta-learner. The default hyper-parameter values have been employed for all learners and both synthetic data generators, in order to focus explicitly on the effect of synthetic data injection. The results for each dataset are respectively presented in the next two subsections.

4.1. Results from the First Dataset

The first dataset consists of 23 inputs and one output (the PL value), having a total of 35,378 instances. The PL values have been produced through the implementation of the Ray-Tracing technique from a commercial software [42]. The operating frequency is set at 900 MHz and the environment is urban. The dataset's input variables contain information regarding the built-up profile of the Line of Sight path between the transmitter and the receiver, the area around the receiver and their coordinates. A detailed presentation of the dataset's input variables can be found in [27]. A train/test ratio of 80/20 has been used in order to split the dataset.

The results are shown in Table 1, while Table 2 contains the error values derived for different ratios of synthetic to raw instances, denoted with a . Table 3 presents the MAE values of the conventional and the TIOsD-SMOGN stacked ensembles, for various combinations of the number of folds, K , and Base Learners, N . Figure 3 demonstrates the effect of TIOsD in the distribution of absolute difference among the base learner predictions, in the meta-learner's training set. It is straightforward to conclude that the number of instances that are associated with low differences between the predictions of the two base learners have not changed. On the other hand, the number of instances that correspond to high differences among the base learners predictions' has risen due to the implementation of TIOsD. Finally, Figure 4 presents the scatter plots and error distribution histograms of both ensembles.

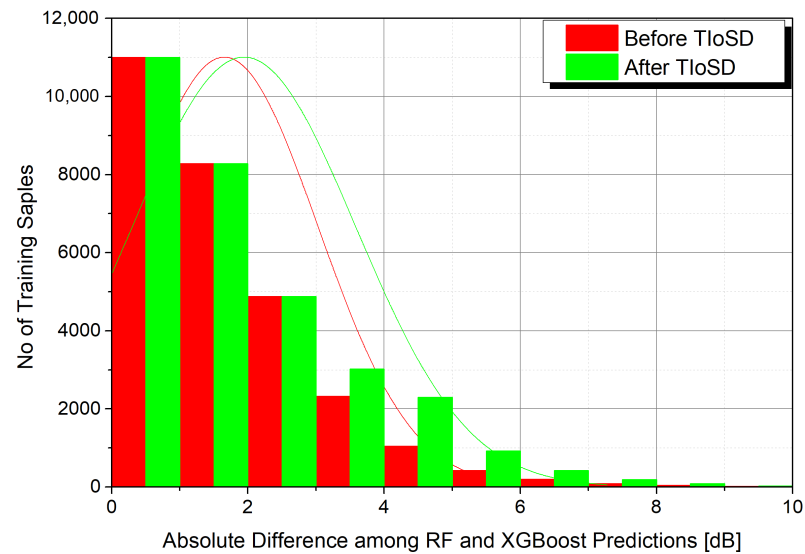


Figure 3. The implementation of TloSD increases the amount of training samples with enlarged distance among their base learner predictions. This diagram corresponds to a ratio of 0.1 between the chosen synthetic instances and the raw ones ($a = 0.1$). The SDG method is SMOGN.

Table 1. Comparative results for the first dataset.

ML Method	MAE [dB]	MAPE [%]	RMSE [dB]	R ²
XGBoost	3.36	3.10	4.43	0.897
Random Forest	3.41	3.15	4.50	0.894
Stacking-Conventional	3.25	3.01	4.28	0.904
Stacking-TloSD (SMOGN)	3.16	2.95	4.18	0.909
Stacking-TloSD (CTGAN)	3.23	3.00	4.26	0.905

For the TloSD implementations, the ratio a was chosen equal to 0.1.

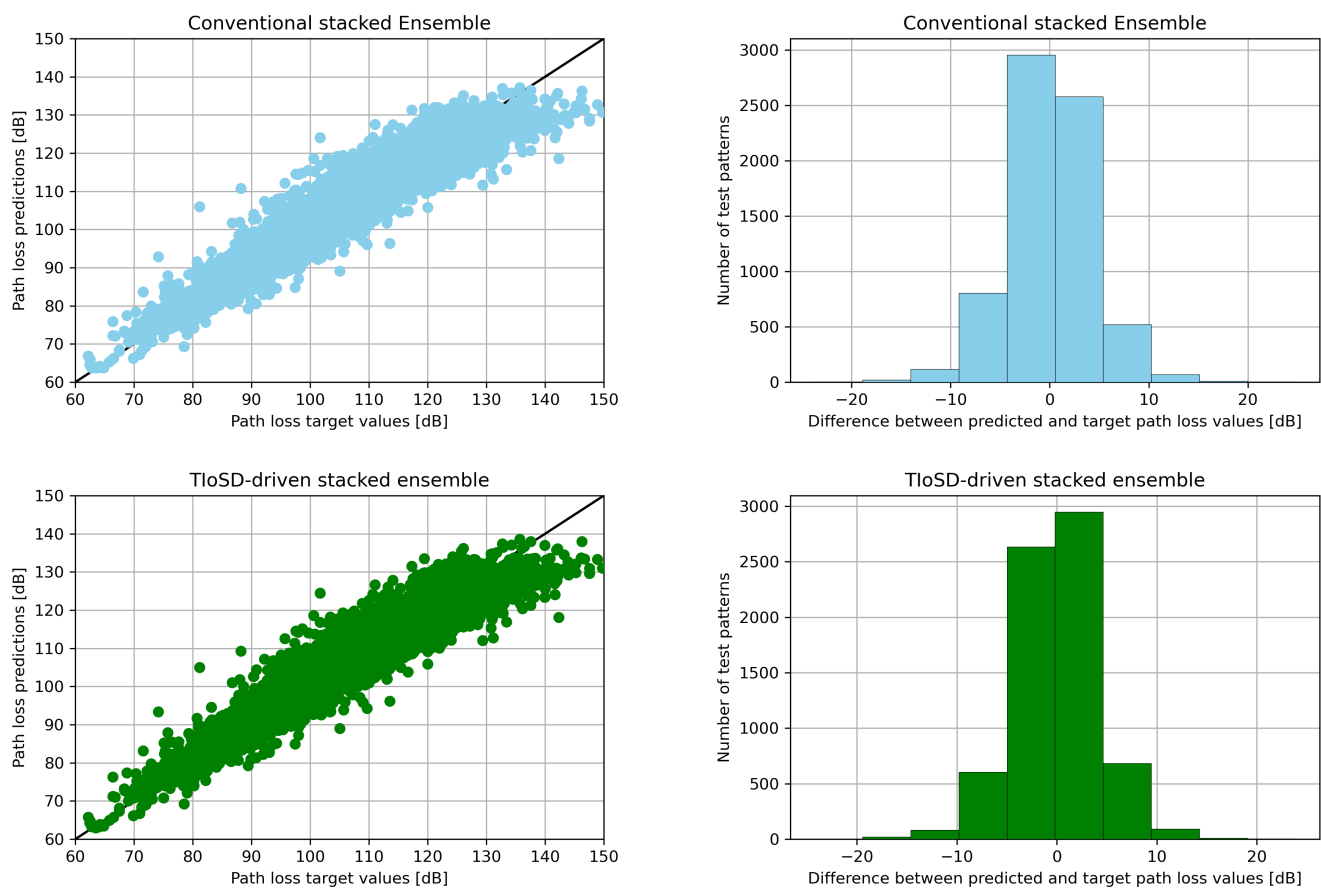
Table 2. Results of the TloSD-based ensembles, for the first dataset, according to the ratio a between synthetic and raw data .

SDG Method	a	MAE [dB]
SMOGN	0.05	3.20
	0.1	3.16
	0.2	3.23
	0.3	3.26
	0.4	3.29
	0.5	3.32
	1	3.38
CTGAN	0.05	3.23
	0.1	3.23
	0.2	3.25
	0.3	3.29
	0.4	3.33
	0.5	3.38
	1	3.41

Table 3. Comparative results for various K and N values (when $N = 3$, the third regressor is LightGBM)

K	N	MAE (Conventional) [dB]	MAE (TlIoSD-SMOGN) [dB]
2	2	3.30	3.18
2	3	3.29	3.17
5	2	3.25	3.16
5	3	3.24	3.16
10	2	3.24	3.15
10	3	3.23	3.14

For the TlIoSD implementations, the ratio a was chosen equal to 0.1.

**Figure 4.** (Left) Scatter plots (Right) error distribution histograms, for the conventional and the TlIoSD-driven ensembles, on the first dataset (using SMOGN and $a = 0.1$)

4.2. Results from the Second Dataset

The second dataset is publicly available in [43] and has been acquired from a measurement campaign in the city of Fortaleza-CE, Brazil [28]. The operating frequency is 853.71 MHz and the propagation environment is urban. A total of nine input variables, describing the coordinates of the receiver, its relative orientation with regard to the transmitter, the terrain elevation and the empirically calculated PL value according to the Okumura-Hata model [44], are used in order to predict PL. The measurements are performed on four different Base Stations. Our experiment uses 2328 data vectors, corresponding to the first Base Station. As with the case of the first dataset, a train/test ratio of 80/20 has been chosen. The results can be found in Tables 4 and 5 and in Figure 5.

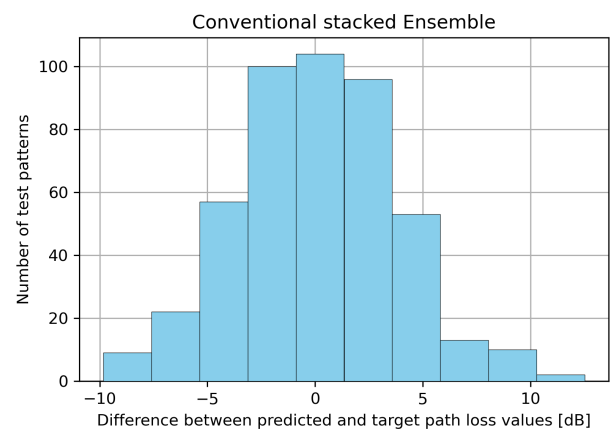
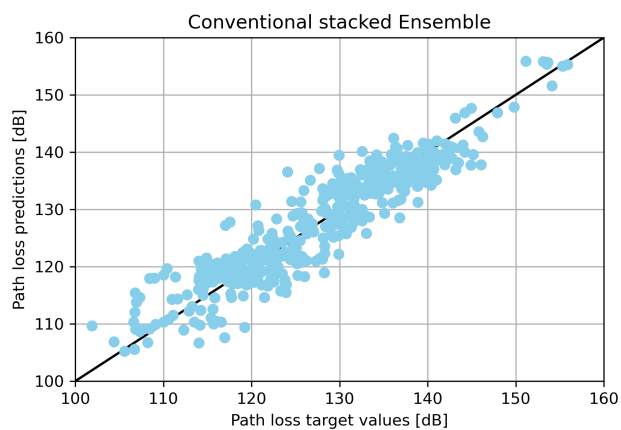
Table 4. Comparative results for the second dataset.

ML Method	MAE [dB]	MAPE [%]	RMSE [dB]	R ²
XGBoost	3.05	2.40	3.87	0.848
Random Forest	3.07	2.42	3.89	0.847
Stacking-Conventional	2.95	2.34	3.76	0.858
Stacking-TIoSD (SMOBN)	2.93	2.32	3.73	0.860
Stacking-TIoSD (CTGAN)	2.95	2.34	3.76	0.858

For the TIoSD implementations, the ratio a was chosen equal to 0.1.

Table 5. Results of the TIoSD-based ensembles, for the second dataset, according to the ratio a between synthetic and raw data.

SDG Method	a	MAE [dB]
SMOBN	0.05	2.95
	0.1	2.93
	0.2	2.96
	0.3	2.98
	0.4	3.01
	0.5	3.03
	1	3.07
CTGAN	0.05	2.95
	0.1	2.95
	0.2	2.97
	0.3	3.01
	0.4	3.03
	0.5	3.05
	1	3.08

**Figure 5.** Cont.

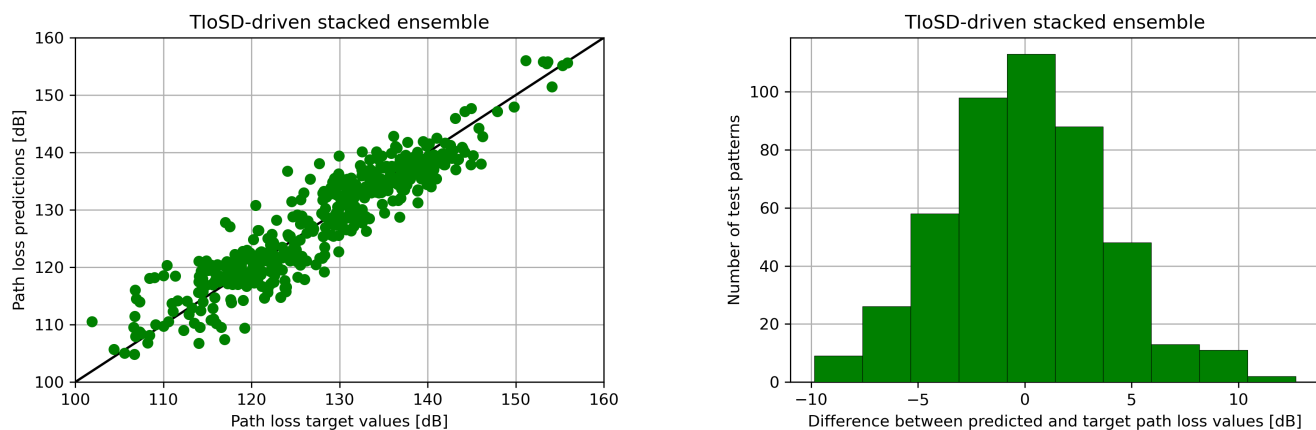


Figure 5. (Left) Scatter plots (Right) error distribution histograms, for the conventional and the TIoSD-driven ensembles, on the first dataset (using SMOGN and $a = 0.1$).

5. Discussion

The ratio a between the injected synthetic data and the raw training data is of crucial importance with regard to the TIoSD-based ensemble's performance. As concluded from the results, a relatively large ratio (more than 0.2 for the first dataset or 0.1 in the second dataset) leads to performance degradation in comparison with conventional stacking. Moreover, the more the ratio a increases beyond that point, the bigger the MAE becomes. Being a distorted version of the real data [45], large amounts of synthetic data tend to negatively influence the ensemble's generalization capability [16].

However, the TIoSD-based ensemble performs better than its conventional counterpart for smaller values of the ratio a , reaching an optimum performance when a becomes equal to 0.1 (for both datasets). That is, the effect of diversity cultivation through the injection of synthetic data, overcomes the negative influence of their synthetic nature, when injected in limited amounts.

Moreover, it is shown through Table 3 that increasing the number K of folds and N of base learners leads to smaller error values. However, the improvement brought by using ten instead of five folds is quite small (0.01 dB). Apart from that, the employment of three instead of two base learners, led also to a marginal improvement of 0.01 dB. It is worth mentioning that the third base learner utilized the LightGBM algorithm, which falls under the boosting category, similar to the first algorithm XGBoost. This is the reason why its incorporation within the ensemble did not have much to offer. For all combinations of K and N , the benefit of applying the TIoSD method led to an error reduction that fluctuated between 0.08 dB and 0.12 dB in comparison with the conventional ensemble.

For both datasets, the SMOGN method has led to better results than CTGAN [46] with regard to the TIoSD-based ensemble's performance. However, since the default hyper-parameter values of both SDG methods were used, CTGAN's performance could be enhanced by hyper-parameter tuning [39,47].

Finally, when comparing the error reduction between the conventional and the TIoSD-based ensemble for both datasets, one can observe an improvement of 0.09 dB for the first dataset and 0.02 dB for the second dataset. This can be attributed to the fact that the second dataset is one order of magnitude smaller than the first one, thus restraining the SDG methods from producing reliable synthetic data [48].

6. Conclusions

It has been shown that the proposed TIoSD method is efficient in reducing the prediction error of stacked generalization ensemble models that perform PL prediction. The method's objective is to cultivate the stacked ensemble's diversity by augmenting the meta-learner's training set through the injection of synthetic data. These data are chosen

according to their ability to further diversify the base learners' predictions. That is, the chosen synthetic data are those that correspond to the largest prediction differences among the base learners.

The ratio of synthetic to raw data instances is of crucial importance; being a distorted version of raw data, synthetic data should be injected to the extent that base learner diversity is cultivated, while the ensemble's performance is not degraded due to their artificial nature.

The optimum ratio of raw to synthetic data has been found to be equal to 0.1 for both experiments. However, further research is needed in order to provide a more systematic way to determine its value in conjunction with the raw dataset's characteristics, as well as with the base learner and SDG configurations.

The proposed method is general and can therefore be used as a framework to enhance the performance of stacked regression ensembles in various domains. Its main difference from other methods that employ SDG, lies in the fact that it facilitates the selection of an ensembling-oriented subset of the synthetic data.

The fusion of SDG and EL techniques through the proposed TIoSD method has led to models with increased diversity and generalization capability. However, the interpretability of the resulting models, as well as their computational complexity, should also be addressed in future research. Moreover, the method's performance is heavily dependent on the quality of synthetic data, which in turn is conditional on the SD generator's performance.

Funding: This research received no external funding

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CTGAN	Conditional Tabular Generative Adversarial Network
EL	Ensemble Learning
GAN	Generative Adversarial Network
LLM	Large Language Model
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
PL	Path Loss
R ²	Coefficient of Determination
RMSE	Root Mean Square Error
SDG	Synthetic Data Generation
SMOEN	Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise
SMOTE	Synthetic Minority Oversampling Technique
SMOTER	Synthetic Minority Oversampling Technique for Regression
TIoSD	Targeted Injection of Synthetic Data

References

1. Alzubaidi, L.; Bai, J.; Al-Sabaawi, A.; Santamaría, J.; Albahri, A.S.; Al-dabbagh, B.S.N.; Fadhel, M.A.; Manoufali, M.; Zhang, J.; Al-Timemy, A.H.; et al. A survey on deep learning tools dealing with data scarcity: Definitions, challenges, solutions, tips, and applications. *J. Big Data* **2023**, *10*, 46. [\[CrossRef\]](#)
2. Khan, M.M.; Hossain, S.; Mozumdar, P.; Akter, S.; Ashique, R.H. A review on machine learning and deep learning for various antenna design applications. *Heliyon* **2022**, *8*, e09317. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Katsikas, G.; Sarafidis, C.; Kioseoglou, J. Machine Learning in Magnetic Materials. *Phys. Status Solidi B* **2021**, *258*, 2000600. [\[CrossRef\]](#)
4. Seretis, A.; Sarris, C.D. An Overview of Machine Learning Techniques for Radiowave Propagation Modeling. *IEEE Trans. Antennas Propag.* **2022**, *70*, 3970–3985. [\[CrossRef\]](#)
5. Lu, Y.; Shen, M.; Wang, H.; van Rechem, C.; Wei, W. Machine Learning for Synthetic Data Generation: A Review. *arXiv* **2023**, arXiv:2302.04062.

6. Figueira, A.; Vaz, B. Survey on Synthetic Data Generation, Evaluation Methods and GANs. *Mathematics* **2022**, *10*, 2733. [\[CrossRef\]](#)
7. 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. [\[CrossRef\]](#)
8. Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; Veeramachaneni, K. Modeling Tabular data using Conditional GAN. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: San Francisco, CA, USA, 2019; Volume 32.
9. Borisov, V.; Seßler, K.; Leemann, T.; Pawelczyk, M.; Kasneci, G. Language Models are Realistic Tabular Data Generators. *arXiv* **2023**, arXiv:2210.06280.
10. Mienye, I.D.; Sun, Y. A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. *IEEE Access* **2022**, *10*, 99129–99149. [\[CrossRef\]](#)
11. Zhang, C.; Ma, Y. *Ensemble Machine Learning: Methods and Applications*; Springer Publishing Company, Incorporated: Berlin/Heidelberg, Germany, 2012.
12. Liu, S.; Qu, H.; Chen, Q.; Jian, W.; Liu, R.; You, L. AFMeta: Asynchronous Federated Meta-learning with Temporally Weighted Aggregation. In *Proceedings of the 2022 IEEE Smartworld, Ubiquitous Intelligence & Computing, Scalable Computing & Communications, Digital Twin, Privacy Computing, Metaverse, Autonomous & Trusted Vehicles (Smart-World/UIC/ScalCom/DigitalTwin/PriComp/Meta)*, Haikou, China, 15–18 December 2022; pp. 641–648. [\[CrossRef\]](#)
13. Wood, D.; Mu, T.; Webb, A.M.; Reeve, H.W.J.; Lujan, M.; Brown, G. A Unified Theory of Diversity in Ensemble Learning. *J. Mach. Learn. Res.* **2023**, *24*, 1–49.
14. Piwowarczyk, M.; Muke, P.Z.; Telec, Z.; Tworek, M.; Trawiński, B. Comparative Analysis of Ensembles Created Using Diversity Measures of Regressors. In *Proceedings of the 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Toronto, ON, Canada, 11–14 October 2020; pp. 2207–2214. [\[CrossRef\]](#)
15. Khan, A.A.; Chaudhari, O.; Chandra, R. A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Syst. Appl.* **2024**, *244*, 122778. [\[CrossRef\]](#)
16. Breugel, B.V.; Qian, Z.; Schaar, M.V.D. Synthetic Data, Real Errors: How (Not) to Publish and Use Synthetic Data. In *Proceedings of the 40th International Conference on Machine Learning*, Honolulu, HI, USA, 23–29 July 2023; pp. 34793–34808.
17. Sun, H.; Azizan, N.; Srivastava, A.; Wang, H. Private Synthetic Data Meets Ensemble Learning. *arXiv* **2023**, arXiv:2310.09729.
18. Sotiroidis, S.P.; Athanasiadou, G.; Tsoulos, G.V.; Christodoulou, C.; Goudos, S.K. Ensemble Learning for 5G Flying Base Station Path Loss Modelling. In *Proceedings of the 2022 16th European Conference on Antennas and Propagation (EuCAP)*, Madrid, Spain, 27 March–1 April 2022; pp. 1–4. [\[CrossRef\]](#)
19. Branco, P.; Torgo, L.; Ribeiro, R.P. SMOGN: A Pre-processing Approach for Imbalanced Regression. In *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, Skopje, Macedonia, 22 September 2017; PMLR: 2017; pp. 36–50.
20. Sotiroidis, S.P.; Athanasiadou, G.; Tsoulos, G.; Sarigiannidis, P.; Christodoulou, C.G.; Goudos, S.K. Evolutionary Ensemble Learning Pathloss Prediction for 4G and 5G Flying Base Stations With UAVs. *IEEE Trans. Antennas Propag.* **2023**, *71*, 5994–6005. [\[CrossRef\]](#)
21. Sotiroidis, S.P.; Boursianis, A.D.; Goudos, S.K.; Siakavara, K. From Spatial Urban Site Data to Path Loss Prediction: An Ensemble Learning Approach. *IEEE Trans. Antennas Propag.* **2022**, *70*, 6101–6105. [\[CrossRef\]](#)
22. Kwon, B.; Son, H. Accurate Path Loss Prediction Using a Neural Network Ensemble Method. *Sensors* **2024**, *24*, 304. [\[CrossRef\]](#)
23. Sani, U.S.; Malik, O.A.; Lai, D.T.C. Dynamic Regressor/Ensemble Selection for a Multi-Frequency and Multi-Environment Path Loss Prediction. *Information* **2022**, *13*, 519. [\[CrossRef\]](#)
24. Thrane, J.; Zibar, D.; Christiansen, H.L. Model-Aided Deep Learning Method for Path Loss Prediction in Mobile Communication Systems at 2.6 GHz. *IEEE Access* **2020**, *8*, 7925–7936. [\[CrossRef\]](#)
25. Kwon, B.; Kim, Y.; Lee, H. A Data Augmentation Approach to 28GHz Path Loss Modeling Using CNNs. In *Proceedings of the 2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC)*, Bali, Indonesia, 20–23 February 2023; pp. 825–829. [\[CrossRef\]](#)
26. Brennan, C.; McGuinness, K. Site-specific Deep Learning Path Loss Models based on the Method of Moments. *arXiv* **2023**, arXiv:2302.01052.
27. Sotiroidis, S.P.; Goudos, S.K.; Siakavara, K. Feature Importances: A Tool to Explain Radio Propagation and Reduce Model Complexity. *Telecom* **2020**, *1*, 114–125. [\[CrossRef\]](#)
28. Timoteo, R.D.; Cunha, D.C.; Cavalcanti, G.D. A proposal for path loss prediction in urban environments using support vector regression. In *Proceedings of the Tenth Advanced International Conference on Telecommunications*, Paris, France, 20–24 July 2014; pp. 1–5.
29. Mahendran, N.; Vincent, D.R.; Srinivasan, K.; Chang, C.Y.; Garg, A.; Gao, L.; Reina, D.G. Sensor-Assisted Weighted Average Ensemble Model for Detecting Major Depressive Disorder. *Sensors* **2019**, *19*, 4822. [\[CrossRef\]](#)
30. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [\[CrossRef\]](#)
31. Shahhosseini, M.; Hu, G.; Pham, H. Optimizing ensemble weights and hyperparameters of machine learning models for regression problems. *Mach. Learn. Appl.* **2022**, *7*, 100251. [\[CrossRef\]](#)
32. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [\[CrossRef\]](#)
33. Freund, Y.; Schapire, R. Experiments with a New Boosting Algorithm. In *Proceedings of the ICML'96: Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, Bari, Italy, 3–6 July 1996.

34. Bourel, M.; Cugliari, J.; Goude, Y.; Poggi, J.M. Boosting Diversity in Regression Ensembles. *Stat. Anal. Data Min.* **2020**. [CrossRef]
35. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
36. 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; pp. 785–794. [CrossRef]
37. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: San Francisco, CA, USA, 2017; Volume 30.
38. Torgo, L.; Ribeiro, R.P.; Pfahringer, B.; Branco, P. SMOTE for Regression. In *Proceedings of the Progress in Artificial Intelligence*; Correia, L., Reis, L.P., Cascalho, J., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2013; pp. 378–389. [CrossRef]
39. Wen, Z.; Wang, Q.; Ma, Y.; Jacinthe, P.A.; Liu, G.; Li, S.; Shang, Y.; Tao, H.; Fang, C.; Lyu, L.; et al. Remote estimates of suspended particulate matter in global lakes using machine learning models. *Int. Soil Water Conserv. Res.* **2024**, *12*, 200–216. [CrossRef]
40. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *arXiv* **2014**, arXiv:1406.2661.
41. Park, N.; Mohammadi, M.; Gorde, K.; Jajodia, S.; Park, H.; Kim, Y. Data Synthesis based on Generative Adversarial Networks. *Proc. VLDB Endow.* **2018**, *11*, 1071–1083. [CrossRef]
42. EDX Wireless Microcell/Indoor Module Reference Manual, Version 7 ©; EDX Wireless: Eugene, OR, USA, 1996–2011.
43. SVR PATHLOSS: Available online: https://github.com/timotrob/SVR_PATHLOSS (accessed on 28 December 2023).
44. Hata, M. Empirical formula for propagation loss in land mobile radio services. *IEEE Trans. Veh. Technol.* **1980**, *29*, 317–325. [CrossRef]
45. Jordon, J.; Szpruch, L.; Houssiau, F.; Bottarelli, M.; Cherubin, G.; Maple, C.; Cohen, S.N.; Weller, A. Synthetic Data—What, why and how? *arXiv* **2022**, arXiv:2205.03257.
46. Espinosa, E.; Figueira, A. On the Quality of Synthetic Generated Tabular Data. *Mathematics* **2023**, *11*, 3278. [CrossRef]
47. Hamad, F.; Nakamura-Sakai, S.; Obitayo, S.; Potluru, V. A supervised generative optimization approach for tabular data. In Proceedings of the Fourth ACM International Conference on AI in Finance, ICAIF '23, Brooklyn, NY, USA, 27–29 November 2023; pp. 10–18. [CrossRef]
48. Optimize Your Training Sample Size for Synthetic Data Accuracy. Available online: <https://mostly.ai/blog/synthetic-data-accuracy-vs-training-sample-size> (accessed on 28 December 2023).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.