Data Imputation In Multi-Objective Semi-Supervised Explanation System

Yuheng Zhu
Department of Computer Science
College of Engineering
North Carolina State University
Raleigh, North Carolina 27606
Email: yzhu63@ncsu.edu

Mengzhe Wang
Department of Computer Science
College of Engineering
North Carolina State University
Raleigh, North Carolina 27606
Email: mwang39@ncsu.edu

Jiayuan Huang
Department of Computer Science
College of Engineering
North Carolina State University
Raleigh, North Carolina 27606
Email: jhuang52@ncsu.edu

***Abstract*—In recent years, semi-supervised learning has become an import research area in machine learning. Among them, multi-objective learning due to its outstanding performance in complex data domains has been particularly favored by researchers. However, missing data inevitably occurs in datasets obtained by sampling the real world. If such problem are not handled before or during the learning process, it can significantly affect the performance of the final model. To address this challenge, we introduces different data imputation methods into a multi-objective semi-supervised explanation system innovated by Chen et al.'s ”SWAY” method. Our proposed system improved the ”SWAY” method by redesigning the distance measurement when there are missing values in datasets, enabling us to obtain more accurate information. Specifically, we demonstrated the performance of our proposed method by experimenting multiple imputation methods on datasets in different densities of noise. We also employed multiple non-parametric tests on our results to verify the effectiveness. Our results suggest that our method has the potential to improve the accuracy and robustness of a multi-objective semi-supervised explanation system, especially when it is facing a dataset with high density of noise.**

***Index Terms*—Multi-objective, Semi-supervised Learning, Data Imputation**

I. Introduction

Increasingly, software engineering (SE) researchers are utilizing search-based optimization techniques to solve SE problems with multiple conflicting objectives. In these optimization problems, configuration parameters of a model need to be tuned such that the model generates ”good” outputs, which are demonstrably better than other possible outputs. While some engineers tend to use heuristic search algorithms like evolutionary algorithms [1] to provide a sufficiently good solution, these algorithms are time-consuming. To address this issue, Chen et al. proposed a method called `SWAY` that applies a sampling technique [2], reducing the total running time by decreasing the number of model evaluations. Based on `SWAY`, Dr. Menzies realized a multi-objective semi-supervised explanation system that can select the best subset of the data.

However, these state-of-the-art methods do not consider the influence of noise. In the process of collecting experimental data, missing or abnormal data often occurs. The usual approach to deal with noise present in many real-world optimization problems is to take an arbitrary number of samples of the objective function and use the sample average as an estimate of the true objective value. For evolutionary algorithms, Cantu-Paz introduced an adaptive sampling for noisy problems [3]. While Cantu-Paz demonstrated that the adaptive sampling could find better solutions, the drawback of that work is that it requires far more computation time.

In order to develop a multi-objective semi-supervised explanation system that can save computing time and deal with noise, we propose a method to improve Dr. Menzies' system. In Dr. Menzies' system, we believe that his approach to dealing with missing data is too arbitrary. Different methods for handling missing data may lead to different clustering, which may affect the final model. Additionally, considering the limited training data in semi-supervised learning, we believe that these outliers should be handled more cautiously. Therefore, we decided to interpolate these missing values to make more effective use of them.

To achieve the above goals, we modified the ”dist” function in the `SWAY` algorithm of Dr. Menzies' system to handle missing data using the MICE and KNN algorithms. We compared the results of the original model with two improved models under different noise conditions. To demonstrate the effectiveness of our approach, we explore the following research questions.

II. Structure of this paper

To demonstrate the effectiveness of our approach, we explored the following research questions.

*1) **RQ1**: Would `SWAY` experience performance degradation due to noise in the dataset?* Here we generated noisy datasets with multiple levels of densities, and use `SWAY` on them to measure the performance difference. Results showed significantly performance decrease as the density of noise increases.

*2) **RQ2**: Will different imputation methods help to improve the performance of `SWAY` on noisy datasets?* We assess the performance of two mainstream data imputation methods, MICE[4] and KNN[5], combining with `SWAY`, on noisy datasets. The clusters generated by `SWAY` combining with MICE and KNN leads to improved performance compared to original `SWAY`.

*3) **RQ3**: Would the above methods produce performance differences on datasets with varying densities of noise, measures in non-parametric significance tests?* At first glance, our results will appear to show that SWAY combining with imputation methods shows better robustness on all noisy datasets, but non-parametric significance tests told us our methods on low noise density datasets do not produce significantly improved results compared to the original method. But our approach produces significantly improved results on datasets with high noise density when applying clustering results onto the whole dataset.

In summary, we say our novel contributions are:

- We demonstrates the performance degradation of SWAY under varying levels of noise density in datasets.
- We proposed an improved SWAY method, which combines it with MICE and KNN imputation methods, evaluated and demonstrated the performance improvement of the combined approach.
- We demonstrated through non-parametric significance tests that our proposed method produces significantly improved results compared to the original SWAY method on datasets with high density of noise.

Our paper presents an innovative data selection method for machine learning that efficiently makes use of records with missing data by implementing intelligent methods. In such a case, we may get a larger number of high-quality data for model training and testing.

By discussing a critical challenge in the age of big data, this innovative approach has the potential to revolutionize the way machine learning models are trained, enabling more accurate predictions and better decision-making across various industries.

III. Related work

A. Multi-Objective Evolutionary Algorithms

When there are many objectives in our model, we call that a multi-objective problem. To solve such a problem, Evolutionary Algorithms create the initial population first, and then execute the crossover and mutation repeatedly until models have reached solutions that suffice for our purposes. Once a population is created, members of the population must now be evaluated according to a fitness function. A fitness function is a function that takes in the characteristics of a member, and outputs a numerical representation of how viable of a solution it is. As to the evaluation operator, the standard approach is, for each decision, run the underlying model to generate objective scores for those decisions. NSGA-II is a common evolutionary genetic algorithm [6], the core innovation in NSGA-II is its method of performing the selection. Candidates are sorted heuristically into bands according to how many other candidates they dominate.

B. SWAY

Chen et al. introduced a method SWAY that recursively clusters the candidates in order to isolate the superior cluster[2]. SWAY just selects a small superior set candidates among a group of candidates. To reduce model evaluation time, SWAY clusters the candidates by their decisions. After candidates are split into two parts according to their decisions, SWAY will prune half of them based on the objectives of the corresponding representatives. There is an important function named BETTER which is used to compare the representatives for two halves of the candidates.

C. Baseline Model

Dr. Menzies created a multi-goal semi-supervised explanation system for this course and we will take it as our baseline model. To select a best subset of the data, this system first calls SWAY in a subset of the data to choose the best data and some samples of the rest data (Using Zitzler's indicator [7] predicate to judge the domination status). Then it calls bins to find ranges that distinguish rows of data from best and rest. After that, it sorts these ranges by their scores (Probability*Support). Finally, it tries to find a best rule (a combination of ranges) to distinguish rows of data from best and rest and applies it to the whole dataset.

D. Semi-supervised Learning

Semi-supervised learning is a machine learning paradigm that lies between supervised and unsupervised learning. In semi-supervised learning, the learning algorithm is provided with a small amount of labeled data and a large amount of unlabeled data. The goal is to leverage both the labeled and unlabeled data to improve the performance of the learning model, particularly in scenarios where obtaining labeled data is expensive, time-consuming, or labor-intensive.

IV. Methods

A. Algorithms

1) SWAY: In general, SWAY is a sampling method to filter and select efficient and high-quality data for software engineering tasks. With detailed testing in different types of SE problems, SWAY is proved as an ideal choice for optimizing SBSE models for its simplicity and fastness. SWAY is a divide-and-conquer process that recursively clusters the candidates in order to isolate the superior cluster. As a result, SWAY is capable of returning a set of data from a large amount of chaotic data for better model training.

2) MICE: MICE stands for Multivariate Imputation by Chained Equations in R. MICE is a R package used for multiple imputations on missing data. There are 2 main approaches involved in MICE: joint modeling (JM) and fully conditional specication (FCS)[4].

3) KNN: The k-Nearest Neighbors (kNN) algorithm is a simple, yet effective, supervised machine learning technique used for classification and regression tasks. It works by finding the k training samples closest in distance to a new, unlabeled data point and determining the label or value for that point based on the majority vote or average of these nearest neighbors. The algorithm is non-parametric, meaning it makes no assumptions about the underlying data distribution, and its performance is highly dependent on the choice of distance metric and the value of k.

	auto2	auto93	china	coc-1000	coc-10000	healthCloseIsses 12mths0001-hard	healthCloseIsses 12mths0011-easy	nasa93 dem	pom	SSM	SSN
#x	19	5	18	20	22	5	5	25	10	13	17
#y	4	3	1	5	3	3	3	4	3	2	2
#rows	93	398	499	1000	10000	10000	10000	93	10000	239360	53662

TABLE I
SUMMARY FOR DATASETS

B. Data

There are 11 datasets used in this project with multiple goals. See in Table I.

For each of the dataset, there are 3 dimensions for illustrating the dataset:

1) What is the point of the data
2) What are its column names
3) How do they group and what do they mean

We use auto2 [8] as an example here:

1) This database appears to be focused on providing detailed information about various car models, covering aspects such as fuel efficiency, safety features, performance, and dimensions. The data could be used for various purposes, including helping buyers make informed decisions, analyzing trends in the automotive industry, or informing regulations and standards.
2) For column table see Table II.

Column names	
maker	type
CityMPG+	HighwayMPG+
Air_Bags_standard	Drive_train_type
Number_of_cylinders	Engine_size
Horsepower	RPM
Engine_revolutions_per_mile	manual_transmission_available
Fuel_tank_capacity	Passenger_capacity
Length	Wheelbase
Width	U-turn_space
Rear_seat_room	Luggage_capacity
Weight-	domestic
Class-	

TABLE II
COLUMN NAMES

3) We may group columns as:
   a) General Information:[maker, type, domestic, Class]
   b) Fuel Efficiency: [CityMPG, HighwayMPG]
   c) Safety Features: [Air_Bags_standard]
   d) Performance and Specifications:
      - Drive_train_type
      - Number_of_cylinders
      - Engine_size
      - Horsepower
      - RPM
      - Engine_revolutions_per_mile
      - manual_transmission_available
      - Fuel_tank_capacity
   e) Capacity and Dimensions:
      - Passenger_capacity
      - Length
      - Wheelbase
      - Width
      - U-turn_space
      - Rear_seat_room
      - Luggage_capacity
      - Weight

V. RESULTS

We divide this section into three parts. First, we analyzed the performance of original SWAY on different densities of multiple datasets. Second, we assessed original SWAY with our imputation-combined approaches on UCI Auto MPG Data Set[8] to verify the improvements. Third, we used Kruskal-Wallis test and Mann–Whitney U test with Bonferroni correction to prove the significant improvement we made. The source code and raw output can be found in our GitHub repository[9].

A. ***RQ1:*** *Would SWAY experience performance degradation due to noise in the dataset?*

To answer this question, we artificially inserted invalid data into multiple datasets as noise by controlling the insertion ratio to create two different datasets: one with a 1% random noise ratio, which we consider as a low noise density dataset, and another with a 5% random noise insertion ratio, which we consider as a high noise density dataset. We performed clustering explanation using the baseline model on these two datasets and compared the results with the data obtained on the original dataset. Results are shown in Table III. ”Better” column means it performs better than that on original datasets, ”worse” means worse and so on. From the table, we can observe that the performance of the baseline model deteriorates as we add noise to the dataset, with a higher noise density leading to worse performance. Upon analyzing the source code of the baseline model, we discovered that the model normalizes all missing values to 1 when calculating the distance between data points, which leads to unstable clustering results as noise increases. This indicates that the baseline model lacks robustness in such scenarios.

B. ***RQ2:*** *Will different imputation methods help to improve the performance of SWAY on noisy datasets?*

To discuss this issue, we need to first select a dataset as a benchmark test set. We chose the UCI Auto MPG Data Set[8] mentioned earlier because it contains multiple attributes, has an appropriate data size, and can well reflect the advantages

Datasets	Low Noise			High Noise		
	Better	Same	Worse	Better	Same	Worse
auto2.csv	2	4	2	1	2	5
auto93.csv	1	2	3	1	2	3
china.csv	0	2	2	0	2	2
coc1000.csv	0	4	6	0	4	6
coc10000.csv	1	0	5	2	0	4
nasa93dem.csv	2	0	6	3	0	5
healthCloseIsses12mths0001-hard.csv	1	3	2	0	2	4
healthCloseIsses12mths0011-easy.csv	0	3	3	2	2	2
pom.csv	2	1	3	0	4	2
SSM.csv	2	0	2	0	1	3
SSN.csv	1	0	3	0	0	4
Total Count	12	19	37	9	19	40

TABLE III
ORIGINAL SWAY PERFORMANCE ON DIFFERENT DATASETS

of the multi-objective semi-supervised model. We integrated the baseline model with two imputation methods, MICE and KNN. Using the baseline as a benchmark, we conducted 20 independent runs on the two datasets with different noise densities mentioned in **RQ1**. The mean values of 20 results are displayed in Table IV and Table V, respectively.

The columns represent multi-objects that the model needs to predict, where the ”+” sign indicates that a higher value in the column is closer to the ideal result, while the ”-” sign indicates the opposite. For each column, ”all” represents the mean value of random sampled data from the dataset, and it should be worse than all other results. The ”top” row represents the results obtained by sorting the entire dataset using the Zitzler predicate[7] and selecting the optimal subset, and it should represent the most ideal result of the model’s performance. For each model, ”sway” means that the model recursively searches for the optimal subset of data from the dataset, while ”xpln” applies the explanation of ”sway” method to the entire set to obtain a subset. We can observe that the sway values of all models are better than the xpln values. This is because sway is free to combine influences from multiple attributes while xpln simplifies that approach which causes information loss.

	CityMPG+	Class-	HighwayMPG+	Weight-
all	21.0	17.7	28.0	3040.0
sway	28.3	9.1	33.7	2198.2
xpln	28.8	11.2	33.4	2341.6
sway+MICE	28.9	9.1	33.5	2208.8
xpln+MICE	29.9	9.5	34.5	2231.0
sway+KNN	28.7	9.0	33.7	2187.4
xpln+KNN	30.1	9.3	34.4	2215.8
top	35.2	8.6	41.7	2045.0

TABLE IV
COMPARE APPROACHES ON LOW NOISE DATASET

	CityMPG+	Class-	HighwayMPG+	Weight-
all	21.0	17.7	28.0	3040.0
sway	28.1	10.0	33.9	2270.8
xpln	25.8	11.8	31.3	2534.2
sway+MICE	28.8	9.4	34.4	2214.2
xpln+MICE	28.9	9.5	33.6	2249.5
sway+KNN	29.6	9.0	34.5	2189.5
xpln+KNN	30.1	9.3	34.4	2197.2
top	36.0	8.6	42.0	2045.0

TABLE V
COMPARE APPROACHES ON HIGH NOISE DATASET

Figure 1 shows the distribution of the result of 20 independent runs of our approaches compared to baseline model on two different datasets. ”ORIGIN” represents baseline model SWAY, ”MICE” represents the model combined with multivariate imputation by chained equation method and ”KNN” represents the model combined with K-Nearest Neighbor method. Most of the box plots show a trend of original $<$ MICE $<$ KNN, and the distributions of MICE and KNN are generally denser than the original. Therefore, we can conclude that the model with the two introduced imputation algorithms can produce better results than the baseline model by effectively removing the outliers to varying degrees in noisy datasets.

Comparing the data between different models, we can see that our method has demonstrated better results on both low and high noise datasets. Moreover, the difference between our method and the baseline model is more pronounced in high noise datasets than in low noise datasets for the same model. From this, we can conclude that introducing different imputation algorithms to the baseline model can effectively improve its performance on noisy datasets.

C. **RQ3:** *Would the above methods produce performance differences on datasets with varying densities of noise, measures in non-parametric significance tests?*

To answer this question, we have introduced the two non-parametric significance testing methods mentioned earlier in

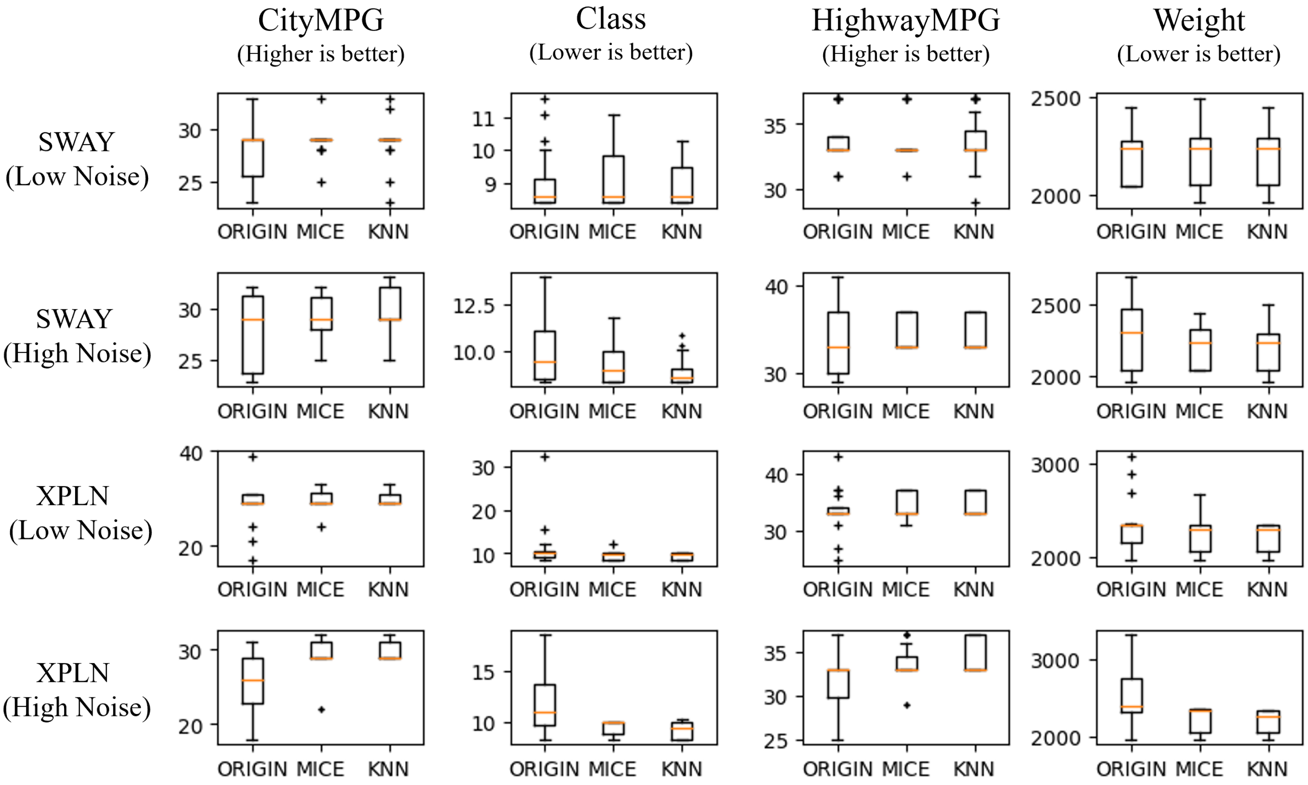

Fig. 1. Box-plots of Approaches on Low And High Density of Datasets

the paper.

1) Kruskal-Wallis Test: We first used the Kruskal-Wallis test to detect differences between the results of different models. We chose a significance level of p=0.05. The results for the low noise and high noise datasets are shown in Table VI and Table VII, respectively. ”=” means there is no significant difference between two model’s results on this attribute and ”≠” means we can say there is significant difference two results and thus we can combine the results from **RQ2** to determine which one is significantly better.

	CityMPG+	Class-	HighwayMPG+	Weight-
all to all	=	=	=	=
all to sway	≠	≠	≠	≠
all to sway+MICE	≠	≠	≠	≠
all to sway+KNN	≠	≠	≠	≠
sway to sway+MICE	=	=	=	=
sway to sway+KNN	=	=	=	=
all to xpln	≠	≠	≠	≠
all to xpln+MICE	≠	≠	≠	≠
all to xpln+KNN	≠	≠	≠	≠
xpln to xpln+MICE	=	=	=	=
xpln to xpln+KNN	=	=	=	=
sway to top	≠	=	≠	≠

TABLE VI
KRUSKAL-WALLIS TEST ON LOW NOISE DATASET

From the two tables above, it can be concluded that under low noise conditions, all three approaches outperform random sampling significantly. However, the model with the added imputation methods did not show a significant statistical difference from the baseline model. In contrast, on the dataset with high noise levels, the two models that applied the interpolation algorithm showed a significant statistical improvement in the results produced when interpreting the dataset, compared to the baseline model. This demonstrates that our method effectively addresses the shortcomings of the baseline model on high noise datasets, improving its accuracy and robustness.

	CityMPG+	Class-	HighwayMPG+	Weight-
all to all	=	=	=	=
all to sway	≠	≠	≠	≠
all to sway+MICE	≠	≠	≠	≠
all to sway+KNN	≠	≠	≠	≠
sway to sway+MICE	=	=	=	=
sway to sway+KNN	=	=	=	=
all to xpln	≠	≠	≠	≠
all to xpln+MICE	≠	≠	≠	≠
all to xpln+KNN	≠	≠	≠	≠
xpln to xpln+MICE	≠	≠	≠	≠
xpln to xpln+KNN	≠	≠	≠	≠
sway to top	≠	≠	≠	≠

TABLE VII
KRUSKAL-WALLIS TEST ON HIGH NOISE DATASET

2) Mann-Whitney U Tests: To further validate our results, we also conducted Mann-Whitney U tests, and the pairwise p value results are shown in Table VIII. By applying pairwise

Mann-Whitney U tests on all dependent attributes in dataset, and we chose a significant difference level of 0.05. Through testing with different noise levels and different attributes, we can obtain the final conclusion, as shown in Table IX. The experimental results show that although our method is significantly superior to the baseline model on high noise datasets in Kruskal-Wallis tests, only the explanation model using KNN imputation can be said to have significant superiority over all other methods in the ”Weight” attribute from the dataset we chose.

	baseline	MICE	KNN
baseline	0.000000	0.016915	0.000727
MICE	0.016915	0.000000	0.036590
KNN	0.000727	0.036590	0.000000

TABLE VIII
P VALUES IN PAIRWISE MANN-WHITNEY U TESTS WITH BENJAMINI/HOCHBERG CORRECTION

	CityMPG+	Class-	HighwayMPG+	Weight-
High Noise	x	x	x	xpln+KNN

TABLE IX
BEST PERFORMANCE MODEL ON DIFFERENT ATTRIBUTES

VI. DISCUSSION

A. Threats to Validity

1) Construct Validity: For instance, we applied the Benjamini/Hochberg correction to the computed p-values prior to conducting the Mann-Whitney U test to ensure their posterior validity. Moreover, we selected a significance level of 0.05 to better demonstrate the significant performance improvement of our method.

VII. CONCLUSION & FUTURE WORK

Finding the correct and high-quality dataset is a prerequisite for successful training in the field of machine learning. Better dataset selection can result in improved model performance. However, when dealing with missing values in the original dataset, ignoring or simply fit them during the selection process can degrade the quality of the selected dataset. This study explores the feasibility of introducing imputation algorithms into a semi-supervised multi-objective explanation system for dataset selection. Through experimentation, we demonstrate that the explanation system incorporating the MICE algorithm or KNN algorithm exhibits significantly higher accuracy and robustness than the original method when noise density increases. However, when using the Mann-Whitney U test for significance testing, we found that our method only demonstrates significant improvement in a small number of dataset attributes, indicating that there is still significant room for improvement. Therefore, while attempting to adjust or introduce other interpolation methods, we also consider using cross-validation and hyperparameter tuning[10] to further enhance the algorithm’s performance and robustness.

REFERENCES

[1] J. H. Holland, “Genetic algorithms,” *Scientific american*, vol. 267, no. 1, pp. 66–73, 1992.

[2] J. Chen, V. Nair, R. Krishna, and T. Menzies, ““sampling” as a baseline optimizer for search-based software engineering,” *IEEE Transactions on Software Engineering*, vol. 45, no. 6, pp. 597–614, 2018.

[3] E. Cantú-Paz, “Adaptive sampling for noisy problems,” Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States), Tech. Rep., 2004.

[4] S. van Buuren and K. Groothuis-Oudshoorn, “mice: Multivariate imputation by chained equations in r,” *Journal of Statistical Software*, vol. 45, no. 3, pp. 1–67, 2011.

[5] U. Pujianto, A. P. Wibawa, M. I. Akbar *et al.*, “K-nearest neighbor (k-nn) based missing data imputation,” in *2019 5th International Conference on Science in Information Technology (ICSITech)*. IEEE, 2019, pp. 83–88.

[6] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, “A fast and elitist multiobjective genetic algorithm: Nsga-ii,” *IEEE transactions on evolutionary computation*, vol. 6, no. 2, pp. 182–197, 2002.

[7] E. Zitzler, M. Laumanns, and L. Thiele, “Spea2: Improving the strength pareto evolutionary algorithm,” *TIK-report*, vol. 103, 2001.

[8] D. Dua and C. Graff, “UCI machine learning repository,” 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[9] yzhu27, “Yzhu27/moss: A multi objective semi supervised explanation system.” [Online]. Available: https://github.com/yzhu27/MOSS

[10] H. Tu, G. Papadimitriou, M. Kiran, C. Wang, A. Mandal, E. Deelman, and T. Menzies, “Mining workflows for anomalous data transfers,” in *2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR)*. IEEE, 2021, pp. 1–12.

APPENDIX A
FEBRUARY STUDY

In our experiment, we implemented the February study as follows: we ran two SWAY models combined with imputation methods to obtain their optimal subset selections. We then used this subset as new input for optimizing the MICE model’s function, re-ran the interpolation model on the original dataset, and finally obtained better results than the first run. This demonstrates that we can better adjust the model in the next experiment based on the results obtained in the first experiment to achieve better results than the previous experiment. The results are shown in Table X. However, it should be noted that the performance improvement using this method will become marginal after multiple repetitions. Therefore, it is not possible to infinitely improve the model’s performance through multiple rounds of knowledge distillation.

	Defects-	Effort-	Kloc+	Months-
all	2007.0	252.0	47.5	21.4
sway	989.0	60.0	31.5	17.6
sway+MICE(1st)	626.0	39.8	20.0	15.4
sway+KNN(2nd)	477.0	40.0	18.9	15.1
top	109.0	10.8	3.5	7.8

TABLE X
FEBRUARY STUDY

APPENDIX B
ABLATION STUDY

In answering RQ2, we used Ablation Study. We first assumed that the baseline model+MICE and baseline model+KNN methods could achieve good results on noisy datasets. After conducting the experiment, we removed the two interpolation methods that were added and re-conducted the experiment. The baseline model performed worse than the above results. Therefore, we concluded that the introduced interpolation methods are important for noisy datasets. The experimental results can be found in Table V.