IEICE TRANS. INF. & SYST., VOL.E106–D, NO.2 FEBRUARY 2023

PAPER

Ensemble-Based Method for Correcting Global Explanation of Prediction Model*

Masaki HAMAMOTO[†a)], **Hiroyuki NAMBA**[†], ***Nonmembers***, ***and*** **Masashi EGI**[†], ***Member***

SUMMARY Explainable artificial intelligence (AI) technology enables us to quantitatively analyze the whole prediction logic of AI as a global explanation. However, unwanted relationships learned by AI due to data sparsity, high dimensionality, and noise are also visualized in the explanation, which deteriorates confidence in the AI. Thus, methods for correcting those unwanted relationships in explanation has been developed. However, since these methods are applicable only to differentiable machine learning (ML) models but not to non-differentiable models such as tree-based models, they are insufficient for covering a wide range of ML technology. Since these methods also require re-training of the model for correcting its explanation (i.e., in-processing method), they cannot be applied to black-box models provided by third parties. Therefore, we propose a method called ensemble-based explanation correction (EBEC) as a post-processing method for correcting the global explanation of a prediction model in a model-agnostic manner by using the Rashomon effect of statistics. We evaluated the performance of EBEC with three different tasks and analyzed its function in more detail. The evaluation results indicate that EBEC can correct global explanation of the model so that the explanation aligns with the domain knowledge given by the user while maintaining its accuracy. EBEC can be extended in various ways and combined with any method to improve correction performance since it is a post-processing-type correction method. Hence, EBEC would contribute to high-productivity ML modeling as a new type of explanation-correction method.

key words: *explainable AI, feature importance, global explanation, model ensembling, Rashomon effect*

1. Introduction

Shapley-value-based explainable artificial intelligence (AI) technology, such as SHAP [1], [2], enables us to quantitatively analyze the whole prediction logic of a machine learning (ML)-based black-box prediction model by revealing its global explanation. With this analysis, in addition to the evaluation of its prediction accuracy, domain experts can evaluate how much the global explanation of the prediction model fits their domain knowledge. However, pseudo or unwanted relationships learned by the model due to data sparsity, high dimensionality, and noise are also shown in the explanation, which deteriorates confidence in the prediction model.

There are two types of approaches to improve the global explanation of the prediction model. The first type is trial-and-error-based improvement that applies various modeling parameters to ML models and selects the best model. This type of approach tends to require a large amount of time to obtain a satisfying result. When the size of the training dataset is small, the global explanation of each trained model greatly varies depending on the applied modeling parameter, which makes obtaining a satisfying result more difficult. The phenomenon that many similar but different models can be built even from the same data set is known as the "Rashomon effect" of statistics [3], [4], and "which model should be selected from those many models" is discussed as a problem of statistical modeling [5], [6]. The other type of approach is regularization-based improvement that introduces an explanation-error term as a penalty term into the objective function of the ML model [7], [8]. With this type of approach, explicit target values of importance score (used for measuring explanation error) are provided by the user, and the optimum model is obtained by minimizing the objective function considering the explanation error. This type of approach solves the productivity issue with the trial-and-error-based approach. However, current methods of this approach can be applicable only to differentiable ML models and importance scores but not to non-differentiable models and scores such as tree-based models (e.g., XGBoost [9] and LightGBM [10]) and Shapley-value-based scores. Thus, such methods are not effective in covering a wide range of ML technology. These methods also need to be applied when the prediction model is trained. Therefore, they are not applicable to black-box prediction models provided by third parties.

We therefore propose the model-agnostic ensemble-based explanation correction method (EBEC) leveraging the Rashomon effect as a new regularization-based approach that works in a post-processing step of model development [11]. EBEC corrects the global explanation of a prediction model by ensembling many similar but different models so that its global explanation becomes close to a desired property given by the user by taking advantage of the Rashomon effect in ML. Since EBEC only adjusts ensemble coefficients, it can work for any type of ML model (including a prediction model provided by a third party) and importance score as long as it can provide their global explanations. We applied EBEC to three public datasets to evaluate its performance and analyzed in more detail how it functions in one of the datasets.

The contributions of this paper are as follows. (i) The evaluation results from public datasets indicate that

Manuscript received June 6, 2022.
Manuscript revised October 21, 2022.
Manuscript publicized November 15, 2022.

[†]The authors are with the Research & Development Group, Hitachi, Ltd., Kokubunji-shi, 185–8601 Japan.

*This paper is extended on the basis of [11], which appeared in proceedings of the 2021 IEEE Symposium Series on Computational Intelligence.

a) E-mail: masaki.hamamoto.qg@hitachi.com

DOI: 10.1587/transinf.2022EDP7095

EBEC performs well in three different tasks, i.e., physics-knowledge-based correction for accuracy, human-intuition-based correction for plausibility, and ethics-based correction for fairness. (ii) The analysis results from one of the public datasets indicate that EBEC can correct a targeted part of the global explanation. (iii) The results also indicate that the accuracy of the prediction model is not so sensitive to the weight values of EBEC when an appropriate function is applied for explanation correction.

The rest of this paper is organized as follows. In Sect. 2., we describe the novelty of EBEC by comparing it with related methods. In Sect. 3., we explain EBEC in more detail. In Sect. 4., we present the experimental settings, results, and discussions. Finally, we conclude this paper in Sect. 5.

2. Related Work

Qualitative constraints: Many methods of controlling an ML model's behavior by applying qualitative constraints (QCs), such as monotonic and synergic influences, to the model have been proposed [12], [13]. A QC is useful, especially when humans have rich knowledge in the field while there are less available data to build a prediction model. Most methods focus on improving the relationship between the input and output of a model that is visualized in an individual conditional expectation plot [14] but not in importance scores or global explanation of the model. Since explainable AI technology is becoming common, a correction method of importance scores as well as global explanation should be provided.

Explanation correction: There are a few explanation-correction methods that correct a model's explanation so that it aligns with domain knowledge provided in a certain format by the user. Right for the Right Reasons (RRR) is the first method for correcting model's explanation [15]. RRR uses input gradients as its importance score and improves generalization performance of the prediction model by lowering the gradients of input features that should not affect the output. The method has been enhanced as Right for the Right Concept so that it can correct concept-level explanations by extending its regularization target from input space to the representation space of the model [16]. Attention branch network (ABN) uses class attribution mapping as the attention mechanism for explanation correction [7], [17]. ABN works with a convolutional neural network for visual input applications to improve its accuracy by taking into account preferable attention maps given by humans. Contextual decomposition explanation penalization (CDEP) uses a CD score as its importance score [8]. CDEP can be applied to any deep neural network model with arbitrary architectures and help users correct those models' explanations. Expected gradients (EG) uses integrated gradients as its importance score and improves model's explainability and robustness to noise by incorporating higher-level expected properties of explanations, such as smoothness and sparsity, into its regularization term for optimization [18]. These methods enable users to correct errors in terms of explanation by directly regularizing the explanations provided from an ML model when the model seems to have incorrectly assigned importance to certain features. However, the methods are applicable only to differentiable ML models and explanations (or importance scores); hence, non-differentiable models, such as tree-based models and Shapley-value-based explanation, are not included. Due to the necessity of retraining the model for correction, these methods are also not applicable to models provided by third parties. Current methods focus on correcting local explanation. Although a global explanation is a set of local explanations, from the aspect of correcting the global explanation of a prediction model, it is necessary to consider the preferable property of the global explanation as well as its local explanation. Therefore, EBEC can correct global explanations.

ML research on Rashomon effect: There have been a few studies that analyzed the Rashomon effect from the theoretical aspects and used its insights to find the best prediction model from the model space that may contain different yet approximately-equally accurate models that might obey various constraints such as interpretability, fairness, or monotonicity [5], [6], [19], [20]. These studies give us much insight in terms of the Rashomon effect in ML. There is another framework, although it does not refer to the Rashom effect, that applies many different hyper-parameters to an ML model to create many different prediction models and finds the most interpretable model from among them [21]. Thus, current methods are mostly focused on finding the simplest or most interpretable prediction model from among the different models obtained from the model space. Our approach is clearly different from the above approaches since EBEC is focused on creating a model that gives a simple explanation by combining the different models obtained from the model space by leveraging the Rashomon effect.

3. Ensemble-Based Explanation Correction

The concept of EBEC is shown in Fig. 1. There are many local minima in the objective function space, and prediction models at each local minimum (denoted as $f_1(\boldsymbol{X})$, $f_2(\boldsymbol{X})$, $\cdots$) have different global explanations for their training data. The key idea with EBEC is not to search the model space for a desired model but create a model that aligns with prior domain knowledge by ensembling various models on the basis of our hypothetical idea of "every model may not be correct but still not be wrong." To do so, we introduce an explanation-error term as a penalty term into the objective function of an ensembled model and obtain ensemble coefficients by minimizing the function. We use the Shapley value as an importance score of features, but applicable importance score for EBEC is not limited to the Shapley value as long as it can provide a global explanation of the model.

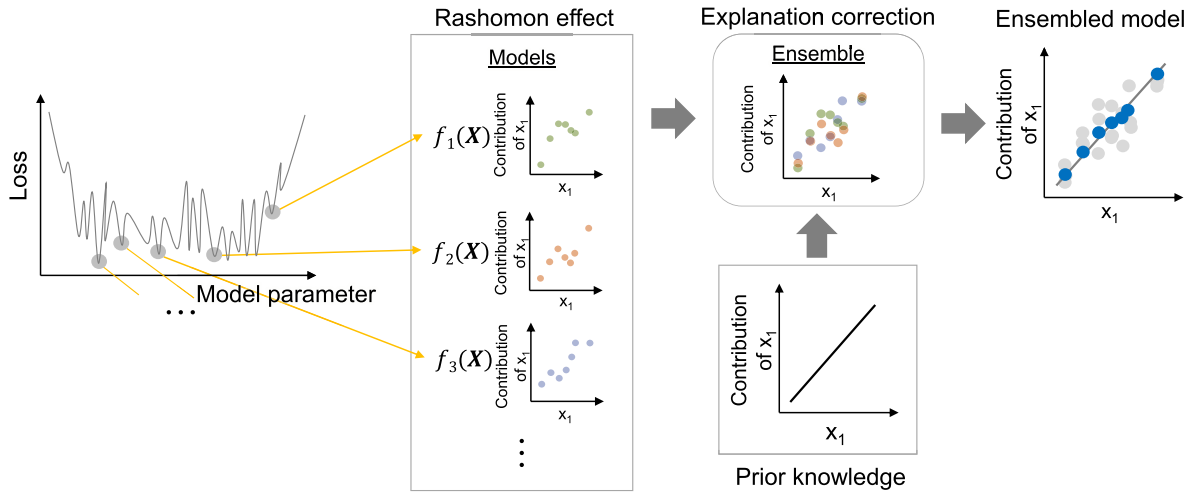


Fig. 1 Concept of ensemble-based explanation correction (EBEC)

3.1 Objective Function

Given N, M, and n as the number of data samples in the training dataset, input features used in a model, and models ensembled, respectively, and $\boldsymbol{a} = [a_1, a_2, \cdots, a_n]^\top$ as a vector of ensemble coefficients, the equations to find the optimum $\widehat{\boldsymbol{a}}$ that minimizes the objective function can be described as

$$\widehat{\boldsymbol{a}} = \underset{\boldsymbol{a}}{\operatorname{argmin}} \left(\mathrm{Loss}_{\mathrm{pred}} + \mathrm{Loss}_{\mathrm{expl}}\right), \tag{1}$$

$$\mathrm{Loss}_{\mathrm{pred}} = \sum_{i}^{N} \left(\bar{G}_i - Y_i\right)^2, \tag{2}$$

$$\begin{aligned}\mathrm{Loss}_{\mathrm{expl}} = &\sum_{i}^{N}\sum_{f}^{M} \lambda_{\mathrm{pref}_f}\left(\bar{R}_{i,f} - Z_{\mathrm{pref}_{i,f}}\right)^2 \\ &+ \sum_{i}^{N}\sum_{f}^{M} \lambda_{\mathrm{base}_f}\left(\bar{R}_{i,f} - Z_{\mathrm{base}i,f}\right)^2,\end{aligned} \tag{3}$$

$$\bar{G}_i = \sum_{k}^{n} a_k G_k\left(\boldsymbol{X}_i\right), \tag{4}$$

$$\bar{\boldsymbol{R}}_i = \sum_{k}^{n} a_k \boldsymbol{R}_k\left(\boldsymbol{X}_i\right). \tag{5}$$

Equation (1) describes the objective function of EBEC, where $\mathrm{Loss}_{\mathrm{pred}}$ is a standard prediction-error term for finding the optimum $\boldsymbol{a}$ from the aspect of prediction accuracy, and $\mathrm{Loss}_{\mathrm{expl}}$ is an explanation-error term for finding the optimum $\boldsymbol{a}$ from the aspect of explanation fitness to prior domain knowledge. Here, $\mathrm{Loss}_{\mathrm{expl}}$ is the term newly added to the objective function for EBEC. Equation (2) describes the details of $\mathrm{Loss}_{\mathrm{pred}}$. Given i as the index number of data samples in the training dataset, $\bar{G}_i$ and Y_i denote the predicted value and its expected value of the ith data sample, respectively. In (3), the details of $\mathrm{Loss}_{\mathrm{expl}}$ are described, where $\bar{R}_{i,f}$ is the importance score of the input feature f for the ith data sample. Given $\bar{\boldsymbol{R}}_i = [\bar{R}_{i,1}, \bar{R}_{i,2}, \cdots, \bar{R}_{i,M}]^\top$ as a vector of the importance score of the ith data sample, $\bar{R}_{i,f}$ can be obtained through (5). $Z_{\mathrm{pref}_{i,f}}$ and $Z_{\mathrm{base}i,f}$ are preferable and base importance scores (or explanations) of feature f for the ith data sample, respectively, and are provided by the user who develops the prediction model. The $Z_{\mathrm{pref}_{i,f}}$ is given as prior knowledge shown in Fig. 1, and any user-defined value can be applied to $Z_{\mathrm{pref}_{i,f}}$. The $Z_{\mathrm{base}i,f}$ is a local explanation given by the base model defined by the user. Any model or just an equally weighted (or simple) ensembled model can be selected as the base model, the global explanation of which is to be corrected. In (3), therefore, the first term is for fitting the prior knowledge, while the second term is for trying to keep the explanation of the base model as is. These two terms are weighted by the hyper-parameters $\lambda_{\mathrm{pref}_f} \in [0, \infty)$ and $\lambda_{\mathrm{base}_f} \in [0, \infty)$ for each f, respectively. Since available prior knowledge is quite limited in most cases, the ensembled model can easily overfit to Y_i and $Z_{\mathrm{pref}_{i,f}}$ and deteriorate its global explanation when n is large. Therefore, by filling the blank of $Z_{\mathrm{pref}_{i,f}}$ with $Z_{\mathrm{base}i,f}$, EBEC can fit its global explanation to the preferable one while maintaining the property of the base model. With this policy, $\mathrm{Loss}_{\mathrm{expl}}$ in (3) can be simply expressed as

$$\mathrm{Loss}_{\mathrm{expl}} = \sum_{i}^{N}\sum_{f}^{M} \lambda_f \left(\bar{R}_{i,f} - Z_{i,f}\right)^2, \tag{6}$$

$$Z_{i,f} = \begin{cases} Z_{\mathrm{pref}_{i,f}} & (\mathrm{PK}(f) = 1), \\ Z_{\mathrm{base}i,f} & (\mathrm{PK}(f) = 0), \end{cases} \tag{7}$$

$$\lambda_f = \begin{cases} \lambda_{\mathrm{pref}_f} & (\mathrm{PK}(f) = 1), \\ \lambda_{\mathrm{base}_f} & (\mathrm{PK}(f) = 0), \end{cases} \tag{8}$$

where $Z_{i,f}$ is a target importance score (or explanation) of feature f for the ith data sample, and $\lambda_f \in [0, \infty)$ is a

Algorithm 1 Ensemble-based explanation correction (EBEC)
Input: $\boldsymbol{D}(\boldsymbol{X},\ Y)$, p_{aug}, p_{range}, $\boldsymbol{\lambda}_{\text{pref}}$, $\boldsymbol{\lambda}_{\text{base}}$
Output: $\widehat{\boldsymbol{a}}$
1: **for** $k = 1$ to n **do**
2: $\quad \boldsymbol{D}_{\text{aug}} = \text{bootstrap}(\boldsymbol{D},\ p_{\text{aug}})$
3: $\quad p = \text{random}(p_{\text{range}})$
4: $\quad G(k,\ \boldsymbol{X}) = \text{generateModel}(\boldsymbol{D}_{\text{aug}},\ p)$
5: $\quad \boldsymbol{R}(k,\ \boldsymbol{X}) = \text{computeGlobalExpl}(G(k,\ \boldsymbol{X}),\ \boldsymbol{D})$
6: **end for**
7: $\boldsymbol{Z}_{\text{base}}(\boldsymbol{X}) = 1/n \sum_k \boldsymbol{R}(k,\ \boldsymbol{X})$
8: $\boldsymbol{Z}_{\text{pref}}(\boldsymbol{X}) = \text{computePrefExpl}(\boldsymbol{Z}_{\text{base}},\ \boldsymbol{X})$
9: $\widehat{\boldsymbol{a}} = \text{argminObjFunc}(\boldsymbol{D},\ G,\ \boldsymbol{R},\ \boldsymbol{Z}_{\text{pref}},\ \boldsymbol{Z}_{\text{base}},\ \boldsymbol{\lambda}_{\text{pref}},\ \boldsymbol{\lambda}_{\text{base}})$

hyper-parameter for weighting the loss value in terms of explanation error for feature f. PK(f) is the availability of prior knowledge for feature f and takes 1 when $Z_{\text{pref}_{i,f}}$ is available, and 0 otherwise. Local explanation for correction ($Z_{\text{pref}_{i,f}}$) and preservation ($Z_{\text{base}i,f}$) can be described as just a target explanation $Z_{i,f}$ as shown in (6). However, their roles ($Z_{\text{pref}_{i,f}}$ and $Z_{\text{base}i,f}$) as well as methods for obtaining their values are clearly different. To easily explain these differences, therefore, we use (3), which separately describes the two error terms for correction and preservation of global explanation, in this paper.

Finally, the $\bar{G}_i$ and $\bar{\boldsymbol{R}}_i$ are obtained from (4) and (5), respectively, where $G_k(\boldsymbol{X}_i)$ and $\boldsymbol{R}_k(\boldsymbol{X}_i)$ denote the prediction value and importance-score vector of the kth model for the input feature $\boldsymbol{X}_i = [x_{i,1},\ x_{i,2},\ \cdots,\ x_{i,M}]^\top$, that is, the input feature vector of the ith data sample. Note that, for (1), scale adjustment between $\text{Loss}_{\text{pred}}$ and $\text{Loss}_{\text{expl}}$ is not necessary. The unit dimension of the Shapley value is the same as that of the predicted value $\bar{G}_i$, and the summation of Shapley values for each input feature f corresponds to $\bar{G}_i - \beta$ (i.e., $\bar{G}_i - \beta = \sum_f \bar{R}_{i,f}$), where β is a constant baseline value of the Shapley value [1]. As long as the Shapley value is used as the importance score of EBEC, $\text{Loss}_{\text{pred}}$ and $\text{Loss}_{\text{expl}}$ should be well balanced. Therefore, the scale adjustment for $\text{Loss}_{\text{pred}}$ and $\text{Loss}_{\text{expl}}$ has not been applied to (1).

3.2 Algorithm

Algorithm 1 describes a pseudo code of EBEC used for our experiments. To build many different models, bootstrap sampling and random hyper-parameters are applied to the training dataset and ML model, respectively, in Algorithm 1. As the input parameters, $\boldsymbol{D}(\boldsymbol{X},\ Y)$ is the training dataset, where $\boldsymbol{X}$ denotes the input feature vector, Y denotes the target feature, and p_{aug} is the augmentation ratio of bootstrap sampling. When p_{aug} is 2, for example, $2N$ data samples are randomly selected from the original training dataset $\boldsymbol{D}$. This augmentation is useful for avoiding accuracy deterioration caused by bootstrap sampling when the size of D is small. The p_{range} is the minimum and maximum values of hyper-parameters used for ML modeling, and $\boldsymbol{\lambda}_{\text{pref}}$ and $\boldsymbol{\lambda}_{\text{base}}$ are weight-value vectors used in the objective function of EBEC, as described in (3). The output of this algorithm is the optimum ensemble-coefficient vector $\widehat{\boldsymbol{a}}$. A detailed description of each step in Algorithm 1 is as follows.

Step 1: in line 2, sub dataset $\boldsymbol{D}_{\text{aug}}$ is obtained by applying bootstrap sampling to $\boldsymbol{D}$ with p_{aug}.
Step 2: in line 3, the hyper-parameters of ML model p are randomly determined within the range defined by p_{range}.
Step 3: in line 4, a model G (G_k in (4)) is created on the basis of $\boldsymbol{D}_{\text{aug}}$ and p. Although we used XGBoost as the ML model in our experiments, any ML model can be applied.
Step 4: in line 5, a global explanation $\boldsymbol{R}$ ($\boldsymbol{R}_k$ in (5)) is computed on the basis of $\boldsymbol{D}$ and $G(k, \boldsymbol{X})$.
Step 5: steps 1 through 4 are repeated n times, and n Gs and n $\boldsymbol{R}$s are created.
Step 6: in line 7, base explanation $\boldsymbol{Z}_{\text{base}}$ (a set of $Z_{\text{base}i,f}$ in (3)) is defined. Although, the average global explanation of all models is set as $\boldsymbol{Z}_{\text{base}}$ in this algorithm, any particular model's global explanation can be used.
Step 7: in line 8, preferable explanation $\boldsymbol{Z}_{\text{pref}}$ (a set of $Z_{\text{pref}_{i,f}}$ in (3)) is computed on the basis of $\boldsymbol{Z}_{\text{base}}$ and $\boldsymbol{X}$. Practically, it might be difficult to directly provide exact values of the preferable global explanation even if the user is an expert in the domain. However, providing its qualitative properties, such as linearity and monotonicity, is much easier. Therefore, computePrefExpl() is a user-defined function that differs depending on its application. For example, when x_1 is expected to linearly contribute to the output, a linear function $f_{\text{linear}}()$ that fits the base model's global explanation in terms of x_1 is computed, then $\boldsymbol{Z}_{\text{pref}}$ is obtained as $f_{\text{linear}}(x_1)$.
Step 8: in line 9, $\widehat{\boldsymbol{a}}$ is obtained by solving the objective function described in (1). We used the conjugate gradient method for solving this minimization problem.

4. Experiments and Analysis

We conducted two types of experiments to evaluate EBEC. One involved using four public datasets to evaluate the performance of EBEC in three different tasks. The other one involved using one of those datasets to analyze in more detail how EBEC functions.

4.1 Experiment 1: Evaluation with Public Datasets

We applied EBEC to three different tasks using four different public datasets to evaluate its performance. The basic experimental settings applied to these tasks are listed in Table 1. Using the training datasets, we created 200 models by using XGBoost [9] and computed the base model's global explanation by using Tree SHAP [2]. The other applied hyper-parameters p_{aug} and p_{range} are listed in Table 1. When the available training dataset is small, the prediction model often has distortion in its global explanation due to pseudo-correlations in the dataset. Under this condition, the benefit of EBEC can be easily observed. To see how EBEC works, we used relatively small training datasets in this experiment.

Task 1: Physics-knowledge-based correction for accuracy with Concrete dataset

The amount of data obtained through physical exper-

Table 1 Basic parameters used in Experiment 1

Parameters	Details
ML model	XGBoost [9]
Importance score	Shapley value (Tree SHAP [2])
n	200
p_{aug}	2
p_{range}	max_depth: 1 to 20, n_estimators: 10 to 1000, learning_rate: 0.01 to 0.50

iments tends to be small because of their high cost, but sometimes humans have rich knowledge about the domain. Assuming such a case, we applied EBEC to a prediction model to improve its accuracy by directly adding domain knowledge to it. The experimental conditions used in task 1 are listed in Table 2. We used the Concrete dataset [22] as a physics-based experimental dataset and the "compressive strength" of concrete was predicted as the target feature. We split the dataset into 10% training data and 90% test data and used a simple (equally weighted) ensembled model as the base model. We also used the main effect of the Shapley values as the score of the preferable and base explanation (i.e., $\boldsymbol{Z}_{\text{pref}}$ and $\boldsymbol{Z}_{\text{base}}$) because the plot of the main effect tends to be simple and have less variation compared with that of the total effect (i.e., Shapley value itself) that may include the interaction effect among input features as well as the feature's main effect. Thus, the main effect is suitable for $\boldsymbol{Z}_{\text{pref}}$ and $\boldsymbol{Z}_{\text{base}}$.

We set the Lyse's equation [23], which describes a property that cement density linearly affects compressive strength of concrete, as the prior knowledge. Thus, the main effect of x_1 (denoted as $\phi_{1,1}$) was approximated by a linear function for $\boldsymbol{Z}_{\text{pref}}$. The $\boldsymbol{\lambda}_{\text{pref}}$ for x_1 was set as 6 and $\boldsymbol{\lambda}_{\text{base}}$ for those other than x_1 were set as 3 in this task.

A comparison of dependence plots of the main effect of the Shapley value (as the global explanations) for x_1 is shown in Figs. 2 (a), (b), and (c). The preferable explanation (b) was given as a function obtained by fitting the dependence plot of (a) to a linear function. With the preferable explanation (b), the dependence plot of the corrected model (c) became much more linear compared with that of the base model (a). Moreover, a comparison of the root-mean-squared error (RMSE) and coefficient of determination (R2) scores among the models is shown in Table 3, where the mean value and standard deviation over ten runs are described. The results indicate that the RMSE of the base model decreased by 3.5% on average as a result of the correction using EBEC.

The t-values of the difference between the scores of corrected and base models are also shown in Table 3. In this task, the t-value was larger than 1.833, which is the value for 95% confidence in the t-distribution. Since the feature x_1 is quite influential for the prediction, the prediction performance improved by correcting the global explanation for x_1 through applying a well-studied knowledge to the explanation. Therefore, physics-knowledge-based correction can improve the generalization performance of the prediction model when the model could not capture the right property of the global explanation from the limited training data.

Task 2: Human-intuition-based correction for plausibility with Boston dataset and Breast cancer dataset

The dataset for predicting a target feature that is not simply based on physics but strongly affected by human senses, customs, or experiences is sometimes too small to build a prediction model with its global explanation fitting the user's intuition. Therefore, in task 2, we applied EBEC to two types of prediction models, which are regression and classification models, to improve its plausibility by reflecting qualitative assumption on the basis of the user's intuition in the model. The experimental parameters used in this task are listed in Table 2.

First, for a case of regression problem as task 2-1, we used the Boston dataset [24] as a non-physics-based dataset and the price of a house "PRICE" was set as the target feature. The dataset was split into 20% training data and 80% test data. Using the training dataset, we created 200 models and a base model by ensembling them with equal weights then obtained dependence plots that visualize the main effect of the Shapley values for all input features. From these plots, we found that the property of the most influential feature, LSTAT (% lower status of the population), had some amount of distortion in its curve, as shown in Fig. 2 (d). We put a qualitative assumption that "LSTAT would affect PRICE in a continuous manner" and chose $\alpha \log x + \beta$ as a simple function to approximate the curve of LSTAT (denoted as x_{13}). Therefore, we obtained $\boldsymbol{Z}_{\text{pref}}$, as shown in Fig. 2 (e). We did not have much confidence in the preferable explanation $\boldsymbol{Z}_{\text{pref}}$ compared with that of task 1; thus, $\boldsymbol{\lambda}_{\text{pref}}$ of x_{13} was set as 1 (as a relatively small value) to preserve the properties of the base model as much as possible. A comparison of dependence plots for the main effect of the Shapley value of x_{13} is shown in Figs. 2 (d), (e), and (f). The dependence plot of the corrected model (f) was smoother compared with that of the base model (d). In terms of accuracy comparison, the RMSE of the corrected model improved by 2.8% (with the t-value of 3.514) on average rather than deteriorated, as shown in Table 3.

We did not expect that the correction would improve prediction performance in this task. However, smoothness can be considered as one of high-level expected properties of explanations and it was reported that emphasizing the smoothness property of explanations improved prediction performance in EG [18]. Thus, improving the smoothness property of explanations can have a positive impact on prediction performance. In this task, therefore, the smoothness property was fortunately matched with the global explanation for x_{13}, which is the most influential feature, and this led the improvement of prediction performance.

Second, for a case of a classification problem as task 2-2, we used the Breast cancer dataset [25], in which the problem is to distinguish malignant (cancerous) from benign (non-cancerous) examples. The dataset was split into 10% training data and 90% test data. Using the training dataset, we created 200 models and a base model by ensem-

Table 2 Main parameters used in Experiment 1

Parameters	Task 1	Task 2-1	Task 2-2	Task 3
Data set	Concrete [22]	Boston [24]	Breast cancer [25]	Adult [25]
Training data	103 samples (10%)	102 samples (20%)	57 samples (10%)	8,997 samples (30%)
Test data	927 samples (90%)	404 samples (80%)	512 samples (90%)	20,994 samples (70%)
Input features	8 features	13 features	30 features	14 features
Base model	Simple ensemble	Simple ensemble	Simple ensemble	Best model
Prior knowledge	Cement (x_1) linearly affects compressive strength (Y) [23]	LSTAT (x_{13}) would affect PRICE (Y) in continuous manner	For Worst texture (x_{22}), the slope connecting the two-value levels would be linear	Race (x_9) and sex (x_{10}) should not affect income class (Y)
$\boldsymbol{Z}_{\text{pref}}$	$\phi_{1,1} = \alpha x_1 + \beta$ (for main effect of Shapley value of x_1)	$\phi_{13,13} = \alpha \log x_{13} + \beta$ (for main effect of Shapley value of x_{13})	if $(20 < x_{22} < 32)$: $\phi_{22,22} = \alpha x_{22} + \beta$ else: $\phi_{22,22} = Z_{\text{base}_{x_{22}}}$ (for main effect of Shapley value of x_{22})	$\phi_9(\boldsymbol{X}) = \phi_{10}(\boldsymbol{X}) = 0$ (for Shapley values of x_9 and x_{10})
$\boldsymbol{Z}_{\text{base}}$	Main effect of Shapley value for each input feature	Main effect of Shapley value for each input feature	Main effect of Shapley value for each input feature	Shapley value for each input feature
$\boldsymbol{\lambda}_{\text{pref}}$	6 for x_1, 0 for others	1 for x_{13}, 0 for others	1 for x_{22}, 0 for others	500 for x_9 and x_{10}, 0 for others
$\boldsymbol{\lambda}_{\text{base}}$	0 for x_1, 3 for others	0 for x_{13}, 3 for others	0 for x_{22}, 3 for others	0 for x_9 and x_{10}, 3 for others

Table 3 Accuracy comparison between base and corrected models (over 10 runs) in Experiment 1

Task	Accuracy criteria	Base model	Corrected model (EBEC)	—t-value| (Corrected - Base)
Task 1: Concrete (Regression)	RMSE	7.419±0.259	7.154±0.246	3.550
	R2	0.802±0.013	0.816±0.012	3.570
Task 2-1: Boston (Regression)	RMSE	4.292±0.455	4.170±0.439	3.514
	R2	0.777±0.042	0.789±0.038	3.348
Task 2-2: Breast cancer (Classification)	Accuracy	0.930±0.013	0.931±0.014	1.342
Task 3: Adult (Classification)	Accuracy	0.862±0.002	0.859±0.002	3.237
	Imp. score (race)	0.044±0.016	0.009±0.002	6.461
	Imp. score (sex)	0.151±0.070	0.015±0.003	6.010

bling them with equal weights then obtained dependence plots that visualize the main effect of the Shapley values for all input features. From these plots, we found that global explanations for most features have two-value levels connected with a steep slope, as shown in Fig. 2 (g). For the property of "Worst texture" (denoted as x_{22}), which is a relatively influential feature, we intuitively assumed that "the slope connecting the two-value levels would be linear" with a small confidence on the basis of its shape. Therefore, $\boldsymbol{\lambda}_{\text{pref}}$ of x_{22} was set as 1 (as same as the regression case) to preserve the properties of the base model as much as possible. A comparison of dependence plots for the main effect of the Shapley value of x_{22} is shown in Figs. 2 (g), (h), and (i). In the dependence plot of the corrected model (i), the linearity of the slope improved compared with that of the base model (g). In terms of accuracy comparison, the accuracy of the corrected model slightly improved, but it was not statistically significant since its t-value for ten runs was 1.342, as shown in Table 3.

Task 3: Ethics-based correction for fairness with Adult dataset

From the aspect of ethics in ML, there is a case in which the user wants to remove unwanted bias from the prediction model that contains sensitive attributes in its input features. This correction is known as "de-bias", and there have been studies on improving the fairness in ML [26]–[28]. Therefore, in task 3, we applied EBEC to a prediction model to improve its fairness by reflecting ethics-based constrains in the model and suppressing the effect of sensitive attributes in it.

The experimental parameters used in this task are listed in Table 2. We used the Adult dataset [25], which involves predicting personal annual income levels as above or below $50,000 based on personal attributes. The dataset was split into 30% training data and 70% test data because there were enough data to build a prediction model with high accuracy, and evaluation with a larger amount of test data is better than that with a small amount as long as there is enough training data. In this task, we chose the most accurate model among the created 200 models as the base model to evaluate how much the bias correction using EBEC would deteriorate its accuracy. As the prior knowledge, "race (denoted as x_9) and sex (denoted as x_{10}) should not affect the prediction of personal income" was applied. The Shapley value was used as $\boldsymbol{Z}_{\text{pref}}$ and $\boldsymbol{Z}_{\text{base}}$ instead of its main effect component because the effects of x_9 and x_{10} including any interaction effect components should be suppressed to zero in this task. Finally, $\boldsymbol{\lambda}_{\text{pref}}$ for x_9 and x_{10} were set as 500 (as strong constraints), while $\boldsymbol{\lambda}_{\text{base}}$ for those others than x_9 and x_{10} were set as 3 in the same manner as the other tasks. A comparison of summary plots that show the mean absolute Shapley value for each feature (as the importance score used in this evaluation) are shown in Figs. 2 (j), (k), and (l). The im-

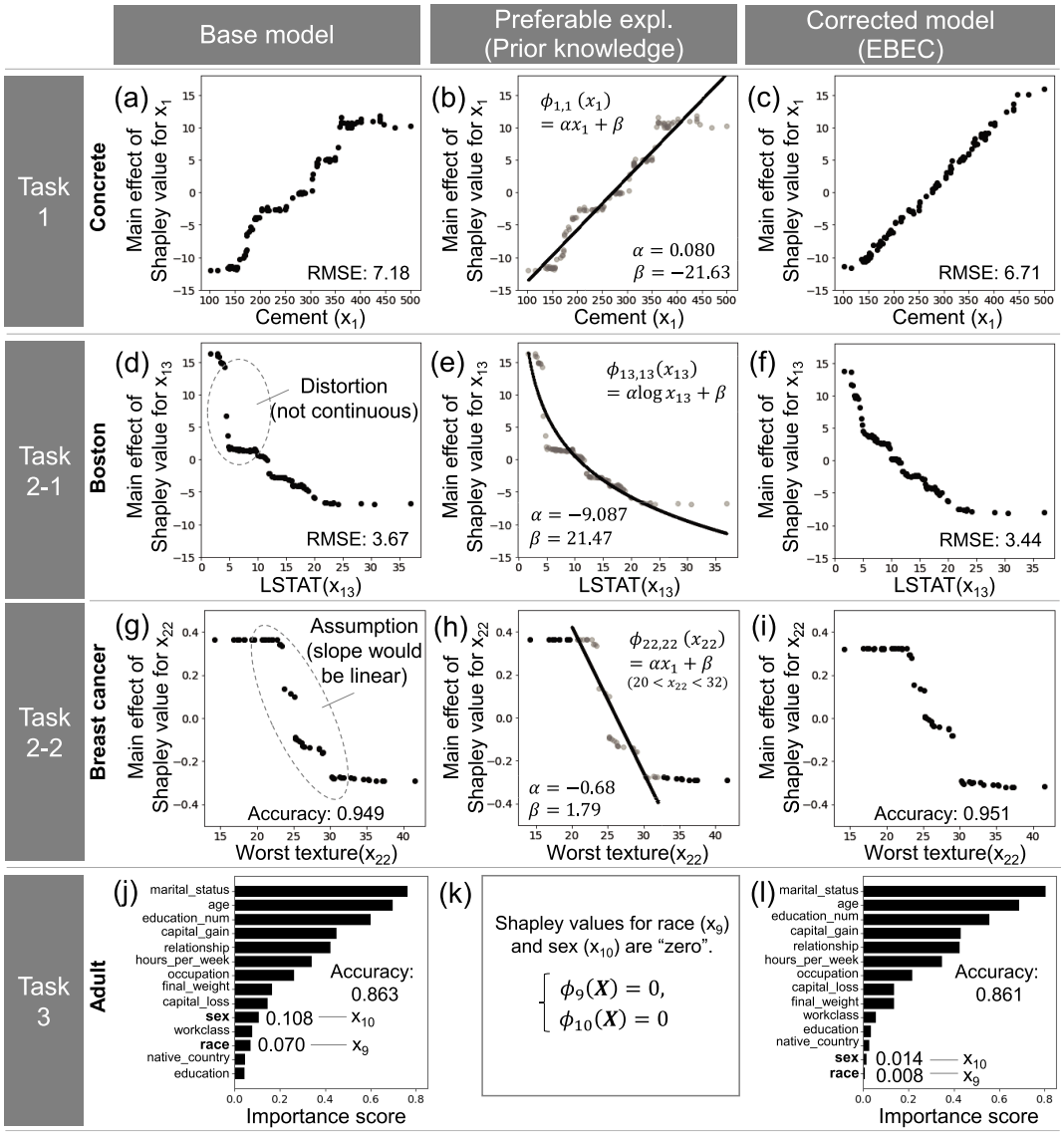


Fig. 2 Results from Experiment 1. (a) and (c) Dependence plots of main effect of Shapley values for x_1 in base and corrected models, respectively. (b) Preferable explanation for task 1. In same manner, (d), (e), (f) and (g), (h), (i) are those for regression and classification problems in task 2, respectively. (j) and (k) Summery plots of Shapely-value-based importance scores for base and corrected models, respectively. (l) Preferable explanation for task 3.

portance scores of the base model for x_9 and x_{10} (shown in (j)) were 0.070 and 0.108 and those of the corrected model (shown in (l)) were 0.008 and 0.014, respectively. Therefore, the results indicate that EBEC successfully suppressed the effect of sensitive attributes x_9 and x_{10} by more than 85% while maintaining its accuracy. The mean value and standard deviation of accuracy and importance scores over ten runs are listed in Table 3, which shows that the importance scores of the sensitive attributes were suppressed by 82% on average with 0.4% accuracy deterioration (with the t-value of 3.237) as a result of the correction using EBEC. Thus, we confirmed that EBEC can de-bias a prediction model in terms of feature attributions to improve its fairness in prediction while maintaining high accuracy.

4.2 Experiment 2: Analysis with Concrete Dataset

To analyze how EBEC functions in more detail, we conducted three evaluations using the Concrete dataset for observing the Rashomon effect in global explanation, effect of the λ_{base} on explanation correction, and performance sensitivity to the parameters of λ_{pref} and λ_{base}. The experimental settings were the same as in task 1.

Observation of Rashomon effect on global explanation: We first analyzed the Rashomon effect in task 1. The dependence plots of the main effect of the Shapley values (on all features) for 200 models are shown in Fig. 3. The gray

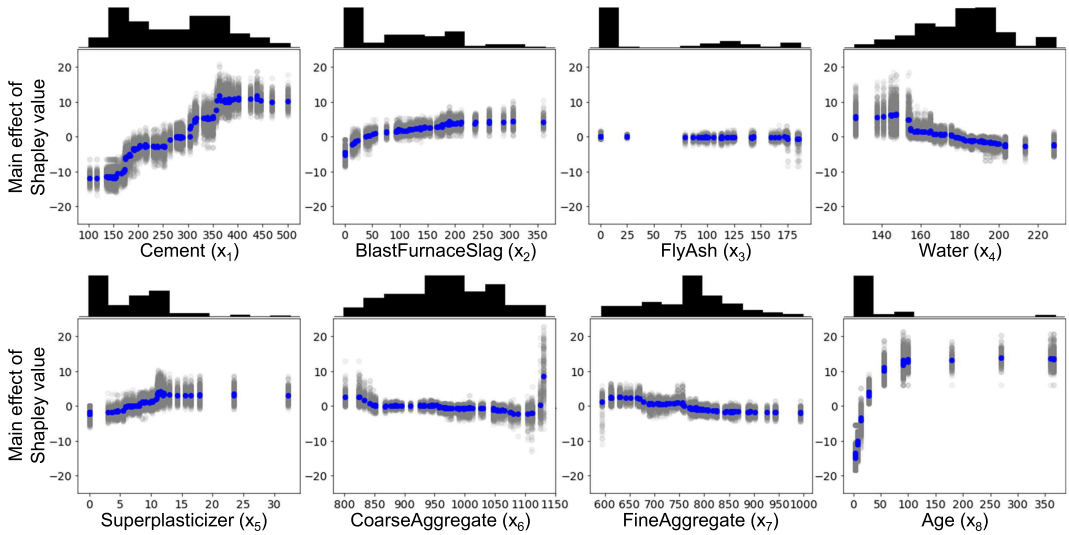


Fig. 3 Dependence plots of main effect of Shapley values for 200 models (Concrete dataset)

points are each model's importance score and the blue points are their mean values. The bar charts above the plots show the histogram of the training data samples for each feature. In these dependence plots, the variation of gray points is caused by the Rashomon effect, and we can see that the variation is relatively large in the region where the number of samples is small. With this variation and diversity of the models, EBEC can correct the explanation of a prediction model.

Effect of $\boldsymbol{\lambda}_{\text{base}}$ on explanation correction: To analyze the effect of the weight-value vector $\boldsymbol{\lambda}_{\text{base}}$ on the explanation correction for each feature, we compared the explanation errors in terms of the main effect of the Sharpley value for each feature obtained by applying zero and three to $\boldsymbol{\lambda}_{\text{base}}$ (denoted as $\lambda_{\text{base}} = 0$ and $\lambda_{\text{base}} = 3$, respectively). The experimental results are shown in Fig. 4. Since the $\boldsymbol{\lambda}_{\text{base}}$ is 0 for x_1 and the $\boldsymbol{\lambda}_{\text{pref}}$ is 0 for other than x_1, we denoted $\boldsymbol{\lambda}_{\text{base}}$ and $\boldsymbol{\lambda}_{\text{pref}}$ as simple scalar weight parameters of λ_{base} and λ_{pref}, respectively. These lines show the mean explanation errors (between the corrected model's explanation and base model's or preferable explanation) of each input feature for the training dataset over ten runs. By comparing Figs. 4 (a) and (b), we can see that the explanation errors of x_2 to x_8 with $\lambda_{\text{base}} = 3$ (shown in (b)) increase much less as λ_{pref} increases than that with $\lambda_{\text{base}} = 0$ (shown in (a)), while the explanation error for x_1 with $\lambda_{\text{base}} = 3$ decreases more rapidly than that with $\lambda_{\text{base}} = 0$. A visual comparison of the dependence plots observed for ten runs is presented in Fig. 5. It compares the global explanation for each feature among the base model, corrected model with $\lambda_{\text{base}} = 0$, and corrected model with $\lambda_{\text{base}} = 3$, which are shown as black, red, and green points, respectively. The global explanation in the base model was mostly preserved in the corrected model with $\lambda_{\text{base}} = 3$ except for that of x_1, while the corrected model with $\lambda_{\text{base}} = 0$ could not preserve the base model's global explanation. Therefore, we confirmed that the weight-value vector $\boldsymbol{\lambda}_{\text{base}}$ works as an anchor to preserve the base model's global explanation, as expected, and EBEC can selectively correct a targeted part of the base model's global explanation.

Performance sensitivity to $\boldsymbol{\lambda}_{\text{pref}}$ and $\boldsymbol{\lambda}_{\text{base}}$: To analyze the sensitivity of the prediction performance to the weight parameters of EBEC, we conducted experiments that applied several values to λ_{base} and λ_{pref} and observed the prediction errors of the corrected model for the test dataset. Applying a larger value to λ_{base} does not affect the prediction model because the role of the parameter is to maintain the property of the base model as is. On the other hand, λ_{pref} changes the property of the base model; thus, the purpose of the analysis is to observe how parameter λ_{pref} affects the prediction model as its value changes. Therefore, we assigned integer values around 1 to λ_{base} as $\lambda_{\text{base}} \in \{0, 1, 2, 3\}$ and a wide range of values to λ_{pref} as $\lambda_{\text{pref}} \in \{0, 0.25, 0.5, 1, 2, 4, 8, 16, 32\}$ to observe the effect of λ_{pref} when the value is small as well as when it is very large.

The results are shown in Fig. 6, where the prediction errors are normalized by that of the base model and the mean value of the prediction errors over ten runs are presented. When both the λ_{pref} and λ_{base} are zero (shown as a orange dot), performance deteriorates because of the overfitting to the training dataset. When the $\lambda_{\text{pref}} > 0$, the performance improves, as we expected. However, even when the λ_{pref} is very large, such as 32, the performance does not deteriorate much for every λ_{base}. Therefore, the results indicate that the performance was not so sensitive to the λ_{pref} as well as λ_{base}. We thought it would be because of the preferable explanation ($\boldsymbol{Z}_{\text{pref}}$) used in the experiment. The larger the λ_{pref}, the closer the targeted feature's global explanation to the preferable explanation provided by the user. Thus, when the preferable explanation is relatively correct, the side effects of λ_{pref} would be still small even if a large value is assigned to it. To confirm this hypothesis, we applied dif-

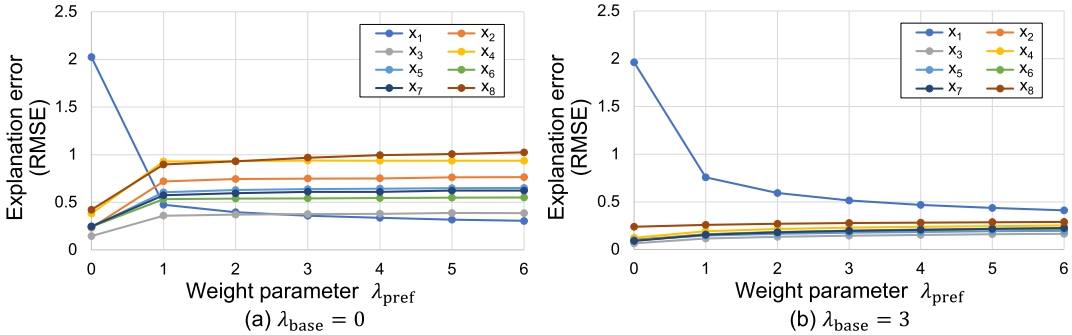


Fig. 4 Explanation error vs. weight parameter of x_1 (λ_{pref}) for (a) $\lambda_{\text{base}} = 0$ and (b) $\lambda_{\text{base}} = 3$ (averaged over 10 runs)

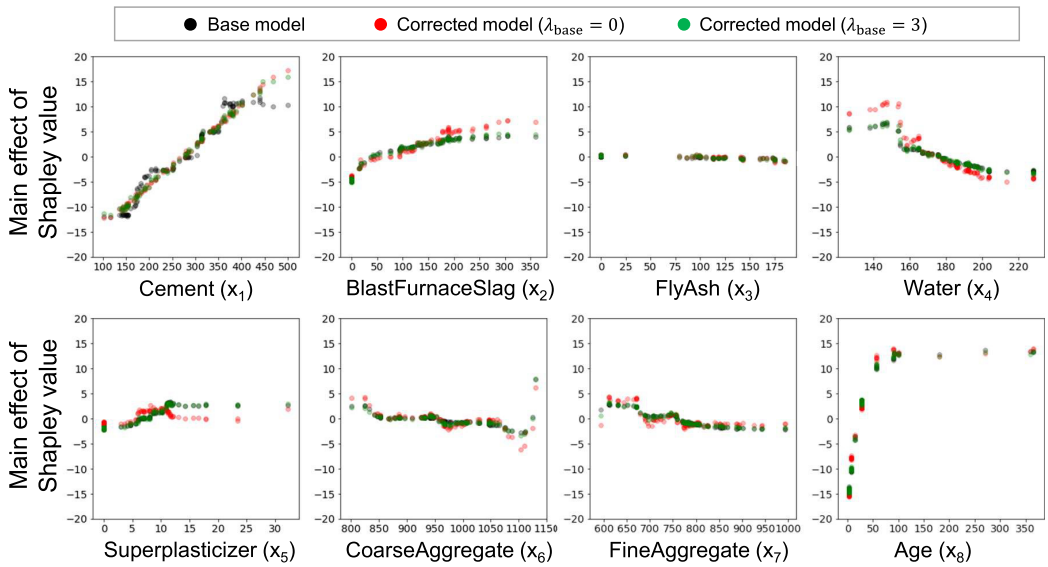


Fig. 5 Comparison of dependence plots of all features among base model, corrected model ($\lambda_{\text{base}} = 0$), and corrected model ($\lambda_{\text{base}} = 3$)

ferent preferable explanations to a case of the ten runs and obtained their prediction performances when $\lambda_{\text{base}} = 3$, as shown in Fig. 7. Simple functions $\alpha \log x + \beta$ (denoted as Log) and $\alpha/x + \beta$ (denoted as $1/x$) were applied to obtain the preferable explanation ($\boldsymbol{Z}_{\text{pref}}$) in addition to the linear function $\alpha x + \beta$ (denoted as Linear). Figure 7 (a) shows the normalized prediction errors (on the test dataset) of the corrected models with the three different functions. As we expected, the prediction performance greatly varies when the λ_{pref} is large, especially for the function of $\alpha/x + \beta$ (i.e., the $1/x$ case). The preferable explanation ($\boldsymbol{Z}_{\text{pref}}$) and corrected global explanation for x_1 of each function case are shown in Fig. 7 (b). The corrected global explanation of the $1/x$ case is largely different from the Linear case, compared with that of the Log case. Therefore, we confirmed our hypothesis that the prediction performance of EBEC is not sensitive to the weight parameters λ_{pref} and λ_{base} when an appropriate function is provided for obtaining the preferable explanation for the targeted feature but sensitive when the provided function is far different from the ideal one for the feature. Considering the risk of choosing a wrong function for the preferable explanation and balance between the prediction and explanation errors in the objective function where the dimension of the importance score (or explanation error) is the same as that of the target variable (or prediction error), especially for the Shapley value, it would be preferable to set the parameters of λ_{pref} and λ_{base} to around 1. Finally, with the Concrete dataset, $\lambda_{\text{base}} = 0$ was the best condition in terms of prediction performance, as shown in Fig. 6. These results indicate that preserving the base model's global explanation as much as possible is not the best approach to improving its prediction performance for every situation, although preserving it does not harm the corrected model's performance. Basically, domain knowledge is necessary for determining which feature's global explanation should be preserved. However, qualitative properties, such as monotonicity and continuity of the dependence plot for each feature, can be a good criterion to determine the global explanation that should be preserved even when the domain knowledge is less available.

5. Conclusion and Future Work

We proposed the model-agnostic ensemble-based explanation correction method (EBEC) for leveraging the

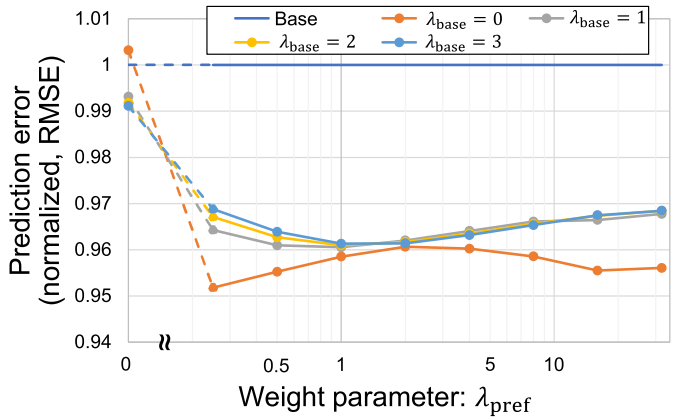


Fig. 6 Prediction error vs. weight parameter of x_1 (λ_{pref}) for various λ_{base} (averaged over 10 runs)

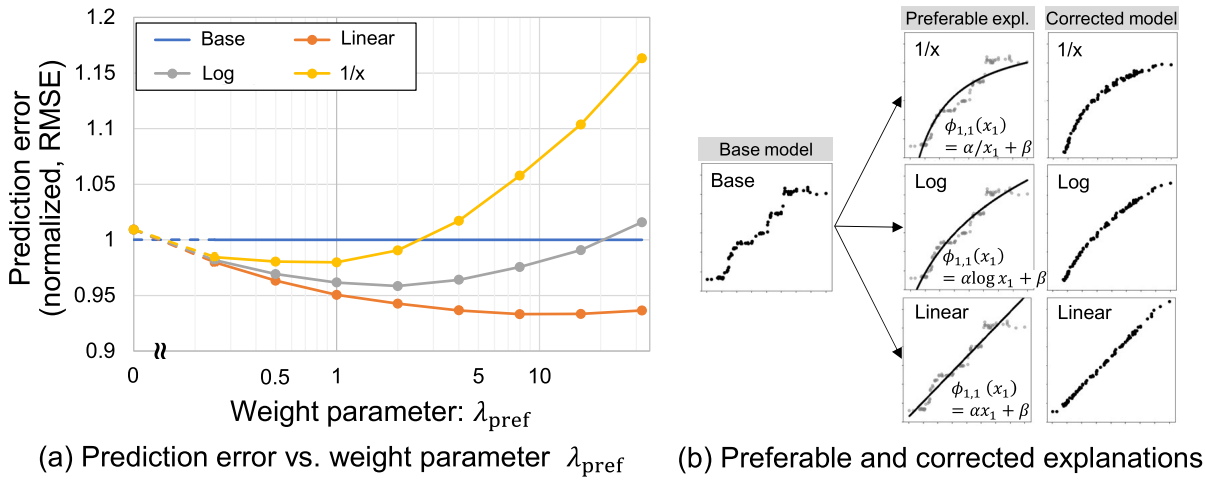


Fig. 7 (a) Prediction error vs. weight parameter of x_1 (λ_{pref}) for different settings on preferable explanation and (b) dependence plots (showing preferable and corrected explanations) of x_1 for each applied function

Rashomon effect of statistics. EBEC can correct the global explanation of any machine learning (ML) model by directly reflecting domain knowledge of experts in the model as long as the global explanation of the model can be computed. The functions of EBEC using XGBoost and Tree SHAP with public datasets was evaluated and analyzed. The evaluation results of three tasks indicate that EBEC can correct the global explanation of the base prediction model while maintaining its accuracy. During analysis with the Concrete dataset, we observed the Rashomon effect in the global explanation. The analysis also confirmed that EBEC can correct a targeted part of the global explanation of the base prediction model while maintaining other parts of it and indicated that the prediction performance of the corrected model is not sensitive to the weight parameters of EBEC when appropriate domain knowledge was given.

We presented a basic form of EBEC in this paper. However, by extending its objective function, it may enable us to tune the model in various ways. For future work, EBEC will be extended to correction of the interaction effects in global explanation, although only the main effects in global explanation were targeted in this paper. Establishing a methodology of EBEC using heterogeneous-model ensemble is also our future direction. Since EBEC is a post-processing-type correction method, it can be combined with any in-processing-type correction method that requires retraining of the model for correction. Revealing the synergy of combining EBEC with those methods may help us explore practical ways in using them.

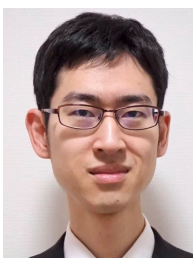
Although the performance of EBEC strongly depends on the diversity of prediction models due to the Rashomon effect, it can contribute to high-productivity ML modeling as a new type of explanation-correction method.

References

[1] S.M. Lundberg and S.I. Lee, "A unified approach to interpreting model predictions," Proc. 31st Conf. Neural Information Processing Systems (NIPS 2017), pp.4768–4777, Dec. 2017.
[2] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.I. Lee, "From local explanations to global understanding with explainable AI for trees," Nat. Mach. Intell. vol.2, pp.56–67, Jan. 2020.
[3] L. Breiman, "Statistical modeling: The two cultures," Statistical Science vol.16, no.3, pp.199–231, Aug. 2001.
[4] J. Wang and Q. Tao, "Machine learning: The state of the art," IEEE

Intell. Syst, vol.23, no.6, pp.49–55. Nov.–Dec. 2008.
[5] L. Semenova, C. Rudin, and R. Parr, "A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning," arXiv preprint arXiv:1908.01755, 2020.
[6] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," J. Mach. Learn. Res. 20 (177), pp.1–81, 2019.
[7] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention branch network: Learning of attention mechanism for visual explanation," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp.10705–10714, 2019.
[8] L. Rieger, C. Singh, W. Murdoch, and B. Yu, "Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge," Proc. 37th Int. Conf. Mach. Learn., PMLR 119, pp.8116–8126, July 2020.
[9] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, pp.785–794, Aug. 2016.
[10] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," Proc. 31st Int. Conf. Neural Information Processing Systems, pp.3148–3156. Dec. 2017.
[11] M. Hamamoto and M. Egi, "Model-agnostic ensemble-based explanation correction leveraging rashomon effect," Proc. 2021 IEEE Symposium Series on Computational Intelligence, pp.01–08, 2021.
[12] H. Kokel, P. Odom, S. Yang, and S. Natarajan, "A unified framework for knowledge intensive gradient boosting: Leveraging human experts for noisy sparse domains," Proc. AAAI Conf. Artifi. Intelli., vol.34, no.4, pp.4460–4468, 2020.
[13] C. Bartley, W. Liu, and M. Reynolds, "Enhanced random forest algorithms for partially monotone ordinal classification," Proc. AAAI Conf. Artifi. Intelli., vol.33, no.1, pp.3224–3231, 2019.
[14] A. Goldstein, A. Kapelnery, J. Bleichz, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," J. Comput. Graph. Stat., vol.24, no.1, pp.44–65, March 2015.
[15] A.S. Ross, M.C. Hughes, and F. Doshi-Velez, "Right for the right reasons: Training differentiable models by constraining their explanations," Proc. 26th International Joint Conference on Artifi. Intelli., pp.2662–2670, 2017.
[16] W. Stammer, P. Schramowsk, and K. Kersting, "Right for the right concept: Revising Neuro-Symbolic concepts by interacting with their explanations," Proc. 2021 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp.3619–3629, 2021.
[17] M. Mitsuhara, H. Fukui, Y. Sakashita, T. Ogata, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Embedding human knowledge into deep neural network via attention map," Proc. 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021) - vol.5: VISAPP, pp.626–636, 2021.
[18] G. Erion, J.D. Janizek, P. Sturmfels, S. Lundberg, and S.I. Lee, "Improving performance of deep learning models with axiomatic attribution priors and expected gradients," Nat. Mach. Intell. vol.3, pp.620–631, May 2021.
[19] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nat. Mach. Intell., vol.1, pp.206–215, May 2019.
[20] C. Rudin, "Do simpler models exist and how can we find them?," Proc. 25th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining, pp.1–2, July 2019.
[21] I. Lage, A.S. Ross, B. Kim, S.J. Gershman and F. Doshi-Velez, "Human-in-the-loop interpretability prior," Proc. 32nd Int. Conf. Neural Information Processing Systems (NIPS 2018), pp.10180–10189, Dec. 2018.
[22] I.C. Yeh, "Modeling of strength of high-performance concrete using artificial neural networks," Cem. Concr. Res., vol.28, no.12, pp.1797–1808, Dec. 1998.
[23] I. Lyse, "Tests on consistency and strength of concrete having constant water content," Proc. American Society for Testing and Materials, vol.32, Part 2, pp.629–636, 1932.
[24] P. Vlachos and M. Meyer, Statlib datasets archive, [Internet], 2005, Available from: http://lib.stat.cmu.edu/datasets/.
[25] D. Dua and C. Graff, UCI Machine Learning Repository, [Internet], 2017, Available from: http://archive.ics.uci.edu/ml.
[26] A. Datta, S. Sen, and Y. Zick, "Algorithmic transparency via Quantitative Input Influence: Theory and experiments with learning systems," 2016 IEEE Symposium on Security and Privacy (SP), pp.598–617, 2016.
[27] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," arXiv preprint arXiv:1908.09635, 2019.
[28] M. Du, F. Yang, N. Zou and X. Hu, "Fairness in deep learning: A computational perspective," IEEE Intell. Syst, vol.36, no.4, pp.25–34, July–Aug. 2020.

Masaki Hamamoto received a B.E. in electrical engineering from Kanazawa University in 2005 and M.S. in computer science from Kobe University in 2007. He is currently at Hitachi, Ltd. as a senior researcher and engaged in research into explainable AI technology.



Hiroyuki Namba received a B.E. in engineering in 2014 and M.S. in information science and technology from Tokyo University in 2016. He is currently at Hitachi, Ltd. as a researcher and engaged in research into explainable AI technology.

Masashi Egi received a B.E. and M.S. from Nagoya University in 1994 and 1996. He is currently at Hitachi Ltd. as a principal researcher. His research interests include data-science, especially machine learning, trustworthy AI, and explainable AI.