IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 June 2024, accepted 9 July 2024, date of publication 16 July 2024, date of current version 24 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3427783

METHODS

Recursive Elimination of "Outliers" to Get Benchmark Dataset

LANGSHA LIU[1], CHUNHUI XIE[2], WENSHENG HU[1], AND YUNQI LI[2]
[1]Department of Information Engineering, Guizhou Communication Polytechnic University, Guiyang 551400, China
[2]College of Materials and Metallurgy, Guizhou University, Guiyang 550025, China

Corresponding author: Yunqi Li (liyq@gzu.edu.cn)

This work was supported in part by the National Science Foundation of China under Grant 22173094, in part by Guizhou Provincial Basic Research Program under Grant Z2024021 and Grant BQW[2024]006, in part by Guizhou University Talents Fund under Grant C0048072, and in part by the Application and Research of Intelligent Rescue Public Service Platform Based on Location Based Services (LBS) through the Science and Technology Project of the Department of Transportation of Guizhou Province under Grant 2022-321-013.

ABSTRACT Benchmark datasets normally have relatively conserved relationships and low fraction of outliers, indicated from higher determination coefficient (R^2) and lower Mean Absolute Error (MAE) in regression model. Here inspired by the process of peeling onions, we introduced a recursive data elimination (RDE) of ''outliers'' strategy to get benchmark dataset. Outliers are labeled using William's plot in residual vs leverage (recorded as RDE_W), and the performance was compared with that using residual alone (recorded as RDE). The validation was performed in single-target and multiple-target ways through the predictions of mechanical properties including Young's modulus, tensile strength, and elongation at break for 643 polyurethane elastomers (the first time this dataset has been released), and compressive strength for 1030 concrete samples. In the single-target way, RDE_W strategy achieved an 8.06% increase in R^2 and a 19.87% reduction in MAE compared to RDE. In the multiple-target way the improvement was approximately 3%. SVM outperformed XGB, NN, RF, Lasso and DT algorithms in the RDE_W strategy. Additional tests also validated the advantages for RDE_W over RDE to generate high-quality benchmark datasets. We released the data and code to facilitate the construction of high quality benchmark datasets and the development of new approaches to better understand, explore and design advanced materials.

INDEX TERMS Benchmark dataset, recursive data elimination, polyurethane elastomer, mechanical properties, regression.

I. INTRODUCTION

In statistical analysis and machine learning studies, an ideal dataset with comprehensive and representative coverage is highly appreciated to deliver theoretical soundness, clear causality and robust conclusions. Benchmark datasets are approaching such ideal state, characterized by high quality and consistency, and a low fraction of outliers. They have been widely used to stringently assess advancements in new models, strategies and algorithms [1], [2], [3], [4]. To build a benchmark dataset from a raw dataset, the Bias-Variance Dilemma (BVD) should be carefully treated to balance the deviation and the sensitivity of different strategies [5], the accuracy of machine learning-based models and their generalization ability [6]. Evaluation performance metrics for machine learning models such as the determination coefficient, R^2 for a regression model and the accuracy for a classification model normally exhibit relatively small changes upon the adding and deletion of partial data in the benchmark dataset, which becomes an important way to evaluate the overall quality of a dataset through cross-validation [7]. In reality, benchmark datasets are highly desired in material science, where new data are massively generated and reported. It is a time-consuming and non-trivial task to get the frontier for given types of materials, to grasp the rules based tunable variables toward advanced materials, and to explore the multivariate and synergistic quantitative relationships. Especially in materials science, high quality benchmark datasets become a top concern before the deployment of statistical analysis and machine learning studies [8], [9], [10], [11]. A general strategy to build high-quality benchmark dataset in material science by reducing outliers is invaluable.

The associate editor coordinating the review of this manuscript and approving it for publication was Jingang Jiang.

Reducing outliers is an essential way to build benchmark dataset, and various subjective and objective reasons lie behind the presence of outliers in raw datasets. The presence of outliers may lead to biased or erroneous conclusions [12], while datasets in materials science typically face high dimensionality, nonlinearity, heterogeneity, non-monotonicity and synergistic impacts,which pose great challenges for outlier detection [13]. It is feasible to reduce outliers in newly designed experiments with prior control, and to detect outliers in posterior treatment. A number of strategies have been reported to detect outliers based on principles in statistics, distance, density, clustering, classification, graph and neural network strategies [14]. Outliers are synonymous with abnormal values according to an assumed probability distribution, which can be identified according to given cutoffs in statistical metrics or confidence intervals, such as the six-sigma criteria and the general 1.5*IQR criteria in a box-plot. Outliers and normal data can be distinguished from their locations in a distribution, according to either parametric methods, such as Gaussian mixture models with global optimal instances [15], subspace learning and Gaussian mixture models [16], or non-parametric methods including kernel density functions [17], kernel local outlier factors [18], fast adaptive kernel density estimator [19] etc. Parametric methods are based on an assumed distribution function, and in non-parametric methods, prior assumption of distribution functions is unnecessary, which rely on the exact probability density instead [20]. Parametric methods rely on prior assumptions for the distributions of data, which may lead to overfitting and low computational efficiency in high-dimensional data with complex correlations. Non-parametric methods do not need such prior assumption but facing challenges to select optimal bandwidth, which are sensitive to noise and suffering from high computational complexity for large datasets with multiple magnitude distributions. Both methods confront challenges in parameter optimization and model interpretability [14], [21], [22]. Another method of outlier detection involves performing univariate anomaly correction on data following a normal distribution, setting a threshold of 99% where data points exceeding this threshold are identified as outliers and subsequently removed [23]. William's plot is a popular graph-based non-parametric outlier detection method, where based on the plot of residual vs leverage, data points located outside the two thresholds for residual and leverage are considered as outliers [24]. Through the recursive elimination of outliers, akin to the removal of outer layers during peeling onionsand following a framework used by recursive feature elimination (RFE) [25], it is possible to get high-quality benchmark datasets with conserved core distributions and correlations. The recursive data elimination (RDE) strategy has been applied to build a benchmark dataset for polyurethane elastomers (PUE) [26], resulting in more conservative multivariate prediction for mechanical properties distributed over four magnitudes.

Here we selected two datasets in material science to train and test the proposed "peeling onion" strategy. It was performed for single-target or multiple-target predictive regression models, built using six machine learning algorithms including RF, XGB, NN, SVM, Lasso and DT. The first dataset records the mechanical properties for PUEs, with 643 samples and each one has Young's modulus (YM, MPa), tensile strength (TS, MPa), and elongation at break (EB, %) which was collected and annotated in recent studies [26], [27]. This dataset is the first time to be released associated with this work. The second dataset contains 1030 concrete samples with compressive strength (CS, MPa) [28]. To predict these four mechanical properties, features in composition, processing, structure and measurements were also well organized. We present the methods, results and discussion in the following sections, the newly proposed strategy, utilizing William's plot for recursive elimination of outliers, shows significant advantages in building high-quality benchmark dataset.

II. METHODOLOGY

The workflow of this study is illustrated in Figure 1. Three stages were performed on statistical analysis, outlier identification, and the comparison of RDE and RDE_W strategies in single-target and multiple-target ways. The statistical metrics for these two datasets are presented in Table 1 and Figure S1 in the supplementary information. The mechanical properties in the PUE dataset have heavily tailed distributions. We then applied a logarithmic transformation to enhance their similarity to a normal distribution, indicating that the normal distribution has a skewness of 0 and kurtosis of 3, and the log values are closer to them. In the following section, where we build machine learning regression predictive models, we predict logYM, logTS and logEB without further notation.

The 643PUE dataset comprised 3 regression predictive properties: logYM, logTS and logEB, each associated with 20 features that accounted for the elements in composition-processing-structure-property-performance (CPSPP) relationships [29], [30]. For the composition, formulation was recorded by hard segment contents and ratios (CHS and R), molecular weight for polyols (PO_MW) and molecular volume (FCVm), topological features and physical parameters were computed using RDkit [31] for constitutional (count of atoms, groups, and bonds, NumNHCO, NumHAcceptors, RingCount etc.), connective (Chi indices), topological (BalabanJ, BertzCT), MOE-type (such as EState_VSA series), and molecular properties descriptors (TPSA) and polarity (MolLogP). Interactions between monomers were labeled using cohesive energy density (CED) and Flory-Huggins interactions (Fchi). Processing settings mainly included reaction temperatures (Tr1, Tr2) and the feed of monomers (PMStep), form methods (Form_method) and measurements settings (CSArea, StrainRate) were recorded.

The 1030 concrete dataset is well known and has one single property, the compressive strength was measured using a unified standard and procedure. There are 8 features in CPSPP, including the formulation of 5 components (Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer),

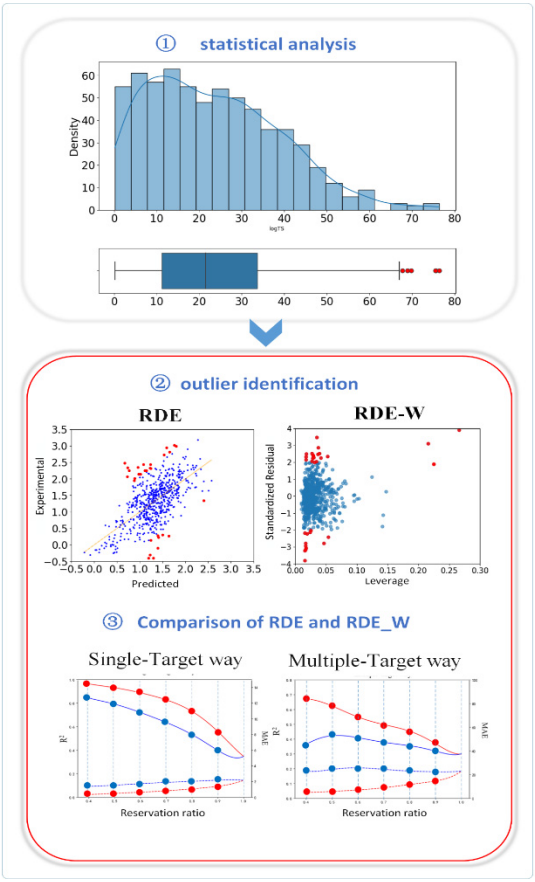

FIGURE 1. The workflow of this work including 1) statistical analysis of the four mechanical properties, 2) RDE using residual alone, RDE_W using both residual and leverage in William's plot to identify outliers, 3) comparison of RDE and RDE_W strategies in single-target and multiple -target ways using six machine learning algorithms.

2 structural parameters for the content of Coarse Aggregate, and Fine Aggregate, as well as the storage time before the measurement (Age).

A. "OUTLIER" IDENTIFICATION

William's plot involves standardized residual and leverage value, which are calculated based on the following definitions. The standardized residual (ze_i) measures the deviation of the observed values from the predicted values, scaled by the standard deviation of the residuals, defined as

$$ze_i = \frac{e_i}{s_e} = \frac{y_i - \widehat{y_i}}{s_e} \tag{1}$$

where ze_i and e_i are the standardized and the absolute residual for the i-th observation, y_i and $\widehat{y_i}$ are the i-th observed and predicted values, and s_e is the estimated standard deviation of the residuals. A general consensus for an absolute ze_i is regarded as an outlier, which is adapted in this work in the RDE strategy. The leverage values (h_i) measure the influence of each observation on the predicted values in a regression model, defined as [32]:

$$h_i = x_i^T (X^T X)^{-1} x_i \tag{2}$$

TABLE 1. Statistical summary for mechanical properties of 643 PUEs, and the concrete dataset with 1030 samples.

properties	No. features	Mean	Median	Variance	Skewness	Kurtosis
YM (MPa)	20	57.18	22.49	13875.39	6.43	57.23
LogYM (MPa)	20	1.31	1.35	0.43	-0.27	0.09
TS (MPa)	20	23.45	21.5	237.36	0.63	-0.05
LogTS (MPa)	20	1.23	1.33	0.17	-1.25	1.88
EB (%)	20	6.29	5.6	13.97	1.38	3.26
LogEB (%)	20	0.72	0.75	0.08	-0.91	1.72
CS (MPa)	8	35.82	34.45	279.08	0.42	-0.31

where x_i is the descriptor row-vector of the i-th compound, x_i^T is the transpose of x_i, X is the descriptor matrix, and X^T is the transpose of X. The warning leverage (h^*) [32] is usually set to:

$$h^* = 3(\mathrm{M}+1)/\mathrm{N} \tag{3}$$

where N is the total number of samples, and M is the number of features involved in the correlation. An outlier is identified by its location according to a combined threshold in (h_i, ze_i).

B. RECURSIVE DATA ELIMINATION (RDE) STRATEGY TO REDUCE OUTLIERS

To predict the four mechanical properties YM, TS, EB and CS, six machine learning algorithms including Lasso [33], Decision Tree [34], Random Forest [35], Neural Network [36], Support Vector Machine [37], and Extreme Gradient Boosting [38] were utilized to build regression predictive models. Five-fold cross-validation was used to train the models and predict the values in the testing set, and the absolute residuals and leverage values for each sample were calculated from the experimental and predicted values. The final values were obtained by averaging across the split of train-test datasets, using 10 different random seeds. The hyperparameters for these models were optimized under a Bayesian inference framework, following the method introduced by Ding et al [26], and the values are presented in Table S1 in the supplementary information.

For each input dataset starting from the raw, regression models were constructed using these well-structured features and mechanical properties. The coefficient of determination (R^2) and mean absolute error (MAE) were calculated from the average of five-fold cross-validation in each iteration. In RDE_W, the thresholds for the residual and leverage are fixed at ($h_i = 0.2, ze_i = 2$), while in RDE, ze_i is adaptive in a range of 3.83 to 2.04, which allows the elimination of the top 5% with the largest predictive residuals. In the single-target way, YM, TS, EB and CS are separately predicted, and the residual and leverage are calculated based on each pair of values from prediction and experiment. In the multiple-target way, YM, TS and EB in the PUE dataset were jointly considered. Firstly, based on the results of RDE and RDE_W, we assigned a quantitative score to each sample. The scoring method involved calculating the number of iterations that

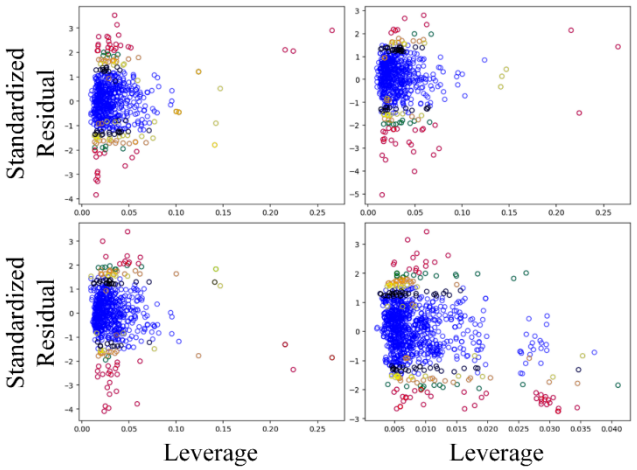

FIGURE 2. Illustration of RDE_W strategy to recursively remove outliers in the prediction of mechanical properties, colors from red, green, orange, yellow and black label the outliers in the 1st to the 5th iterations using LASSO to build regression models.

each sample remained, with a higher score given to samples that can survived more iterations of elimination. A sample assigned a value of 0 indicates it was eliminated at the first iteration due to large residual and leverage values. In order to make a simple comparison, besides the presentation of R^2 and MAE as function of data reservation ratio, we also directly selected their values at a given data reservation ratio of 0.7.

III. RESULTS AND DISCUSSION

Illustrations for the recursive elimination of outliers are presented in Figure 2 for RDE_W strategy. The schematic diagrams illustrating the process of outlier elimination in each iteration for the RDE and RDE_W strategies can be found in Figure S2. They clearly demonstrate that both strategies are efficient in removing data points located in the ''out-layers'' and gradually shifting towards the core region.

In the single-target way, the evolution of R^2 and MAE in the prediction of logYM using six different machine learning algorithms is shown in Figure 3. The corresponding predictions for logTS, logEB, and CS are presented in Figure S3-S5. All predictions show similar trends, where R^2 increases at a lower data reservation ratio, and MAE has the opposite trend. This clearly suggests that the removal of outliers from the raw dataset can increase the convergence for the complex correlations between mechanical properties and the composition, processing, and measurement variables. The robustness of regression models can be significantly improved by removing outliers, and this is a feasible way to build a benchmark dataset through the recursive elimination of outliers from the raw dataset. We also observed that these algorithms have different performances in the prediction of identical properties in these two datasets, and their dependence on the data reservation ratio may be strong for LASSO and DT in the prediction of logYM, logTS and CS, SVM and NN in the prediction of CS. The R^2 may be improved from around 0.4, a poor regression model, up to 0.9 in the test set, indicating a very robust model.

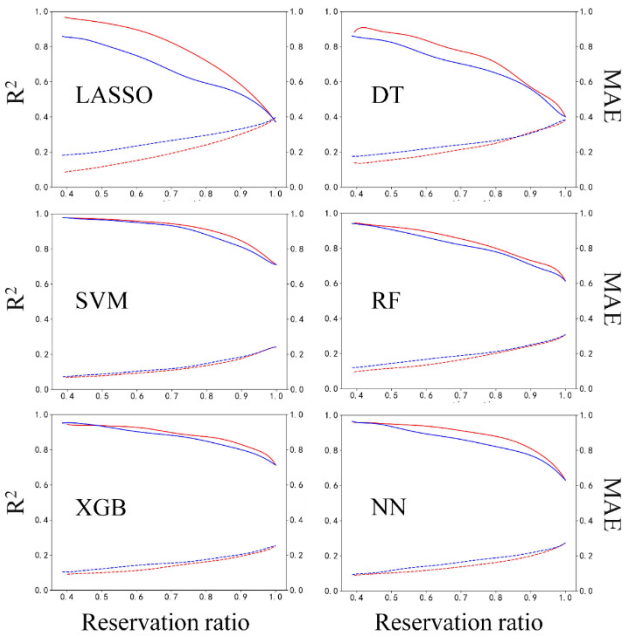

FIGURE 3. Comparison of RDE (blue) and RDE_W (red) in the removal of "outliers" in the prediction of logYM using different machine learning algorithms. Solid lines are R2 and dash lines are MAE as a function of data reservation ratio.

In the consideration of BVD to balance data coverage and the robustness of models, we selected a data reservation ratio of 0.7 to compare these two strategies using different machine learning algorithms. The comparison between R^2 and MAE is presented in Figure 4. The six popular machine learning algorithms exhibit different performance in the regression predictive models for the four mechanical properties. The compression strength distributed in the widely-used concrete dataset shows the most remarkable improvement for RDE_W compared to RDE, which is also indicated by the obvious difference in R^2 and MAE for the six machine learning algorithms. Overall, the RDE_W strategy achieved an advantage of 8.06% over RDE in increasing R^2, and 19.87% in reducing MAE at a fixed data reservation ratio of 0.7.

In the multiple-target way, logYM, logTS and logEB in the PUE dataset were used as multiple constraints to build a benchmark dataset, in comparison to the one reported before [26], [27]. From the iterative regression models, the R^2 and MAE for the reserved samples as a function of data reservation ratio is shown in Figure S6. It indicates that RDE_W only achieved marginal improvement over RDE, and the values were also extracted at a given reservation ratio of 0.7 and summarized in Table 2. The overall improvement of RDE_W over RDE strategy in the multiple-target way was around 3.16% in terms of increased R^2, and 3.37% in terms of reduced MAE. Among the six machine learning algorithms, SVM exhibited the best performance in both single-target and multiple-target ways.

To further check the BVD and the collective contribution of individual samples within the dataset, the ranking score was used to group samples, and the residuals and

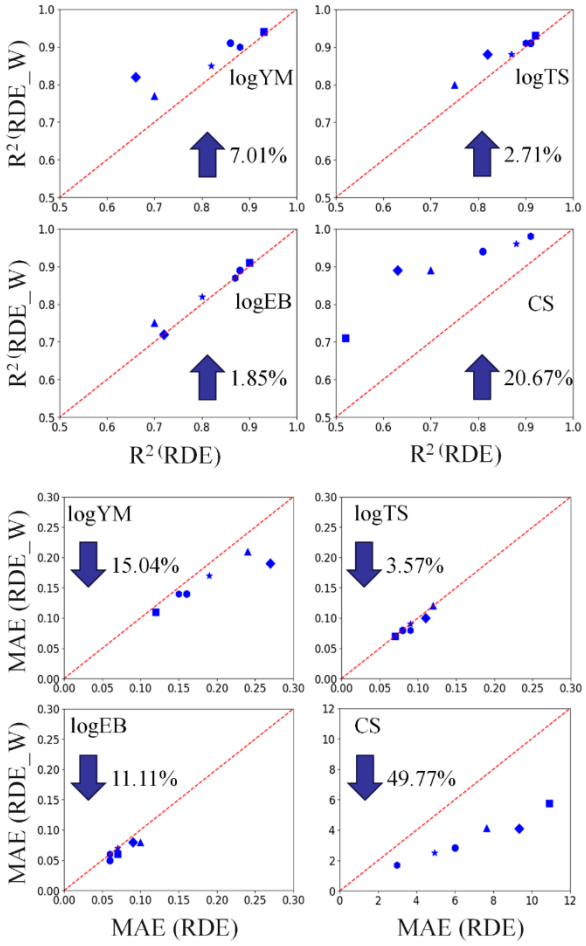

FIGURE 4. Comparison of R^2 and MAE for 4 mechanical properties, 6 algorithms, R^2 and MAE Q-Q plots of RDE vs RDE_W. The following shapes represent the corresponding machine learning algorithms: upward triangle, DT; circle, NN; square, SVM; pentagram RF; diamond LASSO and hexagon for XGB.

TABLE 2. Comparison of RDE_W and RDE strategies in the multiple-target way for the prediction of three mechanical properties in PUE dataset.

Algorithms	R^2		MAE	
	RDE_W	RDE	RDE_W	RDE
LASSO	0.63	0.61	0.19	0.20
DT	0.63	0.59	0.18	0.18
SVM	0.86	0.84	0.11	0.11
NN	0.83	0.81	0.12	0.13
RF	0.77	0.75	0.14	0.15
XGB	0.85	0.83	0.12	0.12

leverage for these grouped samples were analyzed and are shown in Figure 5. It can be seen that at the data reservation ratio between 0.2 to 0.4, both R^2 and MAE may show deleterious changes as more data was removed. This range is empirically regarded as the loss of representatives, similar to the determination of the variance of the projected points in PCA analysis [39]. When more data is reserved, R^2 slightly decreases and MAE slightly increases, these minor changes are derived from multiple reasons. The first one

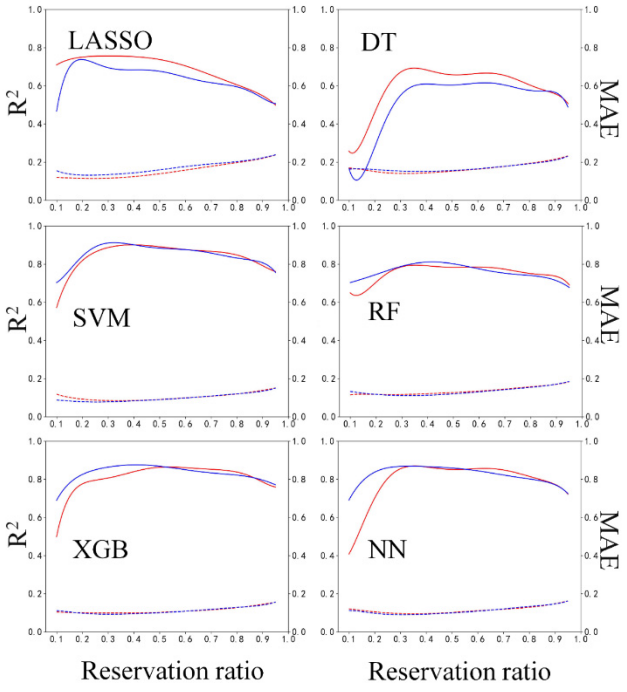

FIGURE 5. Build the benchmark dataset from the raw PUE dataset using YM, TS and EB multiple-target constraints, the R^2 and MAE as a function of data reservation ratio, in the comparison of RDE (blue) and RDE_W (red) strategies.

is the stiffness-extensibility trade-off, where YM and EB over different PUE samples exhibit different trends, it is physically impractical to simultaneously improve them. The second one is the polyurethane elastomers span between plasticity and elasticity according to their application scenarios, they do not have a conserved centroid, so for the distribution of these properties. Therefore, an efficient strategy to remove "outliers" to achieve conserved correlations for either single-target or multiple-target interests in the building of high-quality benchmark dataset, such as the RDE_W proposed here, is valuable for tackling particle problems.

To validate the advantage of RDE_W over RDE in terms of reliability, two well-known datasets from material science were collected for no-reference tests. One was initially used as regression predictive models for the Mohs hardness of naturally occurring ceramic materials (MH, MPa) with 622 samples [40], using atomic and electronic features from mineral compositions and crystal systems. The other dataset, annotated for the prediction of experimental band gap (BG, eV) [41], contains 2483 samples. The distributions for these two datasets are shown in Figure S7 and their statistical metrics were summarized in Table S2.

Following the framework used above, the iterative removal of outliers and the performance of predictive models built by six algorithms were presented in Figure S8 and S9, on the MH and BG datasets respectively. These two validation tests deliver consistent trend as those in PUE and concrete datasets. For the BG dataset, regression model built using RF algorithm is close to that from the original report [41], which

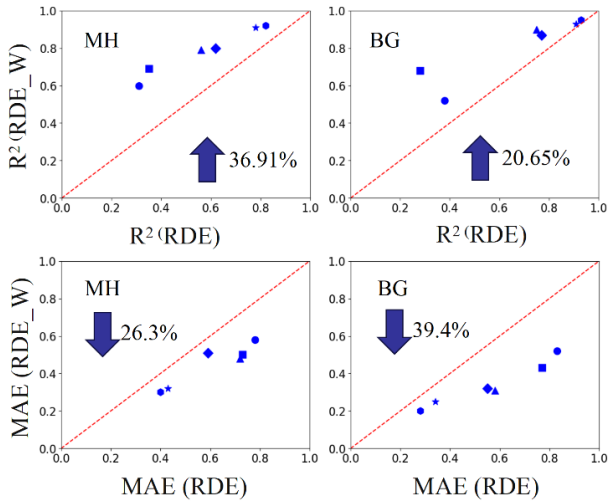

FIGURE 6. Comparison of R^2 and MAE - 2 individual targets, 6 algorithms, R^2 and MAE Q-Q plots of RDE vs RDE_W. The following shapes represent the corresponding machine learning algorithms: an upward triangle represents DT, a circle represents NN, a square represents SVM, a pentagram represents RF, a diamond represents LASSO and a hexagon represents XGB.

reported R^2 and MAE values of 0.81 and 0.44, respectively. By slightly removing 30% of outliers, R^2 can increase to 0.93, and MAE can decrease to 0.25. It indicates the removal of outliers is worthwhile for building robust predictive models. The comparison between RDE_W and RDE to maintain high robustness while reserve raw data as many as possible, was presented in Figure 6. The overall improvement of R^2 and MAE range from 20% to 40%, which aligns with the single-target prediction from the concrete dataset. These two tests again, validate the advantages of RDE_W over RDE to get high quality benchmark datasets through the iterative removal of outliers.

IV. CONCLUSION

In summary, we have released a newly collected dataset contains 643 polyurethane elastomers, with mechanical properties including Young's modulus, tensile strength, and elongation at break from tensile tests, all associated with their composition and processing details. A new strategy named as RDE_W was introduced to construct benchmark dataset with relatively conserved relationships and a higher percentage of raw data reserved. Its performance was validated on three variant datasets for the compressive strength of concretes, the hardness for natural ceramic materials and the band gap. When compared with previously reported RDE strategy, the RDE_W strategy achieved an impressive improvement in regression modeling for both single-target and multiple-target scenarios. It indicates that the consideration of both standardized residues and leverage values is a better way to balance Bias-Variance Dilemma in dataset with complex correlations. The new strategy can be adopted to alleviate the scarcity of benchmark datasets, which is a top concern in data curation prior to the deployment of statistical analysis and machine learning studies.

COMPETING INTERESTS

The authors declare no competing interests.

DATA AND CODE AVAILABILITY

The newly released PUE dataset containing 643 samples is available at https://www.scidb.cn/en/detail?dataSetId=faebba580a2e49efba6aa1eda9259c85, and the concrete dataset containing 1030 samples can be downloaded at https://www.kaggle.com/datasets/niteshyadav3103/concrete-compressive-strength. Python code for RDE and RDE_W are also available at supplementary files with corresponding names.

REFERENCES

[1] B. Koch, E. Denton, A. Hanna, and J. G. Foster, "Reduced, reused and recycled: The life of a dataset in machine learning research," presented at the 35th Conf. Neural Inf. Process. Syst. (NeurIPS), Sydney, NSW, Australia, 2021.

[2] A. N. Henderson, S. K. Kauwe, and T. D. Sparks, "Benchmark datasets incorporating diverse tasks, sample sizes, material systems, and data heterogeneity for materials informatics," *Data Brief*, vol. 37, Aug. 2021, Art. no. 107262, doi: 10.1016/j.dib.2021.107262.

[3] T. Eftimov, G. Petelin, G. Cenikj, A. Kostovska, G. Ispirova, P. Korošec, and J. Bogatinovski, "Less is more: Selecting the right benchmarking set of data for time series classification," *Expert Syst. Appl.*, vol. 198, Jul. 2022, Art. no. 116871, doi: 10.1016/j.eswa.2022.116871.

[4] S. Lu, "CodeXGLUE: A machine learning benchmark dataset for code understanding and generation," presented at the NeurIPS, 2021.

[5] S. B. E. Geman, *Neural Networks and the Bias/Variance Dilemma*. Cambridge, MA, USA: MIT Press, 1992.

[6] S. Schaal and C. G. Atkeson, "Constructive incremental learning from only local information," *Neural Comput.*, vol. 10, no. 8, pp. 2047–2084, Nov. 1998, doi: 10.1162/089976698300016963.

[7] T. Doğan and B. Karaçalı, "Automatic identification of highly conserved family regions and relationships in genome wide datasets including remote protein sequences," *PLoS ONE*, vol. 8, no. 9, Sep. 2013, Art. no. e75458, doi: 10.1371/journal.pone.0075458.

[8] S. M. McDonald, E. K. Augustine, Q. Lanners, C. Rudin, L. C. Brinson, and M. L. Becker, "Applied machine learning as a driver for polymeric biomaterials design," *Nature Commun.*, vol. 14, no. 1, p. 4838, Aug. 2023, doi: 10.1038/s41467-023-40459-8.

[9] L. Zhu, J. Zhou, and Z. Sun, "Materials data toward machine learning: Advances and challenges," *J. Phys. Chem. Lett.*, vol. 13, no. 18, pp. 3965–3977, May 2022.

[10] K. Zhang, X. Gong, and Y. Jiang, "Machine learning in soft matter: From simulations to experiments," *Adv. Funct. Mater.*, vol. 34, no. 24, Jun. 2024, Art. no. 2315177, doi: 10.1002/adfm.202315177.

[11] S. Mishra, B. Boro, N. K. Bansal, and T. Singh, "Machine learning-assisted design of wide bandgap perovskite materials for high-efficiency indoor photovoltaic applications," *Mater. Today Commun.*, vol. 35, Jun. 2023, Art. no. 106376, doi: 10.1016/j.mtcomm.2023.106376.

[12] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artif. Intell. Rev.*, vol. 22, no. 2, pp. 85–126, Oct. 2004, doi: 10.1023/b:aire.0000045502.10941.a9.

[13] S. Kadulkar, Z. M. Sherman, V. Ganesan, and T. M. Truskett, "Machine learning-assisted design of material properties," *Annu. Rev. Chem. Biomol. Eng.*, vol. 13, pp. 235–254, Jun. 2022, doi: 10.1146/annurev-chembioeng-092220-024340.

[14] H. Wang, M. J. Bah, and M. Hammad, "Progress in outlier detection techniques: A survey," *IEEE Access*, vol. 7, pp. 107964–108000, 2019, doi: 10.1109/ACCESS.2019.2932769.

[15] X. Yang, L. J. Latecki, and D. Pokrajac, "Outlier detection with globally optimal exemplar-based GMM," in *Proc. SIAM Int. Conf. Data Mining*, 2009, pp. 145–154.

[16] X.-M. Tang, R.-X. Yuan, and J. Chen, "Outlier detection in energy disaggregation using subspace learning and Gaussian mixture model," *Int. J. Control Autom.*, vol. 8, no. 8, pp. 161–170, Aug. 2015, doi: 10.14257/ijca.2015.8.8.17.

[17] L. J. Latecki, A. Lazarevic, and D. Pokrajac, "Outlier detection with kernel density functions," in *Machine Learning and Data Mining in Pattern Recognition*. Berlin, Germany: Springer, 2007, 61–75.

[18] J. Gao, W. Hu, Z. Zhang, X. Zhang, and O. Wu, *RKOF: Robust Kernel-Based Local Outlier Detection*. Cham, Switzerland: Springer, 2011, doi: 10.1007/978-3-642-20847-8_23.
[19] A. P. Boedihardjo, C.-T. Lu, and F. Chen, "Fast adaptive kernel density estimator for data streams," *Knowl. Inf. Syst.*, vol. 42, no. 2, pp. 285–317, Feb. 2015, doi: 10.1007/s10115-013-0712-0.
[20] E. Eskin, "Anomaly detection over noisy data using learned probability distributions," in *Proc. 17th Int. Conf. Mach. Learn.*, San Francisco, CA, USA, Jul. 2000, pp. 255–262.
[21] M. A. Samara, I. Bennis, A. Abouaissa, and P. Lorenz, "A survey of outlier detection techniques in IoT: Review and classification," *J. Sensor Actuator Netw.*, vol. 11, no. 1, p. 4, Jan. 2022. [Online]. Available: https://www.mdpi.com/2224-2708/11/1/4
[22] S. Borrohou, R. Fissoune, and H. Badir, "Data cleaning survey and challenges—Improving outlier detection algorithm in machine learning," *J. Smart Cities Soc.*, vol. 2, no. 3, pp. 125–140, Oct. 2023.
[23] Y. Wu, C. Qian, and H. Huang, "Enhanced air quality prediction using a coupled DVMD Informer-CNN-LSTM model optimized with dung beetle algorithm," *Entropy*, vol. 26, no. 7, p. 534, Jun. 2024, doi: 10.3390/e26070534.
[24] D. Wang, Y. Yuan, S. Duan, R. Liu, S. Gu, S. Zhao, L. Liu, and J. Xu, "QSPR study on melting point of carbocyclic nitroaromatic compounds by multiple linear regression and artificial neural network," *Chemometric Intell. Lab. Syst.*, vol. 143, pp. 7–15, Apr. 2015, doi: 10.1016/j.chemolab.2015.02.009.
[25] M. Kuhn and K. Johnson, *Discriminant Analysis and Other Linear Classification Models*. New York, NY, USA: Springer, 2013, doi: 10.1007/978-1-4614-6849-3_1.
[26] F. Ding, L.-Y. Liu, T.-L. Liu, Y.-Q. Li, J.-P. Li, and Z.-Y. Sun, "Predicting the mechanical properties of polyurethane elastomers using machine learning," *Chin. J. Polym. Sci.*, vol. 41, no. 3, pp. 422–431, Mar. 2023, doi: 10.1007/s10118-022-2838-6.
[27] F. Ding, T. Liu, H. Zhang, L. Liu, and Y. Li, "Stress-strain curves for polyurethane elastomers: A statistical assessment of constitutive models," *J. Appl. Polym. Sci.*, vol. 138, no. 39, p. 51269, Oct. 2021, doi: 10.1002/app.51269.
[28] I.-C. Yeh, "Modeling of strength of high-performance concrete using artificial neural networks," *Cement Concrete Res.*, vol. 28, pp. 1797–1808, 1998, doi: 10.1016/S0008-8846(98)00165-3.
[29] Y.-Q. Li, Y. Jiang, L.-Q. Wang, and J.-F. Li, "Data and machine learning in polymer science," *Chin. J. Polym. Sci.*, vol. 41, no. 9, pp. 1371–1376, Sep. 2023, doi: 10.1007/s10118-022-2868-0.
[30] L. Liu, F. Ding, and Y. Li, "Big data approach on polymer materials: Fundamental, progress and challenge," *ACTA POLYMERICA SINICA*, vol. 53, no. 6, pp. 564–580, 2022, doi: 10.11777/j.issn1000-3304.2021.21360.
[31] G. Landrum. (2016). *RDKit: Open-Source Cheminformatics Software*. [Online]. Available: https://github.com/rdkit/rdkit/releases/tag/Release_2016_09_4
[32] A. Atkinson, *Plots, Transformations, and Regression*. Oxford, U.K.: Clarendon Press, 1985.
[33] L. Freijeiro-González, M. Febrero-Bande, and W. González-Manteiga, "A critical review of LASSO and its derivatives for variable selection under dependence among covariates," *Int. Stat. Rev.*, vol. 90, no. 1, pp. 118–145, Apr. 2022, doi: 10.1111/insr.12469.
[34] S. B. Kotsiantis, "Decision trees: A recent overview," *Artif. Intell. Rev.*, vol. 39, no. 4, pp. 261–283, Apr. 2013, doi: 10.1007/s10462-011-9272-4.
[35] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/a:1010933404324.
[36] D. F. Specht, "A general regression neural network," *IEEE Trans. Neural Netw.*, vol. 2, no. 6, pp. 568–576, Sep. 1991, doi: 10.1109/72.97934.
[37] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, no. 3, pp. 199–222, Aug. 2004.
[38] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," presented at the 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Aug. 2016.
[39] J. Lever, M. Krzywinski, and N. Altman, "Principal component analysis," *Nature Methods*, vol. 14, no. 7, pp. 641–642, Jul. 2017, doi: 10.1038/nmeth.4346.
[40] J. Garnett, "Prediction of mohs hardness with machine learning methods using compositional features," in *Machine Learning in Chemistry: Data-Driven Algorithms, Learning Systems, and Predictions*. Washington, DC, USA: American Chemical Society, 2019, pp. 23–48.
[41] S. K. Kauwe, T. Welker, and T. D. Sparks, "Extracting knowledge from DFT: Experimental band gap predictions through ensemble learning," *Integrating Mater. Manuf. Innov.*, vol. 9, no. 3, pp. 213–220, Sep. 2020, doi: 10.1007/s40192-020-00178-0.

LANGSHA LIU received the Graduate degree from the School of Computing Science, University of Glasgow, U.K. Currently, he is a Big Data Engineer with the Department of Information Engineering, Guizhou Communication Polytechnic University. With plenty of work experience in IT companies, he is professional in Python and SQL. His current research interest includes developing methods to build high quality benchmark datasets.

CHUNHUI XIE received the B.S. degree in telecommunications engineering from Xidian University, Shaanxi, in 2020, and the M.S. degree in computer science from ShanghaiTech University, Shanghai, in 2023. He is currently pursuing the Ph.D. degree with Guizhou University, focusing on the development of methods for high-quality dataset acquirement using machine learning and graph neural networks in material science. He has published two articles on object detection in computer vision.

WENSHENG HU received the Ph.D. degree in engineering. He is currently with the Department of Information Engineering, Guizhou Communication Polytechnic University, where he is also the Executive Director of the Collaborative Innovation Center for Transportation Big Data. His research interests include software reliability and network security. He has led multiple research projects and published over 20 articles.

YUNQI LI received the B.S. degree in chemistry from Nanjing University and the Ph.D. degree in polymer physics from Changchun Institute of Applied Chemistry (CIAC). He is currently a Professor in polymer materials and engineering with Guizhou University. He was a Postdoctoral Researcher with Kansas University in bioinformatics and with Rutgers University in food chemistry. Since 2013, he has been a Professor with CIAC, dedicating in the structure and big data study of polymers till now. He has published 110 academic articles, four software copyrights, two book chapters, and one U.S. patent.