Applied Intelligence (2022) 52:471–481
https://doi.org/10.1007/s10489-021-02451-x

Lightweight surrogate random forest support for model simplification and feature relevance

Sangwon Kim[1] · **Mira Jeong**[1] · **Byoung Chul Ko**[1]

Accepted: 20 April 2021 / Published online: 3 May 2021

Abstract

In this study, we propose a lightweight surrogate random forest (L-SRF) algorithm that can be interpreted through a new rule distillation method. The common surrogate models replace the existing heavy and deep but high-performance black box model using a teacher–student learning framework. However, the student model obtained in this way must maintain the performance of the teacher model, and thus the degree of model simplification and transparency is extremely limited. Therefore, to increase model transparency while maintaining the performance of the surrogate model, we propose two methods. First, we propose a cross-entropy Shapley value to evaluate the contribution of each rule in the student surrogate model. Second, a random mini-grouping method is devised to effectively distilless important rules while minimizing the overfitting problem caused by a model simplification. The proposed L-SRF based on a rule contribution has the advantage of improving the degree of simplification and transparency of the model by realizing the large distillation ratio against the initial SRF model. In addition, because the proposed L-SRF removes unnecessary rules, it is possible to minimize the loss of the importance and relevance of each feature. To demonstrate the superior performance of the proposed L-SRF method, several comparative experiments were conducted on various data sets. We proved experimentally that the proposed method achieves a more effective performance than black box AI models in terms of model transparency and memory requirement, as well as the interpretation of the feature relevance.

Keywords Explainable artificial intelligence · Surrogate random forest · Rule distillation · Cross-Entropy Shapley

1 Introduction

When classifying artificial intelligence (AI) into four waves, deep neural network (DNN)-based algorithms, for example, a convolutional neural network (CNN), a recurrent neural network (RNN), and a generative adversarial network (GAN), correspond to the second wave and focus on improving the prediction ability by conducting learning with a large amount of data. Despite the excellent recognition performance of DNN-based algorithms, a DNN is greedy in terms of requiring large amounts of training data. Because the learning of a DNN relies on an error backpropagation algorithm, this DNN-based model cannot explain the structure of a black box model or the results of inference. The current generation of AI systems are opaque, non-intuitive, and difficult for people to understand owing to their difficulty in explaining their decisions and actions to users. Therefore, as the third wave of AI, the ability to observe the cause and effect of reasoning in a machine learning model is required, and as a result, the necessity of explainable AI (xAI) or interpretable machine learning (IML) research has emerged [1]. xAI is essential for the decision-making of users because users should be able to understand AI decisions, trust the results, and manage such information effectively.

xAI technologies can be largely divided into a transparent design and post-hoc explainability. An AI model is considered to be transparent if the model structure is understandable by itself. Transparent AI models contain one or all levels of model transparency (e.g., simulatability, decomposability, and algorithm transparency) [2]. Representative algorithms of transparent AI models include decision trees, k-nearest neighbors, and Bayesian models. These methods have an advantage in that there are few variables and

✉ Byoung Chul Ko
niceko@kmu.ac.kr

Sangwon Kim
eddiesangwonkim@gmail.com

[1] Keimyung University, Deagu, 1095 Korea

the relationship between the variables is readable; however, there is a disadvantage in that the prediction performance is lower than that of a DNN. In terms of the model prediction performance, the performance of black box models, such as a DNN, exceeds that of simple machine learning algorithms. However, because black box models cannot meet the model explainability, the goal of a post-hoc explanation is to create a separate explainable subsystem while leaving the black box model as is in terms of how the model applies inference predictions for the inputs.

Post-hoc explainability can be further divided into model-agnostic and model-specific methods [1, 2]. Because model-agnostic methods are not tied to a specific type of ML model, they are suitable for more general-purpose applications. Among several model agnostic approaches, we focus on model simplification, in which the model is simplified by eliminating parameters that approximate a complex model as a transparent model. Because a model simplified by imitating a complex model has some properties of model transparency, an explanation by simplification is possible and has the advantage of not losing the prediction performance of the original model. An explanation by simplification refers to the technique of rebuilding an entirely new system based on the trained model to be explained. Models that have simplified the previous complex model typically attempt to reduce the complexity and maintain a similar prediction performance while optimizing the functionality and similarity of the previous model [2]. In addition, it is possible to describe the feature relevance for training and test data through a simplified approach.

Explanation by simplification is a technique that can be applied most widely in the category of post-hoc model-agnostic methods regardless of the complexity of the black box model [2]. In recent years, there have been many studies on model simplification in the field of xAI, indicating that this approach is expected to continue to play a central role in xAI. Similar studies related to explanation by simplification are as follows.

Bastani et al. [4] proposed model extraction for interpreting the overall reasoning process achieved by a model. Given a model f, the interpretation produced by the proposed approximation is $T(x) \approx f(x)$, where T is an interpretable model. This method takes T as a decision tree, which has been established as highly interpretable. However, an interpretable decision tree incurs an overfitting and a deteriorated performance compared to a complex model. Tan et al. [3] proposed a model distillation algorithm called distill-and-compare. With this method, a transparent student model is trained to mimic the risk score assigned by the black box model as a teacher to gain insight into the black box model. However, this method does not present the difference in prediction performance and feature contribution according to the degree of distillation of the mimic model, which is an important measure of model simplification.

In terms of simplification methods of DNN using knowledge distillation, Zagoruyko and Komodakis [5] defined attention to the CNN, which improves the performance of the student CNN network by mimicking the attention map of a powerful teacher network. Similarly, Xu et al. [6] introduced DarkSight, a dimension reduction technique for interpreting deep classifiers based on a knowledge distillation. DarkSight matches the dark knowledge between students and teachers and compresses the black box teacher classifier into a simple and interpretable student classifier. However, because these methods still rely on compressing the existing DNN model to shallow DNN model, the model's transparency is limited. Kim et al. [7] proposed a method for analyzing and simplifying the black box model of a deep random forest (RF) using the proposed rule removal. The feature contribution provides the basis for determining the impact of a feature on the decision-making process of the rule set, and the black box model can be simplified by selecting top-k important rules based on measuring the feature contribution. As a result, the simplified model has fewer parameters and rules than the original model. Because this method relies on the traditional tree rule evaluation method, the reliability of rule removal is weak. Kim and Boukouvala [8] investigated the effectiveness of a subset selection method for developing a surrogate regression that balances accuracy and complexity. The subset selection produces a sparse regression model by selecting only a subset of the original features, which are linearly combined to produce a different set of surrogate models. However, this method requires high computational cost for feature selection and identification of model parameters, and performance degrades as the dimension of the problem increases. As an application of model simplification, Kim et al. [9] also proposed a lightweight pupil tracking algorithm for on-device ML that uses a fast and accurate cascade deep regression forest instead of a DNN. A pupil estimation is applied roughly in a layer-by-layer regression forest structure and simplifies each regression forest using the proposed rule distillation algorithm to select top-k significant rules that make up the regression forest. The goal of this algorithm is to create a more transparent and adaptable model for application to on-device ML systems while maintaining an accurate pupil tracking performance. However, this method has the disadvantage that the higher the distillation ratio, the more the model performance is overfit or deteriorated.

The various model simplification methods introduced so far still have the following limitations. 1) Surrogate models are still not transparent because they are composed of several unnecessary rules or layers. 2) The higher the model distillation ratio, the more the surrogate model is trained to

overfit and the feature relevance performance decreases. 3) The algorithm for model distillation is heuristic or relies on traditional rule contributions.

In this study, we propose a new xAI method called lightweight surrogate random forest (L-SRF) that simplifies the model by decomposing the black box teacher model and increases transparency at the same time. With a pattern similar to distill-and-compare [3], the L-SRF can replace the existing heavy, deep, but high performing teacher complex model while maintaining the performance of the black box model. Through an L-SRF, it is possible to analyze the feature relevance that affects the inference, and to explain how the L-SRF model structures operate in the inference process. In addition, our approach works with any model family and is independent of the implementation.

In our initial study [10],[1] we introduced a brief SRF to simplify the black box teacher in terms of the model size and prediction performance for solving the classification. However, in this study, we focus on a detailed explanation of how the black box complex model is distilled and rebuilt into an explainable simplification structure and prove the efficiency of the L-SRF model in terms of the transparency and accuracy.

2 Surrogate random forest

DNN models achieve a higher performance as the model becomes deeper and heavier but has a disadvantage in terms of the explainability. In addition, owing to the large number of parameters, the memory usage increases and the speed decreases. To create a model that is explainable and lightweight, a surrogate model based on the teacher–student (T-S) framework [3, 5, 6, 11] was introduced that can construct a shallow student model by reducing the size of the teacher model while maintaining a performance similar to that of the deep and wide teacher model. The approach for creating a surrogate model can be divided into two types depending on which model the user is targeting. If we focus only on reducing the weight of the model, the complex teacher model can be reconstructed into a transparent approach such as a decision tree. This method has an advantage in that the model itself is interpretable and transparent; however, it has a disadvantage in that the performance of the surrogate model is much lower than that of the teacher model when there are numerous classification classes, and the number of dimensions of the feature vector is large. Another method is to reduce the surrogate model itself to a gray box. With this method, the performance of the model is similar to that of the teacher model, and the feature relevance and importance that contribute to the decision of the model can be inferred. In this case, a random forest (RF) [12], gradient boosting method (GBM) [13], XGBoost [14], and CatBoost [15] methods are used as surrogate models.

In this paper, we propose an L-SRF model that can maintain the performance of a complex black box model while having fewer parameters using the T-S framework. In addition, instead of a typical post-hoc based method that must concurrently maintain a black box model for prediction and a surrogate model for explanation, this study aims at achieving a prediction and explanation simultaneously with a single L-SRF. The GBM, XGBoost, CatBoost and an RF are mainly used as student models to create an explainable surrogate model. Unlike the GBM, XGBoost, CatBoost, an RF preserves the properties of the rules that make up the tree, and thus are more effective in eliminating unnecessary rules while maintaining the tree structure. By contrast, boosting-based models change the structure of a tree by using gradient differences, and thus it is difficult to apply to a rule distillation using the characteristics of the rules. To create a surrogate model that achieves a good performance, it is important to develop a teacher model with an excellent performance. Therefore, this paper uses automated machine learning [14] to create a DNN-based teacher model with the highest performance for a given dataset. By following the T-S framework and our proposed rule distillation algorithm, it is possible to create a reduced L-SRF model that inherits the characteristics of the teacher.

The process of generating a student RF model based on the T-S framework is as shown in Fig. 1. The training dataset is divided into dataset A for training the teacher model and dataset B for training the student model. First, the teacher DNN model is obtained by applying dataset A labeled with 0s and 1s (hard target) to AutoML (http://AutoML.org). Then, by inputting the unlabeled dataset B in the trained teacher DNN, a soft target, which is a class-specific probability value output from softmax, is assigned as a label to dataset B. Now, we train the student RF model using dataset B, which is labeled a soft target. The student RF selects the model with the most similar performance as the teacher while controlling the number and depth of the tree. During this process, to prevent an RF overfitting, various RFs are learned using the M-fold-cross-validation method, and the RF with the best performance is selected as the final student RF. The selected student RF model can be trained to consider the inter-class relationships of the teacher DNN model by using training data labeled as a soft target. The student RF model created through a T-S framework learning is called the SRF model.

[1] A shorter version of this paper was presented at the NeurIPS2020 Workshop.

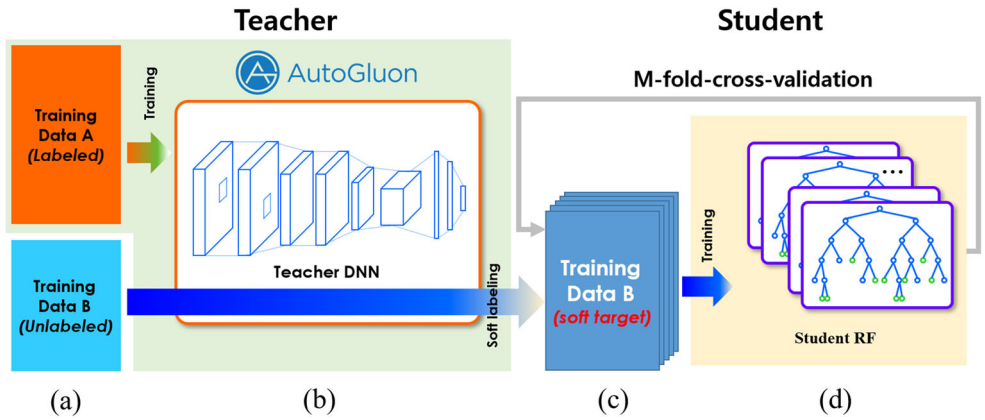

Fig. 1 Teacher-student training framework: **a** the training dataset is divided into labeled dataset A and unlabeled dataset B. The teacher DNN model is trained with dataset A. **b** unlabeled dataset B is applied to the trained teacher DNN and a class-specific probability value is assigned as a label to dataset B. **c** soft target dataset B is used to train the student RF model. **d** the RF with the best performance is selected as the final student surrogate RF using M-fold-cross-validation method

3 Lightness of SRF

A Shapley additive explanation [16] is a representative surrogate model-based feature relevance estimation method. With this method, a data prediction is conducted using a black-box model, and the feature relevance is applied using a surrogate model. Therefore, for an explainable prediction of the data, the black box model and the surrogate model must be used at the same time. However, this typical post-hoc based method is difficult to use in a lightweight system because the size of the model becomes excessively large. Therefore, our proposed L-SRF has the following goals: 1) L-SRF does not maintain a separate black box model, but can preserve the data prediction performance, and 2) the model itself has better transparency than the initial surrogate model. 3) It makes the SRF lighter but maintains the explainability of the feature relevance of the initial surrogate model by eliminating only redundant or unnecessary rules. To further lighten the SRF model obtained from the T-S framework, we proposed a rule distillation method based on the Cross-Entropy Shapley (CES) value.

3.1 Cross-entropy Shapley value

The RF is an ensemble model of decision trees, where each decision tree is a set of rules that are paths from a root to an intermediate and finally to a leaf node [12]. We can reduce the rules of the RF by evaluating the contribution of all the rules constituting the decision trees and eliminating the rules with a low contribution. In this study, we use the Shapley value [17] to determine the contribution of the rule. The original Shapley value was used to measure the contribution of the input feature from a machine learning model. This value measures the difference in accuracy according to the presence or absence of a specific feature, and the greater the difference, the higher the degree of contribution given to the corresponding feature. In this study, instead of determining the contribution of the input feature, the Shapley value is used to determine the contribution to the rule in the SRF.

Because the Shapley value is an algorithm that determines the contribution of individual features, it is necessary to modify the algorithm to evaluate the contribution of the rules constituting the SRF. Therefore, we propose a new CES value to evaluate the contribution of each rule of the SRF. Whereas the existing rule elimination method evaluates the prediction accuracy according to the rule of the tree [10], the proposed CES values can be used to evaluate the more detailed rule contribution by considering the probability for each class of a particular rule used in the tree. First, suppose that the rule set R of SRF consists of N rules. Here, r_j is the j-th rule constituting R, and $\tilde{\mathrm{R}}$ is a subset R composed of $N-1$ rules excluding the r_j rule. In this case, the contribution of the r_j rule in a subset $\tilde{\mathrm{R}}$ can be calculated by considering the classification probability p_i for each class c. In addition, $T_{CE}(\tilde{\mathrm{R}}, r_j)$ represents the cross-entropy between a subgroup g and an individual rule r_j.

$$T_{CE}(\tilde{\mathrm{R}}, r_j) = -\sum_{i=1}^{c} p_i(\tilde{\mathrm{R}} \cup r_j) \log(p_i(\tilde{\mathrm{R}})) \tag{1}$$

The CES value $\Phi(r_j)$ of a j-th rule is the weighted and summed contribution of all possible rule combinations that the j-th rule can contain:

$$\Phi(r_j) = \sum_{\tilde{\mathrm{R}} \subseteq \{r_1,\ldots,r_n\} \setminus r_j} \frac{|\tilde{\mathrm{R}}|!(N - |\tilde{\mathrm{R}}| - 1)!}{N!} T_{CE}(\tilde{\mathrm{R}}, r_j) \quad (2)$$

By measuring the CES value for each rule of the initial SRF, the SRF is reduced by eliminating the rule with a low contribution.

3.2 Rule distillation using mini-grouping

In general, the number of rules of an SRF ranges from tens of thousands to millions of rules, depending on the tree and node depth. The individual rule's contribution check and distillation process for the entire rule not only requires a lengthy processing time, it can also cause an over-fitting of the model. Therefore, in this study, a random mini-grouping method was devised to minimize the overfitting problem caused by rule distillation based on individual contribution checks. In the random mini-grouping, the rules of the SRF are randomly grouped into K mini-groups, as shown in Fig. 2b, and the degree of contribution is evaluated by estimating the CES value for each mini-group. The CES value for each mini-group is measured (Fig. 2c) against rest mini-groups in the same manner as the original Shapley method. This process is repeated H times, and the final contribution of the rule is determined by the average value of each rule in the mini-group, as shown in Fig. 2c. Finally, by eliminating the rules with a low contribution according to the measured contribution, the model can greatly distill the model size while maintaining the existing prediction performance.

Algorithm 1 introduces the rule distillation process through a random mini-grouping using the CES value of the SRF model. Equation 3 of Algorithm 1 was modified to calculate the CES value between the mini-group set and each mini-group g instead of calculating the CES value between a subset $\tilde{\mathrm{R}}$ and individual rule r.

Algorithm 1 Rule distillation using random mini-grouping.

Input: Trained SRF, Number of rules N, Number of mini-groups K, Number of iterations H, Rule set $\mathbf{R} = \{r_1, r_2, ..., r_n\}$, Group set $\mathbf{G} = \{g_1, g_2, ..., g_k\}$, mini-group g, rule elimination rate δ

Output: L − SRF

$\mathbf{R}, \mathbf{G}^h, g_k \leftarrow \emptyset$

$|g| = N/K$ ▷ the number of rules in each mini-group

Generate rule set $\mathbf{R}$ from SRF (Fig. 2a)

for $h = 1$ *to* H **do** ▷ repeat H times

  for $k = 1$ *to* K **do** ▷ repeat K mini-groups

    $g_k = random.choices(\mathrm{R})$ ▷ randomly extract a mini-group g_k from $\mathbf{R}$ (Fig. 2b)

    $\mathbf{G}^h = \mathbf{G}^h \cup g_k$ ▷ stack each mini-group g_k to h-th group set $\mathbf{G}^h$ (Fig. 2b)

  end for

  $\mathbf{G}^h = \{g_1, g_2, ..., g_k\}$ ▷ a generated group set

  for $k = 1$ *to* K **do** ▷ repeat K mini-groups

$$\Phi^h(g_k) = \sum_{\mathrm{G}^h \subseteq \{g_1,\ldots,g_k\} \setminus g_k} \frac{|\mathrm{G}^h|!(K - |\mathrm{G}^h| - 1)!}{K!} T_{CE}(\mathrm{G}^h, g_k) \quad (3)$$

  ▷ calculate the h-th CES value Φ^h of k-th mini group g_k in a group set $\mathbf{G}^h$ (Fig. 2c)

  end for

end for

for $i = 1$ *to* N **do** ▷ repeat N rules

  ▷ average calculation of CES values for each rule r_i belonging to the mini-group g_k (Fig. 2d)

$$\hat{\Phi}(r_i) = \frac{1}{H} \sum_{h=1}^{H} \sum_{k=1}^{K} \{\Phi^h(g_k) \mathbb{1}(\cdot)\} \quad (4)$$

  ▷ where indicator function $\mathbb{1}(\cdot) \leftarrow 1 if g_k \ni r_i$

end for

$\mathbf{R}^* \leftarrow sort(\hat{\Phi}(r_i), ascend)$ ▷ sort existing rules by CES values

L − SRF $\leftarrow build\_RF\_from\_rules(\mathbf{R}^*[: N \times \delta])$ ▷ eliminate weak ruels and rebuild L-SRF

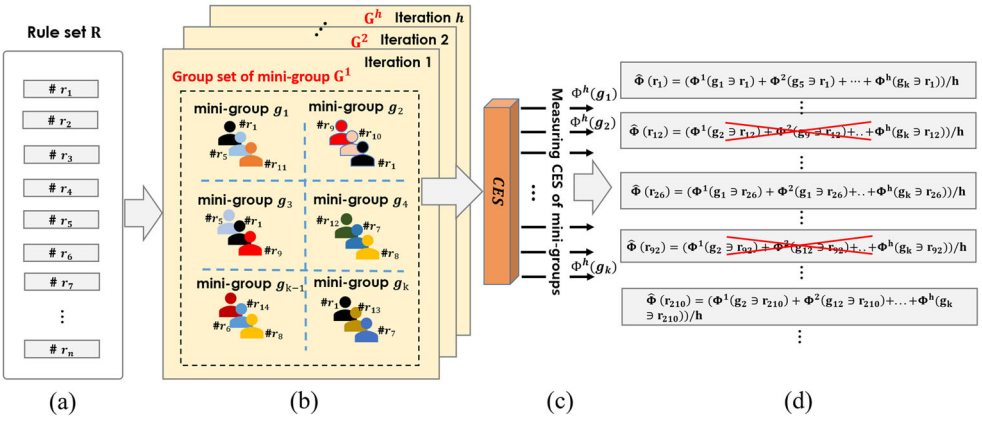

Fig. 2 Rule distillation process using random mini-grouping and CES: **a** a rule set **R** consisting of rules extracted from the SRF and **b** mini-groups randomly generated from **R** and **c** the CES value is measured for each mini-group. This process is repeated H times. **d** Average calculation of CES values for each rule belonging to the mini-group. Rules with small average CES values are eliminated

4 Materials and methods

4.1 Datasets

The UCI repository [18] and the Penn Machine Learning Benchmarks (PMLB) [19] provide several datasets for testing the machine learning and intelligent systems. In this paper, we prove the effectiveness of the proposed method using Adult Income among UCI datasets, and the Phoneme, Car, Mushroom, Chess, and the Mfeat factors among PMLB. The Adult Income dataset predicts whether an individual's income will exceed $50,000 per year, based on demographics of adults aged 16 and older. This dataset contains 48,842 demographic data on people who participated in the 1994 census for 14 attributes such as age, gender, occupation, workclass, and education. The Phoneme is a dataset to distinguish between nasal and oral sounds. It contains 5,404 data for six attributes. The Car is a dataset for evaluating cars according to the conceptual structure, such as the estimated safety of the car, trunk size, number of people carrying, number of doors, etc. It consists of 1,728 samples for seven attributes. The Mushroom is a dataset containing physical properties for classification as poisonous or edible and contains 8,124 samples with 20 properties. The Chess is a dataset for estimating the result of a chess match when only king and pawn remain in the white side and king and rook remain in the black side. It consists of 3,196 samples with 20 attributes. The Mfeat factor is a dataset for recognition of handwritten numerals (0-9). 200 instances per class (for a total of 2,000 samples) have been digitized in binary images with 216 attributes. We conducted experiments using each dataset during the training and testing processing for the models. The UCI adult income dataset was divided into a ratio of approximately 7:3 following the official training/testing split, and the other datasets were divided into five-folds.

4.2 Toolkit and library

In this study, the AutoGluon [20] toolkit is used to generate the AutoML-based teacher model, and Scikit-learn and Python are used to implement the surrogate RF model. In addition, we use the SHAP package in Python to visualize the influence of the feature vectors on the output.

5 Experimental results

Selecting an accurate teacher model is one of the essential factors in the T–S framework because the performance of the student model largely depends on the performance of the teacher model. Various machine learning algorithms can be used as the teacher model, but in this study, AutoGluon, an AutoML toolkit [20] for deep learning, is used. AutoML is highly adaptable to various real-world applications such as images, text, or tabular data, and can automatically utilize the latest deep learning technologies without expert knowledge. In addition, AutoML makes it easy to utilize automatic hyperparameter tuning, model selection/architecture discovery, and data processing. For the student model SRF, Scikit-learn and Python were applied.

5.1 Hyper-parameter evaluation for model simplification

The mini-grouping process requires two hyper-parameters, the number of mini-groups, and the number of grouping iterations. Because the size and performance of the L-SRF depend on two parameters, it is necessary to find the optimal parameters to create a lightweight and generalized L-SRF. First, to find the optimal iteration, we measured the F1-score by changing the number of iterations and the distillation rate for the initial SRF using the UCI Adult Income dataset, as shown in Table 1. At this time, the maximum number of allowed mini-groups was fixed at 50. If the maximum number of allowed mini-groups is too large, an overfitting may occur because the number of rules allocated to one mini-group is too small. Here, the initial SRF model created based on the T-S framework has 100% (1.0) of the rules before rule distillation is applied. From the initial SRF, we repeatedly removed the number of rules by 10% (0.1) and evaluated the relative F1-score.

As shown in Table 1, the difference in F1-score according based on the number of iterations is low. This means that the number of iterations does not significantly affect the

Table 1 Comparison of F1-score performance according to number of iterations for mini-grouping with rule distillation rate for the SRF using UCI Adult Income dataset

Rule distillation rate (δ)	F1-Score				
	Number of Iterations				
	1	3	5	10	15
1.0	90.40	90.40	90.40	90.40	90.40
0.9	90.39	90.41	90.40	90.40	90.39
0.8	90.39	90.43	90.43	90.40	90.38
0.7	90.38	**90.44**	90.42	90.37	90.34
0.6	90.38	90.42	90.41	90.34	90.33
0.5	90.16	90.38	90.42	90.31	90.28
0.4	90.15	90.39	90.41	90.27	90.31
0.3	89.86	89.96	90.34	89.74	90.22
0.2	79.09	89.33	89.46	89.12	89.57
0.1	72.99	81.76	86.12	83.34	87.50

improvement in the performance of the L-SRF. When the number of iterations is 3, the average F1-score of the all rule distillation rate shows slightly higher than other cases. Increasing the number of iterations for L-SRF increases the effort required during the learning process, and thus when the performance is similar, an effective approach is to selectively limit the number of iterations as much as possible and find the optimal number of mini-groups, which is another parameter. Therefore, according to the experimental results listed in Table 1, we determined the optimal number of iterations to be 3. According to the results, the number of iterations was therefore set to 3, and the number of minigroups was repeatedly changed. Table 2 shows the resulting measurement F1-score while adjusting the rule distillation rate and the number of mini-groups. As shown in Table 2, it can be seen that the highest F1-score is obtained when the number of mini-groups is 50 and the rule distillation rate is 0.7. In addition, in terms of F1-score, SRF (0.7) shows an approximately 0.04% higher F1-score than the initial SRF (1.0) without a rule distillation, and thus it can be seen that unnecessary rules or rules with a low contribution are effectively eliminated through the proposed algorithm.

The purpose of the proposed L-SRF is to design a lightweight surrogate model that can operate in a device with low specifications while maintaining the performance of the teaching model. Therefore, we evaluated how the number of parameters and operations decreased according to the rule distillation rate in the SRF. Similarly, a change in performance according to the rule distillation rate was also observed. During the experiment, the number of iterations was set to 3, and the number of mini-groups was set to 50 according to the results of the previous experiments.

Table 2 Comparison of F1-score performance according to number of mini-groups with rule distillation rate for the SRF using UCI Adult Income dataset

Rule distillation rate (δ)	F1-Score					
	Number of mini-groups					
	1	20	30	40	50	60
1.0	90.40	90.40	90.40	90.40	90.40	90.40
0.9	90.40	90.41	90.37	90.41	90.41	90.41
0.8	90.40	90.39	90.37	90.41	90.43	90.41
0.7	90.40	90.39	90.36	90.42	**90.44**	90.39
0.6	90.41	90.39	90.37	90.42	90.42	90.38
0.5	90.40	90.37	90.32	90.42	90.38	90.37
0.4	90.27	90.38	90.37	90.35	90.39	90.39
0.3	90.25	90.26	90.04	90.01	89.96	90.05
0.2	89.57	89.67	89.91	89.00	89.33	89.44
0.1	88.06	84.03	86.05	84.56	81.76	86.69

As shown in the experimental results in Table 3, the number of SRF parameters decreased in proportion to the rule distillation rate. In particular, when we eliminated 70% of the rules from the complete set of rules (SRF(0.3)), the number of parameters was reduced by approximately 53% while maintaining the same level of F1-score. Through these experiments, it can be seen that the proposed model simplification method can effectively distillate the size of the model while maintaining the existing F1-score. In particular, in the case of SRF (0.1), the number of rules is reduced to about 300, so the variables included in the rules can be read, and the size of the rule set can be managed by humans without external assistance.

5.2 Surrogate model comparison

Representative surrogate models used in machine learning include the GBM [13], XGBoost [14], CatBoost [15] and an RF [12]. With the GBM, the gradient informs the weakness of the classifier learned thus far, and the model learns to compensate for the weakness. The GBM has an excellent boosting ability, but the learning is slow and has an overfitting problem. XGBoost was proposed to overcome the shortcomings of the GBM. This method is faster than the GBM and provides regulation and an early stopping function to prevent an overfitting. CatBoost provides a novel gradient boosting scheme for reducing overfitting, as well as this method allows to fast parameter tuning through categorical feature supporting. To test this possibility as a surrogate, we first trained four surrogate models with the output of the same teacher DNN using the same Adult Income dataset. All four methods consisted of 10 trees, and the maximum depth was fixed at 7, and the number of features was set to $\sqrt{d}$ for finding the best splits in each tree nodes.

Table 3 The change in model size according to the rule distillation rate with the UCI Adult Income dataset

Rule distillation rate (δ)	Precision	Recall	F1-score	Number of param.
1.0	**86.15**	95.10	90.40	9,286
0.9	86.08	95.20	90.41	8,901
0.8	86.09	**95.24**	90.43	8,428
0.7	86.13	95.21	**90.44**	7,912
0.6	86.10	95.20	90.42	7,301
0.5	86.08	95.14	90.38	6,661
0.4	86.02	**95.24**	90.39	5,892
0.3	86.03	94.27	89.96	4,898
0.2	86.00	92.92	89.33	3,880
0.1	86.60	77.43	81.76	2,459

Table 4 Comparison of precision, recall, F1-score, and number of parameters for four surrogate models trained with a teach model and the UCI Adult Income dataset

Evaluation Metrics	Teacher DNN [20]	Surrogate GBM [13]	Surrogate XGBoost [14]	Surrogate CatBoost [15]	SRF(1.0)
Precision	88.62	84.81	87.79	87.33	86.15
Recall	93.41	96.72	93.79	94.13	95.10
F1-score	90.79	90.37	90.69	90.60	90.40
Number of param.	–	17,162	15,888	17,820	9,286

As shown in Table 4, the GBM showed a level of F1-score 0.42% lower than that of the teacher model, and XGBoost was 0.1%, and CatBoost was 0.19% which showed no significant difference from the teacher. The SRF showed a 0.39% lower performance than the teacher model, which achieved a performance 0.09% lower than that of XGBoost. In terms of the number of parameters, the GBM requires approximately 1.84-times more parameters and XGBoost requires approximately 1.71-times more parameters, and CatBoot requires approximately 1.91-times more parameters than the SRF. From these results, we confirmed that SRF is a suitable model for model simplification because it inherits the performance of the teacher model more closely than other boosting-based methods and uses a small number of parameters.

5.3 Comparison with machine learning

To prove the excellent performance of the proposed L-SRF, a comparative experiment was conducted with the RF [12], ExtRa [21], k-NN [22], SVMs [23], gcForest [24], AdaBoost [25], GBM [13], XGBoost [14], LGBM [26], CatBoost [15], NgBoost [27] which is a multi-parameter boosting algorithm, and KiGB [28] which is an unified framework for knowledge intensive gradient boosting, and teacher DNN based on Auto Gluon using additional datasets such as the Phoneme, Car, Mushroom, Chess, and the Mfeat factors of PMLB. The evaluation procedure was conducted under Five-fold cross-validation manner. All 13 methods used the same trees, the maximum depth, and the number of features as in previous experiments..

As shown in Table 5, teacher DNN performed the best for all of the dataset. Among the comparison methods except for the teacher DNN, XGBoost showed the best performance for the Car and Chess datasets. Five methods that do not use boosting ([12, 21, 24]) showed an overall lower accuracy than the other boosting-based methods ([25], [13, 15, 26, 28]). Three boosting-based methods (CatBoost [15], NgBoost [27], and KiGB [28]) showed similar results for five datasets. However, these methods

Table 5 Performance comparison with machine learning models using PMLB datasets

Methods	Five-Fold CV Accuracy (%)±std				
	Phoneme	Car	Mushroom	Chess	Mfeat factors
RF [12]	85.12±0.85	84.84±3.54	**100.0**±0.00	94.06±0.73	91.65±1.19
ExtRa [21]	81.18±0.88	85.59±2.54	99.78±0.19	94.96±0.74	93.55±0.48
k-NN [22]	**88.43**±0.38	82.65±6.49	99.98±0.05	95.78±0.42	94.60±0.72
SVMs [23]	83.79±0.52	86.87±4.38	99.74±0.18	97.22±0.56	89.20±2.05
gcForest [24]	87.58±0.43	88.95±4.96	**100.0**±0.00	97.31±0.35	**95.35**±0.41
AdaBoost [25]	86.07±0.78	90.57±2.77	99.43±0.24	94.52±1.05	92.15±1.53
GBM [13]	86.75±0.41	92.31±3.39	**100.0**±0.00	97.90±1.02	91.25±1.17
XGBoost [14]	88.34±0.61	**93.29**±2.52	**100.0**±0.00	**99.12**±0.36	94.55±1.20
LGBM [26]	85.60±0.43	83.17±7.01	99.84±0.14	96.43±0.85	94.05±1.24
CatBoost [15]	86.68±0.52	90.86±2.76	**100.0**±0.00	98.84±0.23	91.50±1.28
NgBoost [27]	86.68±0.35	92.31±0.10	**100.0**±0.00	97.28±0.69	90.15±1.60
KiGB [28]	86.82±0.58	92.71±2.46	**100.0**±0.00	97.78±0.91	90.85±0.82
Teacher DNN (AutoML [20])	91.23±0.49	98.73±0.68	**100.0**±0.00	99.91±0.08	99.90±0.12
SRF(1.0)	86.36±0.73	91.61±2.67	**100.0**±0.00	97.03±0.89	90.60±0.80
L-SRF(0.7)	85.99±0.64	90.22±2.26	**100.0**±0.00	96.84±0.79	90.20±0.60
L-SRF(0.4)	85.21±0.87	88.83±3.57	99.90±0.03	96.71±0.61	88.75±0.79

still showed a 1-4% difference in accuracy compared to XGBoost [14]. Compared to L-SRF (0.7), the accuracy of XGBoost was improved by about 3% overall, but the number of parameters is actually 2 times more required. In particular, the L-SRF (0.7) model obtained through rule distillation showed a similar performance with the original SRF (1.0) model except the Phoneme generated by teacher DNN. These results show that the proposed CES value and random mini-grouping method of L-SRF effectively eliminate unimportant rules that degrade the performance, thereby increasing the performance. However, although L-SRF (0.4) used 1.34 times fewer parameters than L-SRF (0.7), the accuracy was similar or higher than that of gcForest [24] and LGBM [26] for the Car, Mushroom and Chess datasets.

Experimental results showed that the L-SRF model based on the T-S framework can maintain similar performance to the method using only the model itself, although it used a small number of parameters.

5.4 Visualization

Among the post-hoc xAI methods, unlike the DNN-based method, the biggest advantage of the RF-based surrogate model is that it can measure the feature relevance. The contribution of the feature to the output of L-SRF was analyzed through Shapley additive descriptions (SHAP) [16], which can measure the feature relevance, and is an xAI technique. In other words, we use SHAP to quantify how important the features of the L-SRF model are to the results, and based on this, we verify that the proposed method can achieve a model simplification while maintaining the feature relevance. In addition, by comparing the SHAP results of the original SRF (1.0) and L-SRF (0.2), it can be seen that even if the model is light, it does not overfit and preserves important rules well, thereby maintaining the feature relevance similar to that of the original model. To more easily compare the SHAP values of the original SRF (1.0) with those of the simplified L-SRF (0.2), the local importance for individual features was measured and visualized in a plot using the UCI Adult Income dataset, as shown in Fig. 3.

As shown in Fig. 3, the two methods show mostly similar patterns in terms of local importance. The features affecting the output result and having similar patterns between the two models are in order of “EducationNum”, “Race”, and “CapitalGain”, and the positive and negative effects according to the feature values also show similar patterns. Although many rules were eliminated from the initial SRF, the L-SRF (0.2) showed a similar pattern in terms of the feature correlation, and thus we can see that unnecessary rules were effectively eliminated through the proposed CES and mini-grouping

Second, we visualized the correlation between features using SHAP dependence plots to check whether the correlations are preserved even in a simplified L-SRF (0.2) model. Based on a comparison of the correlations between all features, “EducationNum” was found to have the highest correlation with the other features. Figure 4 shows the SHAP values for the combination of the two features, “Age”-“EducationNum” and “WorkClass”-“EducationNum”, which had a high correlation with “EducationNum”. Through this result, we can infer that the proposed L-SRF model achieves a consistent feature relevance even after the simplification process because it maintains correlations between features even in lightweight models. Therefore, the proposed L-SRF method can improve the model transparency by applying a model simplification while maintaining the feature relevance, which is the basic property of xAI explainability.

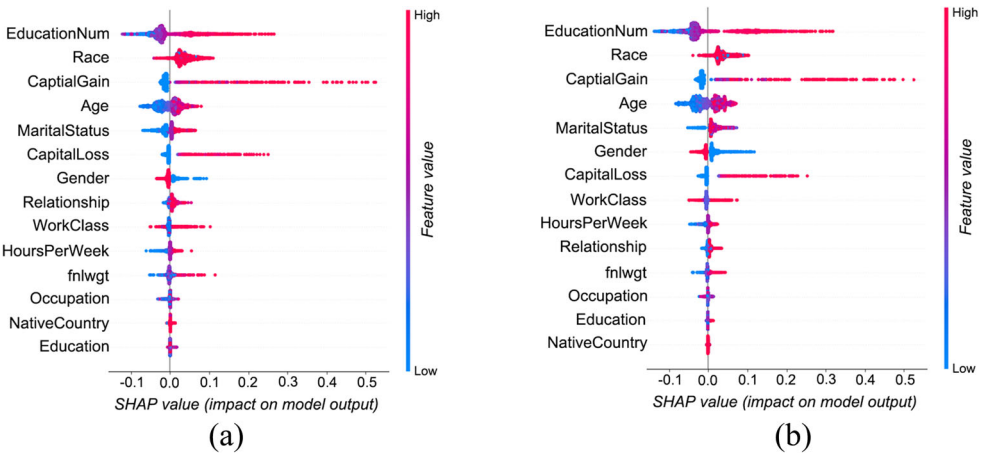

Fig. 3 Visualization for the magnitude of influence on the output of input features according to model simplification based on SHAP value using UCI Adult Income dataset. The x-axis represents the SHAP mean value (global importance) and SHAP value (local importance), and the y-axis represents 14 input features. **a** the initial SRF (1.0) model without rule distillation, and **b** the rule-distillated L-SRF (0.4) model. The importance of each feature for two models shows almost similar results

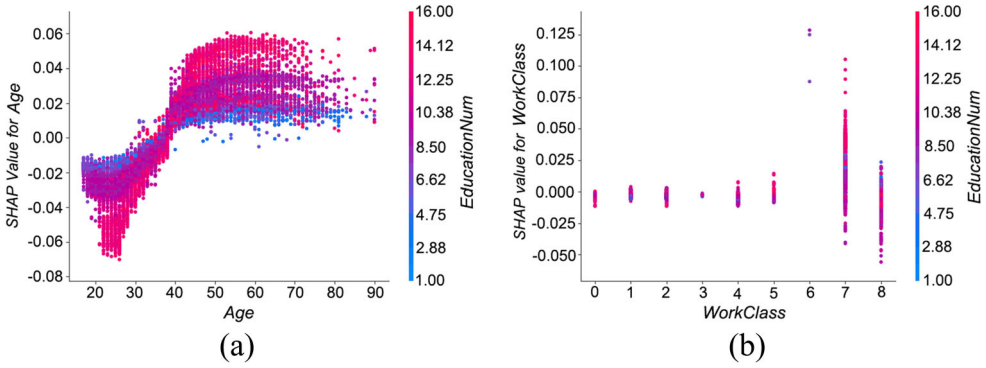

Fig. 4 Visualization of correlation between features estimated from L-SRF (0.4) using UCI Adult Income dataset, **a** SHAP dependence plot between "age" and "EducationNum" feature, **b** SHAP dependence plot between "WorkClass" and "EducationNum" feature

6 Conclusion

In this paper, among several xAI approaches, we proposed a new L-SRF algorithm that can increase the transparency of a complex black box model through a model simplification and analyze the features that influence the prediction through the feature relevance. The proposed L-SRF method has confirmed the ability to compress the model on a small scale while guaranteeing the same prediction performance as the existing complex model. In particular, by applying mini-grouping and a CES proposed in an RF to create a surrogate model instead of XGBoost, GBM or the CatBoost, we were able to design a lightweight surrogate model that can effectively reduce the number of rules and maintain the prediction performance and feature relevance at the same time.

The proposed L-SRF is similar to XGBoost, GBM, and CatBoost in terms of accuracy through experiments on various data sets. In terms of model size, the proposed method effectively eliminated the less important rules, thereby significantly reducing the model size and avoiding the overfitting problem caused by a model reduction. In future research, we will improve the L-SRF model for application to a variety of data, including images and videos, and apply it to an embedded device to test its feasibility in real problems. Furthermore, because the L-SRF is still less accurate than XGBoost even if unimportant rules are removed, it is necessary to devise a lightweight version of XGBoost by modifying the proposed rule distillation method to fit the XGBoost.

Acknowledgements This research was supported by the Bisa Research Grant of Keimyung University in 2021.

References

1. Adadi A, Berrada M (2018) Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access 6:52138–52160
2. Arrieta AB et al (2020) Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. ELSEVIER Inf Fusion 58:82–115
3. Tan S et al (2018) Distill-and-compare: auditing black-box models using transparent model distillation. In: 2018 AAAI/ACM conference on AI, ethics and society. pp 303–310
4. Bastani O, Kim C, Bastani H. (2017) Interpretability via model extraction. arXiv:1706.09773
5. Zagoruyko S, Komodakis N (2017) Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In: ICLR, pp 1–11
6. Xu K et al (2018) Interpreting deep classifier by visual distillation of dark knowledge. arXiv:1803.04042
7. Kim S, Jeong M, Ko BC (2020) Interpretation and simplification of deep forest. TechRxiv, techrxiv. 11661246.v1
8. Kim S, Boukouvala F (2020) Machine learning-based surrogate modeling for data-driven optimization: a comparison of subset selection for regression techniques. Springer Optim Lett 14:989–1010
9. Kim S, Jeong M, Ko BC (2020) Energy efficient pupil tracking based on rule distillation of cascade regression forest. MDPI Sensors 20:1–17
10. Kim S, Jeong M, Ko BC (2020) Is the surrogate model interpretable? In: NeurIPS workshops. pp 1–5
11. Kim SJ, Kwak SY, Ko BC (2019) Fast pedestrian detection in surveillance video based on soft target training of shallow random forest. IEEE ACCESS 7:12415–12426
12. Breiman L (2001) Random forest. Springer Mach Learn 45:5–32
13. Friedman J (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232
14. Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: 22nd ACM SIGKDD International conference on knowledge discovery and data mining. pp 785–794
15. Dorogush AV, Ershov V, Gulin A (2018) CatBoost: gradient boosting with categorical features support. arXiv:1810.11363
16. Lundberg SM et al (2020) From local explanations to global understanding with explainable AI for trees. Nature Mach Intell 2:56–67
17. Shapley LS (1953) A value for n-person games. In: Contributions to the theory of games, vol 2, pp 307–317
18. Dua D, Graff C (2019) UCI Machine learning repository
19. Olson RS et al (2017) PMLB: a large benchmark suite for machine learning evaluation and comparison. BioData mining 10:1–13
20. Erickson N et al (2020) AutoGluon-tabular: robust and accurate automl for structured data. arXiv:2003.06505
21. Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Mach Learn 63:3–42
22. Wilson DL (1972) Asymptotic properties of nearest neighbor rules using edited data. IEEE Trans Syst Man Cybern 3:408–421

23. Cortes C, Vapnik VN (1995) Support-vector networks. Mach Learn 20:273–297
24. Zhou ZH, Feng J (2017) Deep forest: towards an alternative to deep neural networks. arXiv:1702.08835
25. Freund Y, Schapire R (1995) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55:119–139
26. Ke G et al (2017) Lightgbm: A highly efficient gradient boosting decision tree. In: NeurIPS, pp 3146–3154
27. Duan T et al (2020) Ngboost: Natural gradient boosting for probabilistic prediction. In: ICML, pp 2690–2700
28. Kokel H et al (2020) A unified framework for knowledge intensive gradient boosting: leveraging human experts for noisy sparse domains. In: AAAI. pp 4460–4468