IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 8 December 2022, accepted 26 December 2022, date of publication 29 December 2022, date of current version 6 January 2023.

Digital Object Identifier 10.1109/ACCESS.2022.3233194

RESEARCH ARTICLE

Random Interaction Forest (RIF)–A Novel Machine Learning Strategy Accounting for Feature Interaction

CHAO-YU GUO AND YI-JYUN LIN

Division of Biostatistics and Data Science, Institute of Public Health, College of Medicine, National Yang Ming Chiao Tung University, Taipei 112304, Taiwan

Corresponding author: Chao-Yu Guo (cyguo@nycu.edu.tw)

This work was supported by the National Science and Technology Council under Grant 111-2118-M-A49-005.

ABSTRACT If an interaction exists in medical and health sciences, a proper statistical approach is required to avoid an erroneous conclusion. For example, different genders may introduce modified therapeutic effects of drugs, or an adverse interaction between two medicines changes the pharmacological activity, reduces the therapeutic effect, or induces toxicity. Therefore, if the analysis does not account for the impact of the interaction, it may introduce significant prediction errors or bias. Regression models deal with a two-way interaction by adding the product of the two interactive variables. Since machine learning models demonstrate a superior predictive ability to regression models, this study proposes a new method based on the random forest to account for interaction, called random interaction forest (RIF). This new strategy modifies the structure of the random forest, where the interaction features are forced to be in the first two nodes. Simulation studies examined the predictive ability of the linear regression model, logistic regression model, random forest, and the RIF under various scenarios. The results showed that the RIF consistently outperforms random forest and logistic regression when interactions are present. The RIF also performs better in many scenarios than the linear regression model. When the effect of interaction is more significant, the performance of RIF could be superior.

INDEX TERMS Interaction, random forest, linear regression, logistic regression, machine learning.

I. INTRODUCTION

A. THE DEVELOPMENT OF MACHINE LEARNING MODELS

The application of machine learning models has been booming and favored in many research fields [1]. Previous research [2], [5] showed that novel machine learning models have better predictions or performance than traditional statistical approaches in various situations. In particular, Guo and Chang [5] recently proposed a novel algorithm, the Extreme Gradient Boosting Machine for Feature Interaction (XGB-FI), to find big data's most significant feature interaction, which outperforms the conventional statistical model.

The supervised machine learning method makes the tree-based model easy to understand and highly interpretable. The tree model uses a series of if-else rules to generate prediction results from one or more decision trees.

The decision tree is a supervised machine learning model with simple logic, intuition, and high execution efficiency that is applicable for both regressions with a continuous outcome and classification with a categorical dependent variable. One of the most popular algorithms is Classification and Regression Trees (CART). CART belongs to a binary classification tree structure and is the basis for establishing a random forest [6].

However, the decision tree algorithm can easily cause overfitting. As a result, the random decision forest Ho [7] and random forest [8] avoided overfitting issues without the need to prune the trees. Random forest is an ensemble method. The ensemble method collects multiple weak classifiers to create a strong classifier [9]. Hence, the random forest is a supervised machine learning model composed of numerous

The associate editor coordinating the review of this manuscript and approving it for publication was Joey Tianyi Zhou.

CART decision trees. To build CART trees, the technique used is BAGGING (Bootstrap aggregating) [10]. The random forest performs *split-variable randomization* where each time a split is to be performed, the search for the split variable is limited to a random subset of ($m = \sqrt{p}$). Note that p is the number of all variables in the original training set

If it is a classification problem, the random forest will summarize the prediction results of all CART trees and determine the final classification result by majority voting. When the output is continuous, the random forest calculates the average of all CART tree predictions as the predicted value. Out-of-bag error evaluates the performance of the random forest.

B. THE ISSUE OF FEATURE INTERACTION

In research, an interaction could alter the results or introduce biases. For example, genders may show different therapeutic effects of the same medicine. Interactions can be seen everywhere in life, such as gastrointestinal discomfort by eating crabs and persimmons simultaneously because crabs are rich in protein and persimmons are rich in tannins. In particular, recent discoveries of interaction impact on heart diseases draw significant attention [11], [14]. Therefore, if the analysis does not adequately handle the interaction effect, it could result in prediction errors.

The conventional statistical method for the interaction problem is the regression model [15]. A coefficient for estimating the multiplication of two interaction variables is added to the regression model. If the influence of one independent variable (x_1) on the dependent variable (y) is affected by other independent variables (x_2), then it is said that there is an interaction between (x_1, x_2) [16]. If this coefficient is statistically significant, one of the two variables modifies the effect between the other variable and the outcome of interest. When the interaction is known, we could replace the interacting features with a single ''engineered'' feature that reflects the interaction in machine learning strategies, which is identical to the conventional regression model that adds the interaction term in the predictors.

Tree models contain the partial interaction effect due to their design since the decision tree structure is composed of top-to-bottom recursive branching rules. This rule allows the decision tree to be hierarchical. The method considers the interaction between variables. Specifically, when a variable is selected as an internal node, if the two branches after the split have different behaviors in the subsequent selection of variables, there may be information indicating an interaction between the variables.

However, Wright, et al. [17] suggested that the tree model could not initially pick two interaction variables in simulating multiple interactions. It is an internal node, especially when the marginal effect of the interaction variable is small. Although the tree model can handle a partial interaction effect, such modification is easily affected by the marginal effect. In addition, the random forest is composed of many decision trees. When building the trees, the random forest only selects some variables as the nodes. As a result, each tree may not include the interaction variables, and the prediction could be biased.

Besides tree-based models, numerous machine learning strategies exist, such as the support vector machines (SVM) [3], [18]. The SVM is more sensitive to missing data and efficient in processing big data. Artificial Neural Networks (ANN) [19] provide a satisfying predictive ability with a complicated structure. The ANN outperforms the logistic regression [20] and could apply to traffic predictions [21], incidence clearance [22], and environmental research [23]. The most critical problem of ANN is the unexplained behavior of the network.

C. THE RESEARCH MOTIVATION

The ANN does not explain why and how when it produces a probing solution. Since machine learning outperforms the conventional statistical models in many situations and the random forest has the most suitable structure to deal with interactions, we choose the random forest as the foundation to develop a new interaction model.

After the XGB-FI [5] determines which variables introduce feature interaction, the subsequent step is to perform the analysis that properly accounts for the feature interaction. This research aims to establish a tree-based model with better predictive performance under the interaction impact, named the random interaction forest (RIF). The programming language is R Software (4.0.3) [24], which uses the (mvrnorm) function in the MASS package to generate simulation data, and extends the ''randomForest'' package to construct the RIF model. This research also aims to provide a free R code to implement the RIF approach.

II. METHODS

In contrast to the random forest that randomly selects some variables to create the trees, the RIF targets the interaction in the earliest stage. Therefore, the RIF employed a restricted structure that forces the first and second nodes to be the two variables (also known as features) introducing the interaction. If the interaction variable is continuous, the median is the cutoff point. We assessed various cutoffs such as the mean, the first quartile, the third quartile, or extreme values. However, we discovered that the median is the optimal choice since it is robust to skewed distributions and has the best overall performance.

After the data passes through the first two nodes, the four leave nodes consist of the four subsets of original data according to the four combinations of the first two nodes. In other words, this step stratifies the data into four subsets, removing the interaction's impact. The random forests are implemented independently within each of the four leaf nodes. The prediction of the RIF is the bagged results from the four random forests.

Analysis flow:

1. Determine the two interaction features X_1 and X_2

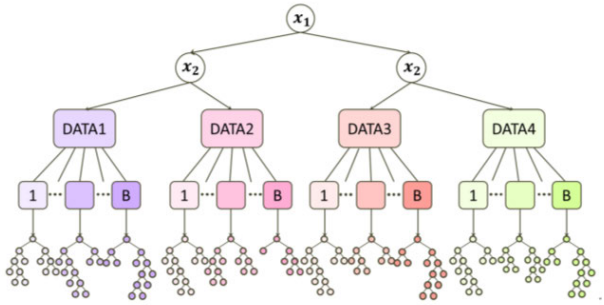

FIGURE 1. Data Structure of the Random Interaction Forest (RIF).

2. Let X_1 be the first node and X_2 be the second node (the order of the two feature interaction does not alter the results)

3. Generate random forests within the four strata (DATA1, DATA2, DATA3, and DATA4) with B bootstrapped samples

4. Bagging the predictive results from all observations

Figure 1 reveals the model structure of the RIF with two-way interaction. The extension to a three-way or other higher-order interaction is straightforward. The RIF would have more restricted nodes in the earliest stage when creating the trees for a higher-order interaction. For example, the RIF deals with a three-way interaction with the first three nodes forced to be the three features introducing the interaction. Therefore, the n-th order interaction has the top n nodes forced into the restricted structure.

Since the RIF is an extension of the random forest with a restricted structure, the RIF could deal with both categorical and continuous outcomes. If the output variable is continuous, the average value of the four branches is the final predicted value. In contrast, if the output is a dichotomous value, the majority voting by all leaf nodes determines the final classification. The RIF avoids the probability that the two interaction variables may not be selected as the closest nodes in the earliest stage.

To evaluate the performance of the RIF, we compare the predictive ability between the RIF, random forests, and statistical approaches using linear regression for the continuous outcome or logistic regression for a dichotomous variable in different settings.

For a continuous output, the root mean square error (RMSE) is used for evaluating the performance. We define two ratios of the RMSE according to two different comparisons. The RMSE Ratio1 is for the comparison between the new model and the random forest as the following: $RMSE\ Ratio1 = \frac{RMSE\ of\ random\ forest}{RMSE\ of\ RIF}$. The RMSE Ratio2 is for the linear regression and RIF. $RMSE\ Ratio2 = \frac{RMSE\ of\ linear\ regression}{RMSE\ of\ RIF}$. When the RMSE Ratio is greater than 1, then the prediction of the RIF is better.

If the output is dichotomous, we compare the accuracy between the two models. The Accuracy Ratio1 compares the RIF to the random forest, which is defined as $Accuracy\ Ratio1 = \frac{Accuracy\ of\ RIF}{Accuracy\ of\ random\ forest}$. The Accuracy Ratio2 compares the RIF to the logistic regression, which is defined as $Accuracy\ Ratio2 = \frac{Accuracy\ of\ RIF}{Accuracy\ of\ logistic\ regression}$. If the Accuracy Ratio is greater than 1, then the prediction of the RIF is better.

TABLE 1. Scenarios for the simulation study.

	Y	x1	x2
Scenario1	continuous	continuous	continuous
Scenario2	continuous	continuous	dichotomous
Scenario3	continuous	dichotomous	dichotomous
Scenario4	dichotomous	continuous	continuous
Scenario5	dichotomous	continuous	dichotomous
Scenario6	dichotomous	dichotomous	dichotomous

A. SIMULATION STUDY

For simplicity, the computer simulations examined only the two-way interaction. The simulated data comprises one dependent variable (Y) and ten independent variables (X). The independent variables (X) include two variables with interaction (X_1 and X_2), three variables with marginal effects (X_3, X_4, and X_5), and five noise variables ($X_6 - X_{10}$) that are independent of Y and other predictors ($X_1 - X_5$). The outcome (Y) and the two interaction variables (X_1 and X_2) could be continuous or categorical. Table 1 displays the six scenarios for the combination of data types we examined. The sample sizes are 500 and 1000.

Following a similar simulation scheme of the XGB-FI [5], we assume that Y and X follow a multivariate normal distribution and simulate Scenario1. The covariance matrix determines the relationship between Y and X. Figure 2 shows five structures (case1 to case5) adopted for Y and X_1 to X_5. The five noise variables ($X_6 - X_{10}$) are uncorrelated with any variable in the dataset. Therefore, the covariance matrix did not include the five noises. Case1 assumes a high correlation between the two interaction features, and the correlation between X_1 (X_2) and Y is 0.3 (0.5). Case2 increases the marginal effect of X_3-X_5 to Y by changing the correlation from 0.2 to 0.6. The correlation between X_1 (X_2) and Y was reduced to 0.1 (0.2) in Case2. Case3 allows different marginal effects of X_3-X_5 to Y (0.3, 0.1, 0.6). Case4 assumes an equal correlation between X_1-X_5 and Y (0.2). Case5 also assumes an equal correlation between X_1-X_5 and Y, but with a higher value than 0.2.

Regarding the data of Scenario2, we dichotomize one of the continuous features with interaction (X_1 or X_2) in Scenario1. Note that the median is the cutoff point. If the original values of X_1 or X_2 are higher than the median, then the new interaction variable is coded as 1 (event). Otherwise, the new interaction variable would represent normal subjects with 0.

If the two interaction variables (X_1 and X_2) are dichotomous, Y is generated as four groups according to the four combinations of the two interaction variables. Table 2

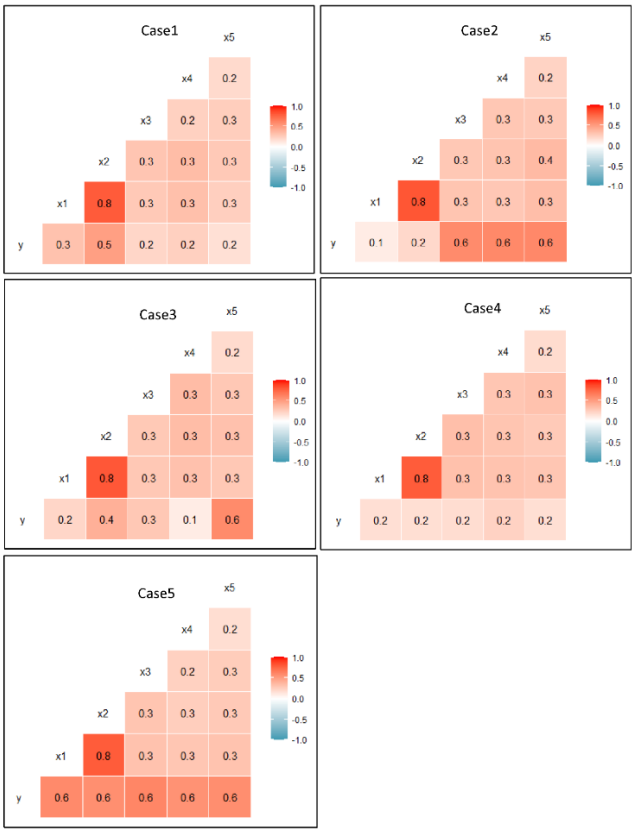

FIGURE 2. Correlation matrix for case1 to case5.

TABLE 2. Distribution of Y for scenario3.

4 leaf nodes	No interaction	Weak interaction	Strong interaction
1 (x1 = 1 & x2 = 1)	Normal(10, 1)	Normal(6, 1)	Normal(1, 1)
2 (x1 = 1 & x2 = 0)	Normal(10, 3)	Normal(10, 3)	Normal(10, 3)
3 (x1 = 0 & x2 = 1)	Normal(10, 5)	Normal(10, 5)	Normal(10, 5)
4 (x1 = 0 & x2 = 0)	Normal(10, 1)	Normal(0, 1)	Normal(0, 1)

displays the normal distribution parameters with different means and standard deviations for Scenario3. Various parameter settings represent different intensities of interaction effects. The interaction effect is absent when the means of Y are identical in the four leaf nodes. For the weak interaction, the mean of the interaction group (the first leaf node) is reduced from 10 to 6. Lastly, the more substantial interaction effect further reduces the mean from 10 to 1. Note that the weak interaction accounts for 40% of the variance in Y. These scenarios assess whether the model will have a better prediction performance as the interaction effect increases.

When the dependent variable (Y) is categorical, the simulation process is similar to the continuous scenario, except that the Y is generated after simulating X. The independent variables include two variables with interaction, three marginal effect variables, and five confounding variables. A multivariate normal distribution generates X. Four Covariance Matrixes (case6 to case9) are simulated

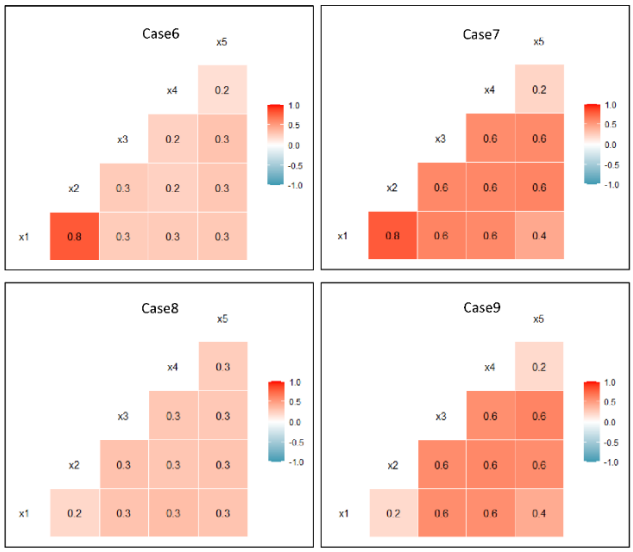

FIGURE 3. Correlation matrix for case6 to case9.

TABLE 3. Scenario4∼scenario6, probability of the event of Y.

Four leaf nodes	No interaction	Weak interaction	Strong interaction
1 (x1 = 1 & x2 = 1)	0.2	0.6	0.1
2 (x1 = 1 & x2 = 0)	0.2	0.4	0.6
3 (x1 = 0 & x2 = 1)	0.2	0.4	0.6
4 (x1 = 0 & x2 = 0)	0.05	0.05	0.1

(Figure 3). Case6 assumes a high correlation between the two interaction features. In Case7, we increase the correlation between X_1 to X_5. Case8 reduces all correlation coefficients. Lastly, Case9 has a weak correlation between X_1 and X_2, but the others are high.

We assign the four leaf nodes (data1 to data4, as shown in Figure 1) with different binomial distribution probabilities. The three scenarios are no interaction, weak interaction, and strong interaction (Table 3). According to the binomial distribution, the value indicates the probability of assigning Y to be the event. There is no interaction effect when the first three leaf nodes have the same probability (0.2). There is weak interaction if the first leaf node (0.6) probability is slightly higher than the second and third leaf nodes (0.4). Strong interaction assumes that the first leaf node (0.1) probability is lower than the second and third leaf node (0.6).

In addition to simulations, we applied the RIF to the "Early Stage Diabetes Risk Prediction Dataset" from the University of California, Irvine (UCI) machine learning Repository. This data was collected from a direct questionnaire of patients from the Diabetes Hospital in Sylhet, Bangladesh. It contains a total of 520 people with diabetes. Related symptoms are in the reference, of which 320 people have diabetes, and 200 do not [25]. The first step adopted a logistic regression model to discover whether there is an interaction between the two variables. In the second step, the two variables with the

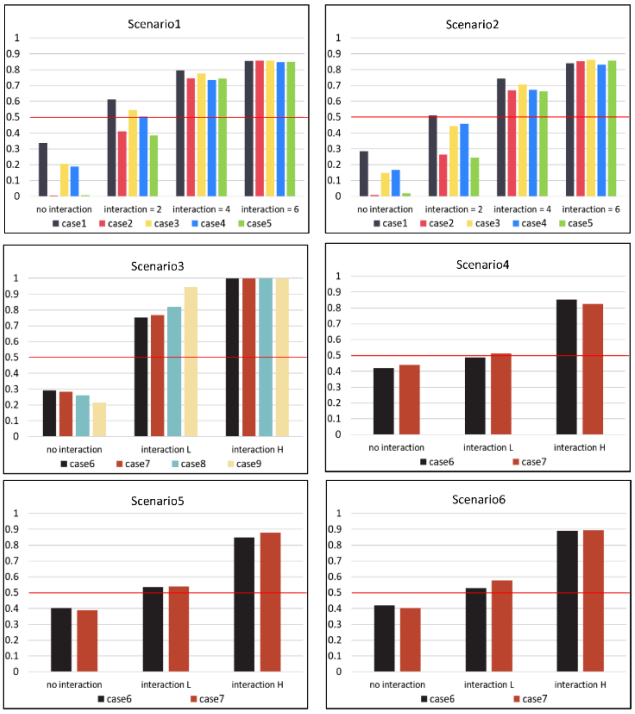

FIGURE 4. Probability of RIF is better than RF for the continuous (Scenario1-3) and dichotomous (Scenario4-6) outcomes.

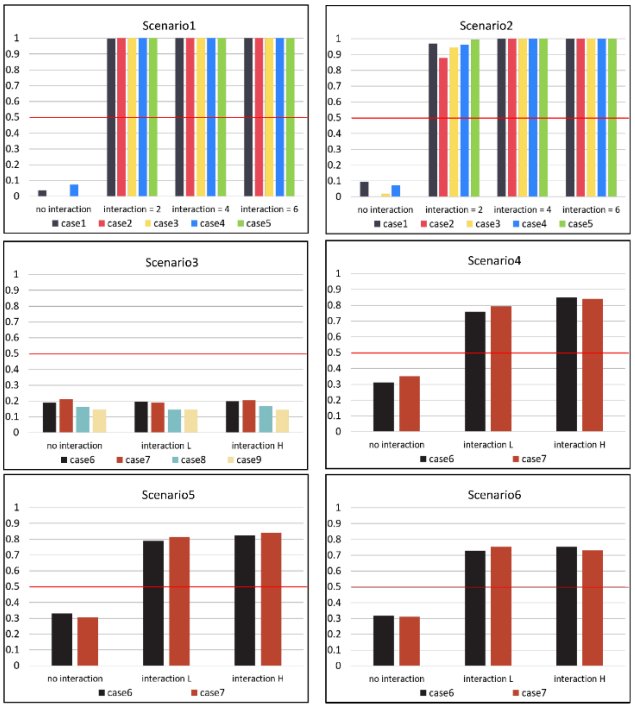

FIGURE 5. Probability of RIF is better than Linear (Scenario1-3) or Logistic Regression (Scenario4-6).

most significant interaction effects (Gender and Alopecia) are selected for the restricted structure of the RIF in the first two nodes. The interaction effect (Age:Alopecia1) is moderate since the p-value of the interaction term is the 7th significant predictor (details in the Appendix, Table 5).

We randomly split the data into 80% training and 20% testing to evaluate predictive ability. We repeated this process ten times and recorded the results. We could conclude which method is superior, and the random split does not bias the conclusion. Here we use the accuracy rate as the evaluation with ten repetitions.

III. RESULTS

All simulations had 1000 repetitions with 80% training and 20% testing data. In Figure 4, the probability of each bar is the number of times the Accuracy Ratio or RMSE Ratio is higher than 1. Therefore, we translate this probability into the number of times the RIF predicts better than a random forest with or without the interaction effect. In the horizontal axis, interaction = 2, 4, or 6 indicates the elevated mean value of Y due to the interaction effect for scenario1 and scenario2. The superiority performance of the RIF increases with the magnitude of interaction.

For an estimated probability over 54%, we could claim that the RIF significantly outperforms other strategies at the significance level of 5% since the 95% confidence interval of p=0.54 is (0.509, 0.571). This confidence interval does not contain the null value of 50%, which means the RIF outperforms others by chance. In contrast, if the probability is under 46%, the RIF is significantly inferior to others because the 95% confidence interval of p=0.46 is (0.429, 0.491). Most situations in Figure 4 indicate that the RIF significantly outperforms others with interaction.

The median of RMSE Ratio1 and Accuracy Ratio1 show similar comparisons (Figure 6). The more substantial interaction results in a larger median of RMSE Ratio1 and Accuracy Ratio1. Therefore, the RIF outperforms the random forest under the impact of the interaction.

In Figure 5, the RIF is compared with regression models. The RIF is better than logistic regression in the situations we simulated. However, regardless of the magnitude of the interaction, the RIF cannot outperform the regression model under Scenario3. The nature of linear regression with two dichotomous variables may be the optimal model for tackling the interaction. However, this conclusion requires deliberate theoretical research and forms a great future work. Therefore, we recommend the conventional linear regression model when two dichotomous features introduce the interaction.

The difference between case6 and case7 is minor, which means that the correlation between marginal effects does not change the results much. In scenario1 and scenario2 of Figure 7, the median RMSE Ratio1 and Ratio2 for case2 and case5 are higher than in other cases. This result reveals that the RIF performs better with the marginal effect than the linear regression. Therefore, more significant marginal effects would prevent the linear regression from obtaining unbiased results under scenario1 and scenario2, but the RIF could tackle such impact properly.

In addition, increasing the number of noise variables from 5 to 10 enhances the prediction ability of RIF compared

TABLE 4. Application to the real data.

Repetition	RIF	RF	Logistic
1	0.9711538	0.9519231	0.8653846
2	1	1	0.9038462
3	0.9903846	0.9807692	0.9423077
4	0.9711538	0.9807692	0.9230769
5	0.9711538	0.9519231	0.8942308
6	0.9903846	1	0.9230769
7	0.9326923	0.9423077	0.9038462
8	0.9903846	0.9807692	0.9326923
9	1	1	0.9711538
10	0.9903846	0.9615385	0.9326923

to random forests (Figure 8). The reason is that more noises would prevent the random forests from employing the two interaction variables as the closest nodes in the early stage. As a result, the random forest generates a higher prediction error under the impact of the interaction in this situation. Compared with the regression model, the RIF also shows a better prediction. With the 6-fold interaction effect, the prediction of the RIF is three times better than the regression model.

The simulation results show that under the impact of interaction, regardless of whether the outcome variable is continuous or categorical, the prediction ability of the RIF is higher than that of the random forest. As the interaction effect increases, the prediction ability of the RIF is higher. The RIF consistently outperforms the logistic regression model for a dichotomous outcome under the interaction impact. For a continuous measure, the RIF also demonstrates better results than the linear regression model in most situations with interaction. All methods in the sample size of 1000 have higher predictive ability than the smaller sample size of 500 (Figure 9).

In the real-life application of the RIF, the accuracy of the three models is displayed in Table 4. Among the ten repetitions, the RIF has the highest accuracy in seven repetitions. The results are consistent with the simulations of scenario6 with a weak interaction.

IV. DISCUSSION

This research proposes a new machine learning model, random interaction forest (RIF), which extends the random forest to a restricted structure in the first few nodes by a known interaction effect. Although we studied two-way interactions in computer simulations, the extension to a three-way or higher-order interactions is straightforward. The RIF is superior to random forest or regression models under the impact of the interaction. The higher interaction effect results in a better prediction of RIF than the other two strategies.

Although the RIF has slightly inferior performance than the random forest when the data does not show any sign of

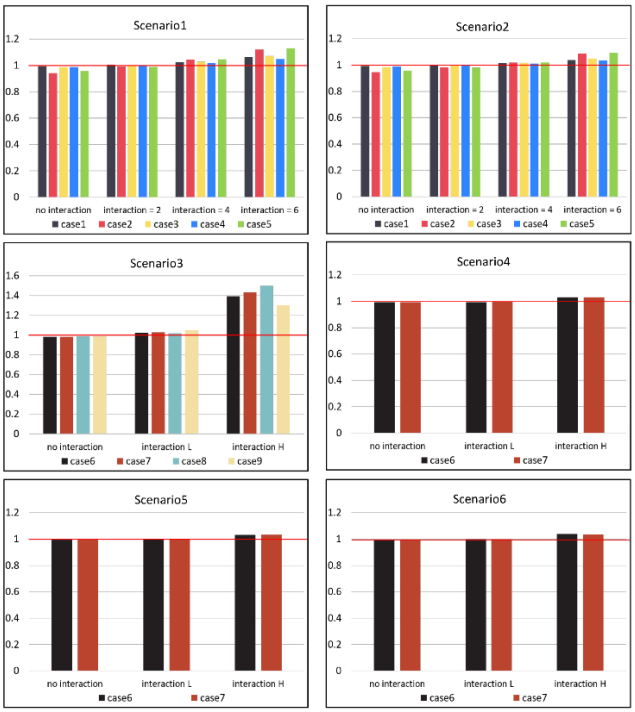

FIGURE 6. Median of RMSE Ratio1 (Scenario1-3) and Accuracy Ratio1 (Scenario4-6).

interaction, under scenario1, the RIF revealed approximately 6% loss compared to the random forest. However, the RIF outperforms the random forest in most scenarios with about 13% gains under the impact of the interaction. We recommend the RIF for more accurate results if the interaction effect is suspectable or observed.

In this simulated data, the median of the interaction variable is used as the threshold, and the interaction effect is added to the corresponding strain number. In the structure of the RIF, the median is also used as the cut point. However, there are various real-life data, and the median may not necessarily be the optimal cut point. The future study could examine if the RIF's threshold of the interaction variable node can be adjusted according to the data.

A freely available R code with example files could effortlessly implement the random interaction forest in the supplementary materials, and this new machine could be applied in numerous research fields. The hyperparameter setting of the RIF is the same as the default setting in the random forest package. m $= \sqrt{p}$ and 500 trees in the forest

Research topics in the K-Nearest Neighbors (KNN), SVM [3], XGBoost Machine [2], [5], and ANN [23] revealed that each method has pros and cons concerning efficiency, accuracy, and feasibility in various settings. In particular, the random forest has an excellent performance in missing data imputation. The ''missForest'' imputation is non-parametric missing value imputation using the random forest [4], [26]. However, the ''missForest'' imputation assumes no feature interaction. Since the RIF is a restricted model based on the random forest, the RIF could also develop another imputation

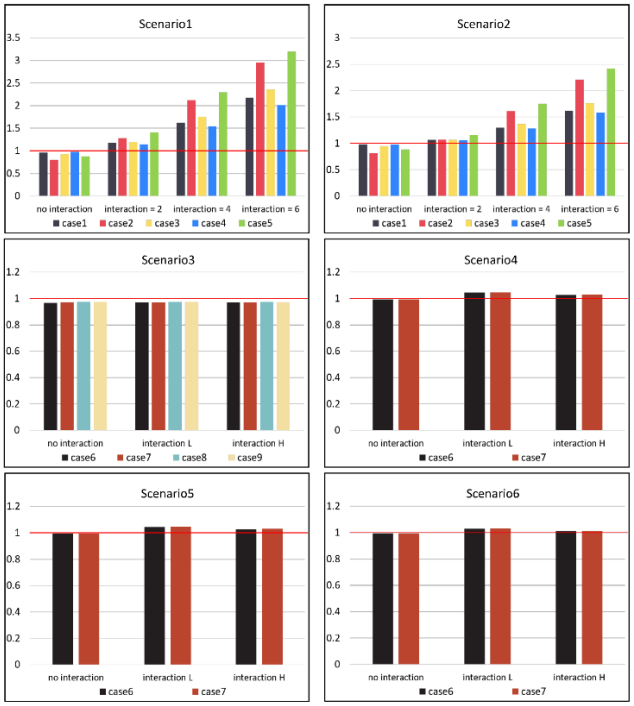

FIGURE 7. Median of RMSE Ratio2 (Scenario1-3) and Accuracy Ratio2 (Scenario4-6).

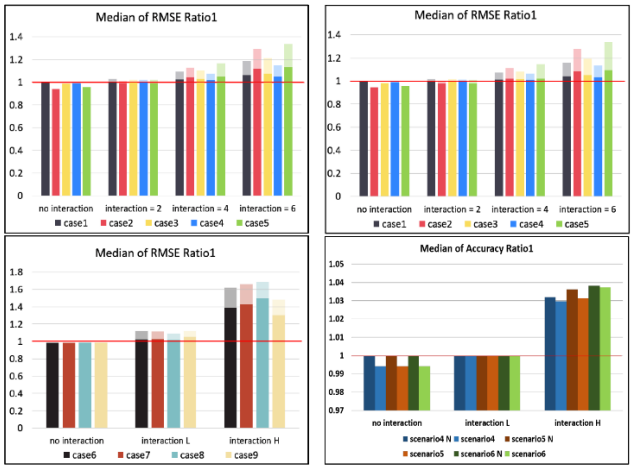

FIGURE 8. Under Scenario1 (top left), Scenario2 (top right), Scenario3 (bottom left), Scenario4-6 (bottom right), Changes of Median of RMSE Ratio1 and Accuracy Ratio 1 after adding more variables with noise. The light color represents the changes.

strategy in future works. Besides, comprehensive research that compares all machine learning strategies [23] under the impact of feature interaction is desired.

V. CONCLUSION

This research proposes a novel machine learning strategy to tackle interaction's impact in medicine, health sciences, and all research fields. According to simulation studies with interaction impact, the RIF generally outperforms the random forest, linear, and logistic regression models. The only exception is the continuous outcome with two dichotomous features introducing the interaction when the linear regression model performs best. A desirable future work extends the RIF to accommodate missing data imputation.

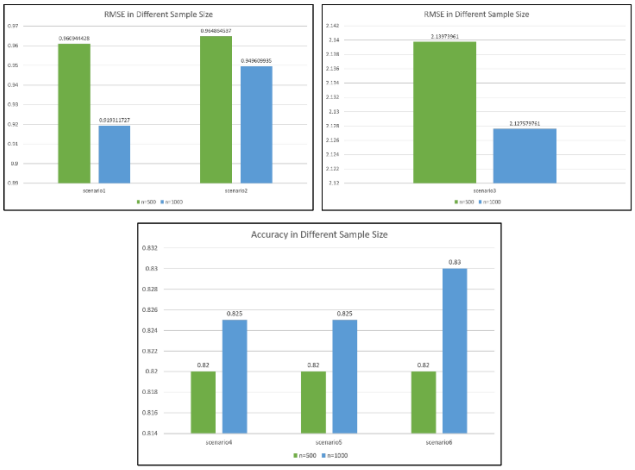

FIGURE 9. RMSE in the sample size of 500 and 1000.

TABLE 5. Signif. codes: 0 '∗∗∗' 0.001 '∗∗' 0.01 '∗' 0.05 '.' 0.1 ' ' 1.

Coefficients:					
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.258209	1.322053	0.195	0.845151	
Age	0.009269	0.032075	0.289	0.772605	
Gender1	-4.756019	0.684836	-6.945	3.79E-12	***
PolyuriaYes	4.89864	0.784407	6.245	4.24E-10	***
PolydipsiaYes	5.378	0.899976	5.976	2.29E-09	***
sudden.weight.lossYes	0.190014	0.56265	0.338	0.735579	
weaknessYes	0.622764	0.55466	1.123	0.261528	
PolyphagiaYes	1.257745	0.525117	2.395	0.016613	*
Genital.thrushYes	2.067883	0.600639	3.443	0.000576	***
visual.blurringYes	1.296656	0.657755	1.971	0.048686	*
ItchingYes	-3.255807	0.748511	-4.35	1.36E-05	***
IrritabilityYes	2.82751	0.664493	4.255	2.09E-05	***
delayed.healingYes	-0.577709	0.585352	-0.987	0.323671	
partial.paresisYes	1.194119	0.512386	2.331	0.019779	*
muscle.stiffnessYes	-0.483334	0.582392	-0.83	0.406589	
Alopecia1	7.029956	2.437479	2.884	0.003925	**
ObesityYes	-0.357444	0.558253	-0.64	0.521984	
Age:Alopecia1	-0.137432	0.046232	-2.973	0.002952	**

VI. DECLARATIONS

Conflicts of interest: The authors declare no conflicts of interest related to this article's subject matter or materials.

Ethics approval: Our study did not require ethical board approval because It is a computer simulation study.

Guarantor: Not applicable.

Consent for publication: Not applicable. It is a computer simulation study.

Contributorship:

Chao-Yu Guo proposed the research concept, supervised the project, and wrote the manuscript.

Yi-Jyun Lin conducted the analysis and prepared figures and tables. All authors read and approved the final manuscript.

APPENDIX

See Figures 6–9 and Table 5.

ACKNOWLEDGMENT

None

REFERENCES

[1] T. M. Mitchell, "Machine learning," Tech. Rep., 1997.

[2] C.-Y. Guo, M.-Y. Wu, and H.-M. Cheng, "The comprehensive machine learning analytics for heart failure," *Int. J. Environ. Res. Public Health*, vol. 18, no. 9, p. 4943, May 2021, doi: 10.3390/ijerph18094943.

[3] C. Y. Guo and Y. C. Chou, "A novel machine learning strategy for model selections—Stepwise support vector machine (StepSVM)," *PLoS One*, vol. 15, no. 8, 2020, Art. no. e0238384, doi: 10.1371/journal.pone.0238384.

[4] C.-Y. Guo, Y.-C. Yang, and Y.-H. Chen, "The optimal machine learning-based missing data imputation for the cox proportional hazard model," *Frontiers Public Health*, vol. 9, Jul. 2021, Art. no. 680054, doi: 10.3389/fpubh.2021.680054.

[5] C.-Y. Guo and K.-H. Chang, "A novel algorithm to estimate the significance level of a feature interaction using the extreme gradient boosting machine," *Int. J. Environ. Res. Public Health*, vol. 19, no. 4, p. 2338, Feb. 2022. [Online]. Available: https://www.mdpi.com/1660-4601/19/4/2338

[6] L. Breiman, *Classification and Regression Trees*. 1984.

[7] T. K. Ho, "Random decision forests," in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, vol. 1, 1995, pp. 278–282.

[8] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[9] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits Syst. Mag.*, vol. 6, no. 3, pp. 21–45, 3rd Quart., 2006.

[10] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[11] B. Schnegg, D. Robson, M. Fürholz, T. Meredith, C. Kessler, S. H. Baldinger, and C. Hayward, "Importance of electromagnetic interactions between ICD and VAD devices—Mechanistic assessment," *Artif. Organs*, vol. 46, no. 6, pp. 1132–1141, Jun. 2022, doi: 10.1111/aor.14167.

[12] H. Kawashima, P. W. Serruys, H. Hara, M. Ono, C. Gao, R. Wang, S. Garg, F. Sharif, R. J. de Winter, M. J. Mack, D. R. Holmes, M.-C. Morice, A. P. Kappetein, D. J. F. M. Thuijs, M. Milojevic, T. Noack, F.-W. Mohr, P. M. Davierwala, and Y. Onuma, "10-year all-cause mortality following percutaneous or surgical revascularization in patients with heavy calcification," *JACC, Cardiovascular Intervent.*, vol. 15, no. 2, pp. 193–204, Jan. 2022, doi: 10.1016/j.jcin.2021.10.026.

[13] J. P. Curtain, A. M. Jackson, L. Shen, P. S. Jhund, K. F. Docherty, M. C. Petrie, D. Castagno, A. S. Desai, L. E. Rohde, M. P. Lefkowitz, J. Rouleau, M. R. Zile, S. D. Solomon, K. Swedberg, M. Packer, and J. J. V. Mcmurray, "Effect of sacubitril/valsartan on investigator-reported ventricular arrhythmias in PARADIGM-HF," *Eur. J. Heart Failure*, vol. 24, no. 3, pp. 551–561, Mar. 2022, doi: 10.1002/ejhf.2419.

[14] G. D. Pinna, E. Robbi, C. Bruschi, M. T. La Rovere, and R. Maestri, "Interaction between arousals and ventilation during Cheyne–Stokes respiration in heart failure patients: Insights from breath-by-breath analysis," *Frontiers Med.*, vol. 8, Dec. 2021, Art. no. 742458, doi: 10.3389/fmed.2021.742458.

[15] K. J. Preacher, P. J. Curran, and D. J. Bauer, "Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis," *J. Educ. Behav. Statist.*, vol. 31, no. 4, pp. 437–448, 2006.

[16] P. D. Allison, "Testing for interaction in multiple regression," *Amer. J. Sociol.*, vol. 83, no. 1, pp. 144–153, Jul. 1977.

[17] M. N. Wright, A. Ziegler, and I. R. König, "Do little interactions get lost in dark random forests?" *BMC Bioinf.*, vol. 17, no. 1, pp. 1–10, Dec. 2016.

[18] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Oct. 1995.

[19] M. Anthony and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[20] M. H. Hassoun, *Fundamentals of Artificial Neural Networks*. Cambridge, MA, USA: MIT Press, 1995.

[21] J. Tang, F. Liu, Y. Zou, W. Zhang, and Y. Wang, "An improved fuzzy neural network for traffic speed prediction considering periodic characteristic," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 9, pp. 2340–2350, Sep. 2017.

[22] J. Tang, L. Zheng, C. Han, W. Yin, Y. Zhang, Y. Zou, and H. Huang, "Statistical and machine-learning methods for clearance time prediction of road incidents: A methodology review," *Anal. Methods Accident Res.*, vol. 27, Sep. 2020, Art. no. 100123.

[23] C.-Y. Guo, T.-W. Liu, and Y.-H. Chen, "A novel cross-validation strategy for artificial neural networks using distributed-lag environmental factors," *PLoS ONE*, vol. 16, no. 1, Jan. 2021, Art. no. e0244094, doi: 10.1371/journal.pone.0244094.

[24] *R: A Language and Environment for Statistical Computing*, R Core Team, R Found. Stat. Comput., Vienna, Austria, 2014. [Online]. Available: http://www.R-project.org/

[25] M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood prediction of diabetes at early stage using data mining techniques," in *Computer Vision and Machine Intelligence in Medical Image Analysis*. Springer, 2020, pp. 113–125.

[26] D. J. Stekhoven and P. Buhlmann, "MissForest-non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, Jan. 2012.

CHAO-YU GUO received the Ph.D. degree in biostatistics from the Department of Biostatistics, School of Public Health, Boston University, USA, in 2004.

From 2004 to 2007, he was appointed as a Research Assistant Professor at the Department of Mathematics and Statistics, Boston University. From 2007 to 2010, he was an Assistant Professor with the joint appointment of Children's Hospital Boston and Harvard Medical School. He is currently an Associate Professor with the Biostatistics and Data Science, National Yang Ming Chiao Tung University. He has published 92 research articles available in PubMed.

Dr. Guo is also a Guest Editor of the *International Journal of Environmental Science and Pollution Research* (2021 Impact factor 4.614) in the Special Issue for Machine Learning Analytics for Cardiovascular Diseases. He is also a Guest Editor of *Symmetry* (2021 Impact Factor 2.94).

YI-JYUN LIN received the M.S. degree from the Division of Biostatistics and Data Science, National Yang Ming Chiao Tung University, in 2021.

From 2019 to 2021, she was a Teaching Assistant for biostatistics courses at the Graduate School and taught the R software coding lectures. She is currently a Professional Statistician in high-tech company, Tainan, Taiwan.