Pattern Recognition 120 (2021) 108192

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

GAMI-Net: An explainable neural network based on generalized additive models with structured interactions

Zebin Yang[a], Aijun Zhang[b,*], Agus Sudjianto[b]

[a] Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong
[b] Corporate Model Risk, Wells Fargo, USA

ARTICLE INFO

Article history:
Received 4 November 2020
Revised 28 June 2021
Accepted 4 July 2021
Available online 20 July 2021

Keywords:
Explainable neural network
Generalized additive model
Pairwise interaction
Interpretability constraints

ABSTRACT

The lack of interpretability is an inevitable problem when using neural network models in real applications. In this paper, an explainable neural network based on generalized additive models with structured interactions (GAMI-Net) is proposed to pursue a good balance between prediction accuracy and model interpretability. GAMI-Net is a disentangled feedforward network with multiple additive subnetworks; each subnetwork consists of multiple hidden layers and is designed for capturing one main effect or one pairwise interaction. Three interpretability aspects are further considered, including a) sparsity, to select the most significant effects for parsimonious representations; b) heredity, a pairwise interaction could only be included when at least one of its parent main effects exists; and c) marginal clarity, to make main effects and pairwise interactions mutually distinguishable. An adaptive training algorithm is developed, where main effects are first trained and then pairwise interactions are fitted to the residuals. Numerical experiments on both synthetic functions and real-world datasets show that the proposed model enjoys superior interpretability and it maintains competitive prediction accuracy in comparison to the explainable boosting machine and other classic machine learning models.

1. Introduction

Deep learning is one of the leading techniques in artificial intelligence (AI). Despite its great success, a fundamental and unsolved problem is that the working mechanism of deep neural networks is hardly understandable. Without sufficient interpretability, it would be risky to apply these AI systems in real-life applications. A well-trained deep neural network is known to usually have accurate predictive performance on the data at hand. However, it may perform abnormally as the data is slightly changed, as its inner decision-making process is unknown. Some recent examples can be referred to the adversarial attacks, where a convolutional neural network can be easily fooled by its attackers [1,2].

Interpretable machine learning is an emerging research topic that tries to solve the aforementioned problem and opens up the black-box of complicated machine learning algorithms [3]. Two categories of interpretability are generally investigated, i.e., post-hoc interpretability and intrinsic interpretability. In the post-hoc analysis, a fitted model is interpreted using external tools. Examples of this category include the partial dependence plot (PDP; [4]), local interpretable model-agnostic explanations (LIME; [5]), SHapley Additive exPlanations (SHAP; [6]), and visual explanation of deep neural networks [7,8]. In contrast, intrinsic interpretability aims at making the model intrinsically interpretable. Many statistical models belong to this category, e.g., generalized linear model, decision tree, and naïve Bayes classifier. In this paper, we limit our focus to the second type of interpretability.

The generalized additive index model (GAIM) is such an intrinsically interpretable model when proper constraints are imposed. It was first proposed by [9] in the name of projection pursuit regression. GAIM is shown to have close connections with feedforward neural networks [10], which has universal approximation capability as the number of hidden nodes is sufficiently large [11]. The functional relationship between raw features $\boldsymbol{x} \in \mathbb{R}^p$ and the response y is represented by

$$g(\mathbb{E}(y|\boldsymbol{x})) = \mu + \sum_{j=1}^{M} h_j(\boldsymbol{w}_j^T \boldsymbol{x}), \tag{1}$$

where g is a pre-specified link function, μ is the intercept, and M is the number of additive functional components. For each $j = 1, \ldots, M$, $\boldsymbol{w}_j \in \mathbb{R}^p$ denotes the projection index and h_j is the so-called ridge or nonlinear shape function. Conventionally, GAIM is estimated by the backfitting algorithm, which iteratively estimates a pair of $\{\boldsymbol{w}_j, h_j\}$ at a time, with other pairs fixed. Nonparametric regression (e.g., smoothing splines) is used to fit the shape func-

* Corresponding author.
E-mail address: ajzhang@umich.edu (A. Zhang).

https://doi.org/10.1016/j.patcog.2021.108192
0031-3203/

tions in (1). Such a greedy procedure yields the sub-optimal solution. Recently, GAIM has been reformulated to be an explainable neural network (xNN[12];). In xNN, a fully-connected multi-layer perceptron is disentangled into a projection layer followed by multiple sub-modular networks, where each subnetwork represents a nonlinear shape function in (1). The interpretability of xNN is further enhanced by imposing sparsity, orthogonality and smoothness constraints [13]. As a result of using neural network parametrization of shape functions in xNN, it is more likely to obtain a globally optimal solution through full network training.

The generalized additive model (GAM; [14]) is another intrinsically interpretable model of the form

$$g(\mathbb{E}(y|\boldsymbol{x})) = \mu + \sum_{j=1}^{p} h_j(x_j), \tag{2}$$

which is a special case of (1). Regarding the expressive power, GAIM is indeed much more competitive than GAM; however, for some specific applications, the fitting results of GAIM are not easily interpretable. The main problem lies in interpreting the projection $\boldsymbol{z}_j = \boldsymbol{w}_j^T \boldsymbol{x}$. For features with different practical meanings, their linear combination could be non-intuitive and not interpretable. For example, a weighted sum of stock prices can be accepted by human beings; while the weighted sum of feature values of different types (e.g., stock price and temperature) is not directly interpretable. To avoid GAIM with such hard-to-interpret projections, we consider GAM (2) whose each component is a function of the original interpretable feature.

An empirical study of GAM based on machine learning datasets was presented by [15], which suggested that using tree ensembles to fit nonlinear shape functions in (2) may achieve better predictive performance than using regression splines. Recently, it draws our attention that [16] proposed to use neural network representation for the shape functions in GAM, which is the same idea as in xNN [12,13].

The interaction effects between individual features can be incorporated into the GAM for performance improvement [17,18]. Among them, the generalized additive models with pairwise interactions (GA^{2}M) proposed by [17] is a state-of-the-art extension of (2) plus pairwise interactions, which is also known as the explainable boosting machine (EBM) with a fast implementation by Microsoft Research [19]. EBM is similar to [15] by using tree ensembles to fit either main effect $h_j(x_j)$ or pairwise interaction $f_{jk}(x_j, x_k)$, and it comes with a fast procedure for pairwise interaction detection. It is shown by [19] that EBM has an overwhelming prediction performance when compared to some black-box models based on five classification datasets.

In this paper, a novel xNN structure is proposed by using neural network parametrization for both main effects and pairwise interactions, and we call it GAMI-Net. Unlike EBM based on tree ensembles, we suggest modeling each main effect or pairwise interaction by a fully-connected subnetwork consisting of one or two input nodes, respectively. These subnetworks are then additively combined to form the final output. Each subnetwork can be easily visualized by 1D and 2D plots for the purpose of interpretation. In addition to neural network parametrization, the interpretability of GAMI-Net is enhanced with the following three constraints,

- **Sparsity**. Model parsimony is an essential factor for an interpretable model. In GAMI-Net, only non-trivial main effects and pairwise interactions are included. Pruning of trivial effects is also helpful for reducing the degree of overfitting.
- **Heredity**. The classic heredity principle in statistics is introduced to enhance structural interpretability. That is, a pairwise interaction can only be included in the final model if at least one of its parent main effects is important.
- **Marginal clarity**. The first two constraints are both employed to select important main effects and pairwise interactions, while marginal clarity serves as a regularization to avoid potential confusion between main effects and their corresponding child pairwise interactions.

A three-stage adaptive training algorithm is proposed for GAMI-Net estimation. First, the main effect subnetworks are trained and pruned. Second, important pairwise interactions are selected, fitted, and pruned. Finally, all the important main effects and pairwise interactions are collectively fine-tuned. Numerical experiments on both synthetic functions and real-world datasets are conducted. The superiority of GAMI-Net is reflected in both predictive performance and intrinsic interpretability. In the synthetic functions, GAMI-Net achieves the best predictive performance and the visualized model fits are close to the ground truth. In the real-world datasets, GAMI-Net also shows close predictive performance to black-box models, and its interpretability is demonstrated by two case studies. Therefore, the proposed GAMI-Net can serve as a promising tool for interpretable machine learning.

This paper is organized as follows. Section 2 presents the proposed GAMI-Net methodology, including the network architecture, the training algorithm, and the interpretability. One synthetic function and multiple real-world datasets are used to test the GAMI-Net performance in Section 3. Finally, Section 4 concludes this paper.

2. GAMI-Net methodology

This section first introduces the proposed GAMI-Net architecture, interpretability constraints, and computational algorithm. Further discussions are then provided, regarding the model interpretation and hyperparameter tuning guidelines. Finally, we compare GAMI-Net with its counterpart models from various perspectives.

2.1. Network architecture

In GAMI-Net, a complex functional relationship is formulated via its lower-order representations, including nonlinear main effects and pairwise interactions[1]. Let S_1, S_2 denote the sets of active main effects and pairwise interactions, respectively. Then, the proposed GAMI-Net is formulated as follows,

$$g(\mathbb{E}(y|\boldsymbol{x})) = \mu + \sum_{j\in S_1} h_j(x_j) + \sum_{(j,k)\in S_2} f_{jk}(x_j, x_k). \tag{3}$$

Note that (j, k) and (k, j) refer to the same pairwise interaction. Here, each main effect and pairwise interaction is assumed to have zero mean, i.e.,

$$\begin{aligned} &\int h_j(x_j) dF(x_j) = 0, \forall j \in S_1, \\ &\int f_{jk}(x_j, x_k) dF(x_j, x_k) = 0, \forall (j,k) \in S_2, \end{aligned} \tag{4}$$

where $F(x_j)$ and $F(x_j, x_k)$ represent the corresponding cumulative distribution functions. Besides, each pairwise interaction $f_{jk}(x_j, x_k)$ is desired to be nearly orthogonal to its parent main effects $h_j(x_j)$ and $h_k(x_k)$, subject to the marginal clarity constraint to be discussed later. Note that the zero mean and marginal clarity constraints are both introduced for identifiability consideration. In particular, the zero mean constraint can be easily implemented by normalizing the subnetwork outputs at the end of each training

[1] Higher-order interactions can be treated similarly by GAMI-Net, but for simplicity we focus on pairwise interactions only. Meanwhile, we believe that higher-order interactions are usually rare in practice (unlike pairwise interactions), and when they exist the interpretation is not straightforward.

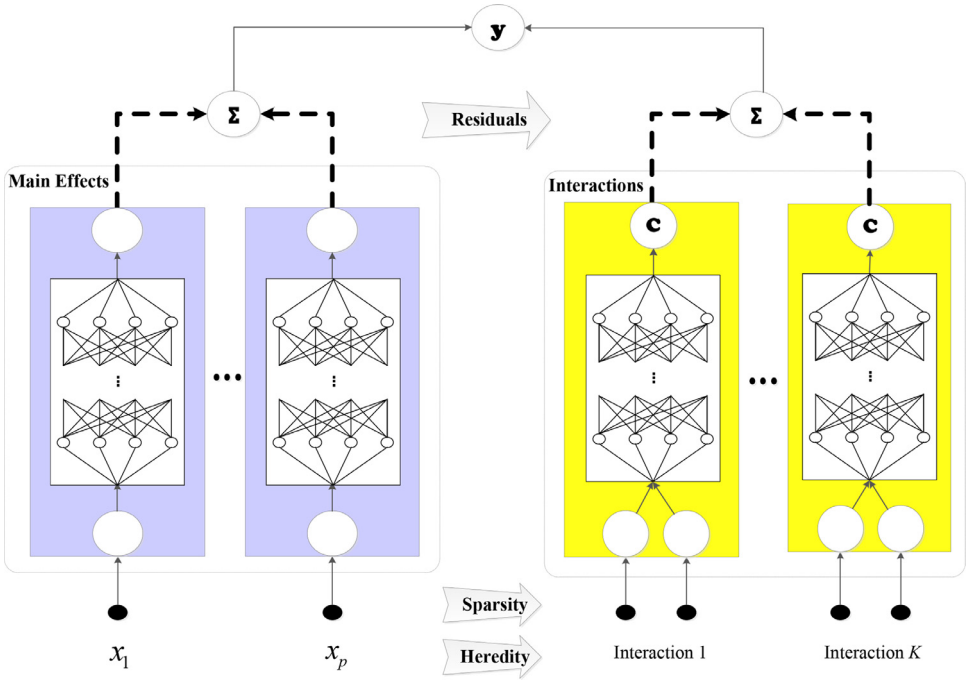


Fig. 1. The GAMI-Net architecture. The main effects are fitted first, then the top-K ranked pairwise interactions are selected and fitted to the residuals, subject to the heredity constraint. The dashed arrows pointing to the $\boldsymbol{\Sigma}$ nodes denote the sparsity constraints, where the trivial subnetworks are pruned. Finally, the marginal clarity is imposed for regularizing pairwise interactions, denoted by the symbol "C".

iteration; while the marginal clarity constraint is achieved by regularizing the outputs of pairwise interaction subnetworks during the training iterations.

The GAMI-Net architecture is presented in Fig. 1. It consists of a main effect module and a pairwise interaction module. Each main effect $h_j(x_j)$ in (3) is captured by a subnetwork consisting of one input node, multiple hidden layers, and one output node. Each pairwise interaction $f_{jk}(x_j, x_k)$ in (3) is captured by a subnetwork with two input nodes. All these networks are linearly combined (plus a bias node for capturing the intercept μ) to produce the final output. More specifically, each main effect subnetwork fits a 1D curve, while each interaction subnetwork approximates a 2D surface. When approximating an arbitrary curve or surface, we can use a single-hidden-layer feedforward neural network with a sufficiently large number of hidden nodes, while modern deep learning training techniques make it feasible to use multiple hidden layers to achieve superior predictive performance. In fact, the multi-layer subnetworks are flexible enough to capture any form of functions upon proper network configuration. Besides, the categorical variables are preprocessed using one-hot encoding. The subnetworks used for fitting the main effects of categorical variables can be simplified to multiple bias nodes, where each node captures the intercept effect of a corresponding dummy variable.

2.2. Interpretability constraints

The proposed GAMI-Net is developed with sparsity, heredity, and marginal clarity constraints. Specifically, sparsity and heredity constraints are introduced to enhance the interpretability of the fitted model; while the marginal clarity constraint is introduced to make main effects and their child pairwise interactions uniquely identifiable.

Sparsity constraint. The principle of parsimony is commonly assumed in statistical machine learning. The sparse models, upon reduction of unnecessary model complexity, not only enjoy computational benefits but also prevent overfitting problems. Moreover, sparsity is an essential building block for model interpretation. For example, a shallow decision tree that uses a few explanatory variables is generally thought to be easily interpretable; however, a deep decision tree involving multiple variables and many leaf nodes can be hardly understandable. For high-dimensional data, the GAM involving all the variables can be too complex to interpret. Therefore, it is critical for GAMI-Net to remove unnecessary main or interaction effects, in order to benefit from efficient computation and enhanced interpretability.

The importance of a main effect or pairwise interaction can be quantified by the variation it explains. Empirically, the variation of the j-th main effect can be measured by the sample variance,

$$D(h_j) = \frac{1}{n-1}\sum h_j^2(x_j), \tag{5}$$

where n is the sample size. We treat the main effect functions with very small variation as trivial effects, and enforce them to zero, which results in the sparse GAM [20]. Alternatively, given an integer parameter s_1 (between 1 and p), GAMI-Net is designed to select the top-s_1 main effects ranked by $D(h_j)$ values, as listed by the index set S_1.

Similarly, the sparsity of pairwise interactions can also be induced by selecting the top-s_2 pairwise interactions according to $D(f_{jk})$ defined by

$$D(f_{jk}) = \frac{1}{n-1}\sum f_{jk}^2(x_j, x_k), \tag{6}$$

for all the pairwise interactions. We use the index set S_2 to denote the list of selected top-s_2 pairwise interactions.

Heredity constraint. In addition to sparsity constraint, hierarchical and hereditary principles are essential rules for modeling main effects and low-order to high-order interactions. The hierarchical principle states that lower-order effects are generally more important than higher-order effects. The principle of heredity further requires a more strict hierarchical structure between main effects and interactions [21], whereas the model violating the heredity principle is thought to be insensible [22]. The heredity principle has also been used in the variable selection literature [23–26].

There are two versions of the heredity principle, namely strong heredity, and weak heredity. In the case of main effects (indexed by S_1) and pairwise interactions (indexed by S_2), the strong heredity imposes the constrain that

$$\forall (j,k) \in S_2 : j \in S_1 \quad \text{and} \quad k \in S_1,$$

while the weak heredity imposes that

$$\forall (j,k) \in S_2 : j \in S_1 \quad \text{or} \quad k \in S_1.$$

That is, a pairwise interaction can be included by S_2 only if a) both of its parent main effects are included (strong heredity) by S_1, or b) at least one of its parent main effects are included by S_1.

In GAMI-Net, the weak heredity constraint is employed for the following reasons. First, the search space (of pairwise interactions) can be reduced and hence it brings computational efficiency. Second, the resulting model can be improved with enhanced interpretability in the sense of the heredity principle. Third, the heredity principle is empirically supported in statistical modeling literature; see, for instance, the meta-analysis conducted by [27] for a large number of data sets from published factorial experiments.

Marginal clarity. For model identifiability, each main effect or pairwise interaction is assumed to have zero mean in (4). However, without further assumptions, the main effects can be easily absorbed by their child interactions and vice versa. There could be multiple representations for a given model, which makes the model estimation unstable and leads to confusion in model interpretation.

The marginal clarity constraint is accordingly introduced to make the model more identifiable. It is motivated by the functional ANOVA decomposition, in which the original function can be uniquely decomposed into orthogonal components. The weighted functional ANOVA decomposition [28] is proposed for handling explanatory variables with empirical distributions, where the orthogonality condition for the j-th main effect and corresponding pairwise interaction (j,k) is presented as follows,

$$\int h_j(x_j) f_{jk}(x_j, x_k) dF(\boldsymbol{x}) = 0. \tag{7}$$

The symbol $F(\boldsymbol{x})$ denotes the joint cumulative distribution function. Empirically, the degree of non-orthogonality can be defined by

$$\Omega(h_j, f_{jk}) = \left| \frac{1}{n} \sum h_j(x_j) f_{jk}(x_j, x_k) \right|. \tag{8}$$

The smaller the value of $\Omega(h_j, f_{jk})$, the more clearly the marginal effect h_j is separated from its child interaction f_{jk}. The perfect case is when $\Omega(h_j, f_{jk}) = 0$; in practice, it is acceptable to have $\Omega(h_j, f_{jk})$ slightly greater than zero. Hence in GAMI-Net, we penalize the non-orthogonality $\Omega(h_j, f_{jk})$ for all $j \in S_1$ and their corresponding child interactions $(j,k) \in S_2$, in a way for pursuing the marginal clarity. Note that a similar interaction purifying method is proposed in [29], which is based on post-hoc processing and only suitable for piecewise constant functions.

2.3. Computational aspects

In this section, we discuss the computational procedures for estimating GAMI-Net. All the unknown parameters in the proposed model are denoted by $\boldsymbol{\theta}$. For each sample $\boldsymbol{x}$, the prediction is denoted by $\hat{y} = \mathbb{E}(y|\boldsymbol{x}; \boldsymbol{\theta})$. Combining all the interpretability constraints, GAMI-Net is estimated by solving the following constrained optimization problem,

$$\begin{aligned} &\min_{\boldsymbol{\theta}} \mathcal{L}_\lambda(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) + \lambda \textstyle\sum_{j \in S_1} \sum_{(j,k) \in S_2} \Omega(h_j, f_{jk}), \\ &\text{s.t.} \int h_j(x_j) dF(x_j) = 0, \forall j \in S_1, \\ &\int f_{jk}(x_j, x_k) dF(x_j, x_k) = 0, \forall (j,k) \in S_2, \end{aligned} \tag{9}$$

where the active sets of main effects and pairwise interactions S_1, S_2 are determined subject to sparsity and heredity constraints. The empirical loss $l(\boldsymbol{\theta})$ is determined by the type of tasks (e.g., regression or classification). The second term is the marginal clarity regularization, and the regularization strength is denoted by $\lambda \geq 0$.

Referring to Fig. 1, an adaptive training algorithm is introduced to sequentially estimate the main effects and pairwise interactions, which can be summarized as the following three stages.

1) Train all the main effect subnetworks for some epochs and prune the trivial main effects according to their contributions and validation performance.
2) Select at most K pairwise interactions for training and then prune the trivial pairwise interactions according to their contributions and validation performance.
3) Fine-tune all the network parameters for some epochs.

Training main effects. In the first stage, all the main effect subnetworks are simultaneously estimated while the pairwise interaction subnetworks are frozen to zero. The trainable parameters in the network are updated by mini-batch gradient descent with adaptive learning rates determined by the Adam optimizer, which is scalable to very large datasets.

The training would stop as the maximum number of training epochs is reached or the validation performance does not get improved for a certain number of epochs. Each main effect is then normalized to have zero mean such that the bias node of the output layer represents the overall mean. The trivial main effect subnetworks are then pruned according to the sparsity constraint. Given a null model that only contains the intercept term, we evaluate its performance on the validation set, denoted by l_0. The most important main effect is then added, and we evaluate its validation performance l_1. Next, the other important main effects are added one-by-one in the descending order of their contributions (5). The list $\{l_0, l_1, \cdots, l_p\}$ represents the corresponding validation loss.

In general, when more and more main effects are added, the validation loss would show a decreasing trend. However, including too many main effects may lead to overfitting, as reflected by the turning trend in the validation loss curve. According to the sparse modeling principle, those main effects after the turning point should be pruned. In practice, a tolerance threshold η is introduced to balance the level of sparsity and predictive performance. Then, s_1 is set to be the minimal index whose validation loss is smaller than or equal to $(1+\eta)\min\{l_0, l_1, \cdots, l_p\}$. The active set S_1 is determined as the list of top-s_1 important main effects. Note that the aforementioned threshold can be useless when the minimal loss is zero. In that case, we may need to manually determine the validation loss threshold.

Training interactions. After the top-s_1 important main effects are captured, the next step is to train the pairwise interaction subnetworks. In total, there exist $p(p-1)/2$ possible pairwise interactions that can be tested, which is extremely time-consuming, especially for a large p. According to the weak heredity constraint, we consider pairwise interactions with at least one of their parent main effects belonging to S_1. This reduces the computational complexity a lot when s_1 is much less than p. Besides, an interaction filtering procedure is introduced to remove the pairwise interactions, which are less likely to be important. There exist many interaction detection methods in the literature, to list a few, hierarchical Lasso [23], shallow tree-like model-based pairwise interaction ranking [17], and neural network-based interaction detection [30].

In GAMI-Net, we employ the interaction ranking algorithm proposed in [17], subject to the heredity constraint. The modified pairwise interaction filtering algorithm selects the top-K pairwise interactions through the following steps.

1) Obtain the prediction residuals from the main effect training stage.

2) For each $j < k$ with $j \in S_1$ or $k \in S_1$, evaluate the strength of interaction (j, k) by building shallow tree-like models between variables (x_j, x_k) and the residuals; the strength of interaction (j, k) is set to the minimal fitting loss across all evaluated trees [17].
3) Rank all the evaluated pairwise interactions and obtain the top-K pairwise interactions.

Next, the selected top-K pairwise interactions are simultaneously trained using the mini-batch gradient algorithm, subject to the marginal clarity regularization. Note that in this stage, the main effect subnetworks are fixed. Each estimated pairwise interaction is normalized to have zero mean, for which the offset is added to the bias node in the output layer.

The pruning of pairwise interaction effects is similar to that of the main effects. We start from the pre-trained model with the intercept term and active main effects. The top-ranked pairwise interactions are sequentially added to the model, together with the record of their corresponding validation losses. For simplicity, we use the same tolerance threshold η as for main effect pruning, in order to balance the level sparsity and predictive performance. Thus, s_2 can be determined accordingly, and the active set S_2 is formed as the list of the top-s_2 important pairwise interactions.

Fine tuning. The first two stages perform a structured variable selection. In the final stage, a fine-tuning procedure is implemented to jointly retrain all the active subnetworks, where the marginal clarity regularization is still imposed between main effects and pairwise interactions. All the main effects and pairwise interactions are re-normalized. We find such a fine-tuning step is helpful to solve the following two problems:

1) The removal of trivial main effects or pairwise interactions may lead to biased estimation;
2) The pairwise interactions estimated separately are conditional on the pre-trained main effects (subject to marginal clarity regularization), which can limit the predictive performance.

These two problems can be mitigated via jointly retraining all the selected main effects and pairwise interactions so that the predictive performance of GAMI-Net can be further improved. When applying the estimated GAMI-Net for prediction, data outside the training range are clipped to make the prediction stable. Assume the training range of x_1 is $[0, 1]$. As new data comes, data with x_1 smaller than 0 or greater than 1 would be clipped to 0 or 1, respectively, before inputted into the model.

2.4. Interpretability of GAMI-Net

The proposed GAMI-Net is intrinsically interpretable in the following aspects.

Importance Ratio (IR). Given an estimated GAMI-Net, we can inspect the contribution of each individual variable to the overall prediction. The IR of each main effect can be quantitatively measured by

$$\mathrm{IR}(j) = D(h_j)/T, \tag{10}$$

where $T = \sum_{j\in S_1} D(h_j) + \sum_{(j,k)\in S_2} D(f_{jk})$. Similarly, the IR of each pairwise interaction can be measured by

$$\mathrm{IR}(j, k) = D(f_{jk})/T. \tag{11}$$

The IR's of all the effects sum up to one. In practice, we can sort the effect importance according to the IR values in descending order. The effects of large IR values are more important.

The definition of IR is related to that of Sobol indices [31]. The main difference lies in that Sobol indice is derived under the assumption that all the variables are independent and uniformly distributed, while IR is based on the empirical distributions of explanatory variables.

Global interpretation. In addition to measuring the importance of each estimated effect, we can further inspect the relationship between one / two individual variables and the response by visualizing the fitted shape functions. Unlike the post-hoc diagnostic tool PDP [4], the partial dependence relationships can be directly obtained from GAMI-Net. We suggest using the 1D line plots for numerical variables and the bar charts for categorical variables to show the input-output relationship, which can be linear, convex, monotonic, and any other forms. These plots can be directly drawn based on the final estimates of $h_j(x_j)$ for $j \in S_1$. Moreover, we suggest using the 2D heatmap for visualizing each estimated pairwise interaction, which shows the joint effect of the two underlying variables. See, e.g., Fig. 4 (middle panel) for such kinds of plots.

Local interpretation. The prediction by GAMI-Net is also easy to be locally explained, leading to a transparent decision-making system. Given a sample $\boldsymbol{x}$, the model not only outputs the final decision but also the concrete function form (1) with the input $\boldsymbol{x}$. The values of each additive component, i.e., marginal main or pairwise interaction effects, can be directly obtained. These marginal effects can be rank-ordered for understanding the decision for the input $\boldsymbol{x}$ specifically. Besides, the sensitivity of prediction to small changes of an explanatory variable can be quantitatively investigated by the corresponding 1D line plots (or bar charts) and 2D heatmaps.

2.5. Hyperparameters

Some hyperparameters for GAMI-Net can be configured with the following default settings (for numerical experiments in the next section). The maximal number of pairwise interactions is set to $K = 20$. For simplicity, each subnetwork is configured to have 5 ReLU hidden layers, with 40 nodes per layer. It is worth mentioning that the choice of activation would affect the resulting functional forms of the fitted model. Using ReLU, the fitted curves are piecewise linear; while using hyperbolic tangent, the fitted curves can be more smooth.

The subnetwork weights are initialized using the Gaussian orthogonal initializer. The initial learning rate of the Adam optimizer is set to 0.0001. The numbers of training epochs for the three training stages are set to 5000, 5000, and 500, respectively. The mini-batch sample size is determined according to the sample sizes of different datasets. A 20% validation set is split for early stopping, and the early stopping threshold is set to be 50 epochs. The tolerance threshold η is set to be 1% of the minimal validation loss. The marginal clarity regularization strength λ can be empirically selected from 0.0001 to 1.

Finally, a demo implementation of the proposed GAM-Net is publicly available, which can be found on Github[2], including the numerical examples presented in this paper. This package is based on the *TensorFlow 2.0* platform using the Python language.

2.6. Comparison with related methods

Unlike traditional spline-based GAMs, GAMI-Net uses neural networks to model the non-parametric shape functions. The proposed GAMI-Net is also closely related to the explainable boosting machine (EBM; [17]), as both of them are based on main effects and pairwise interactions.

Base models. In EBM, each main effect or pairwise interaction is estimated via gradient boosted shallow trees, which is modified from the standard gradient boosting model [4]. Therefore, the estimated shape functions by EBM are all piecewise constant. Empirically, gradient boosted shallow trees are shown to have strong ap-

[2] https://github.com/SelfExplainML/GamiNet

proximation ability, which makes EBM even comparable to black-box models [17]. Despite its predictive performance, EBM sometimes outputs shape functions with unexpected jumps, which are hard to explain. Such a problem may become worse when there exist outliers or noisy samples.

In contrast, spline-based GAMs and neural network-based GAMI-Net usually output continuous shape functions for numerical variables. Using splines, the smoothness of the fitted functions can be partially controlled by the choice of spline orders and the roughness penalty, while for GAMI-Net, the fitted functions can be piecewise linear (e.g., when using ReLU) or more smooth (e.g., when using sigmoid). Such continuous or smooth shape functions can prevent unexpected jumps and therefore warrant the model interpretability.

Model estimation. EBM is estimated using a boosting algorithm, in which all variables are sequentially fitted to the residuals for multiple iterations until the stopping rule is reached. Spline-based GAM models are usually estimated by backfitting, where each variable is fitted at a time in a cycle sequential way. In GAMI-Net, the training algorithm is composed of three stages. The first two stages not only select important main effects and pairwise interactions but also provide a good initialization. The third stage fine-tunes all the network parameters. Unlike boosting or backfitting, the main effects and pairwise interactions in GAMI-Net are jointly optimized, and it is more likely to find the global optimum. In addition, the GAMI-Net fitting algorithm shares the advantages of modern deep learning training techniques and is easily scalable to extremely large datasets.

Interpretability constraints. The proposed GAMI-Net tends to be more efficient and more interpretable than EBM. In EBM, all the main effects are included in the final model; the number of active pairwise interactions in EBM can only be pre-specified (which is not flexible) or tuned by cross-validation (which is time-consuming). The resulting model from EBM can be extremely complex for high-dimensional data. Besides, without marginal clarity constraint in EBM, an estimated main effect and its corresponding child pairwise interactions may be mutually absorbed, which leads to non-identifiable results. Such problems can be well addressed by GAMI-Net's interpretability constraints, including sparsity, heredity, and marginal clarity.

3. Numerical experiments

In this section, the proposed GAMI-Net is tested on a synthetic example and an extensive list of real-world datasets.

3.1. Experimental setup

Several benchmark models are included for comparison, including EBM, spline-based GAM, generalized linear models (GLM), multi-layer perceptron (MLP), random forest (RF), and extreme gradient boosting (XGBoost). Specifically, EBM is implemented by the open-source Python package *interpret* [19]. The spline-based GAM is based on the implementation of the *pyGAM* package [32], and we use pyGAM to denote spline-based GAM in the remaining part of this paper. For the other benchmarks, GLM, MLP, and RF are all available in the *Scikit-learn* package, and XGBoost is implemented by the *xgboost* package. In particular, GLM uses Lasso for regression tasks and ℓ_1-shrinkage logistic regression for binary classification tasks. The comparative results are grouped into two categories, intrinsically interpretable models (GAMI-Net, EBM, pyGAM, and GLM) and black-box models (MLP, RF, and XGBoost).

By default, we split a dataset into training (80%) and test (20%) sets upon random permutation. For hyperparameter tuning, a 20% hold-out validation set is further split from the training set. GAMI-Net is configured and trained using the settings described in Section 2.5. By default, the strength of the marginal clarity regularization is set to 1 in the simulation study and 0.1 in all the real-world datasets. The rationale of such a setting is further justified through ablation studies. In EBM, the number of interactions is set to 20, and all the other hyperparameters are set to the default values. In pyGAM, the smoothness regularization strength is tuned within the package's recommended range. The ℓ_1 regularization strength for Lasso (or logistic regression) is tuned within $\{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}\}$. For black-box MLP, the hidden layer architecture is set to $[40, 20]$ with hyperbolic tangent nodes. Finally, the number of base estimators is set to 500 for both RF and XGBoost; and for each of them, the maximum tree depth is tuned within $\{3, 4, 5, 6, 7, 8\}$. The predictive performance is measured by the root-mean-square error (RMSE) for regression tasks and the area under the ROC curve (AUC) for binary classification tasks. All the experiments are repeated 10 times, and we report the average results. Each time the data is reshuffled before getting split. For a fair comparison, all the models are given the same training and test sets for the same repetition.

3.2. Simulation study

A synthetic function is used to demonstrate the proposed method, in which both main effects and pairwise interactions are included, as follows,

$$y = 8\left(x_1 - \frac{1}{2}\right)^2 + \frac{1}{10}e^{(-8x_2+4)} + 3\sin(2\pi x_3 x_4) + 5e^{-2(2x_5-1)^2-\frac{1}{2}\left[15x_6+12(2x_5-1)^2-13\right]^2} + \varepsilon, \tag{12}$$

where the response is calculated via complicated nonlinear transformations of the explanatory variables plus a noise term generated from the standard normal distribution. In addition to $(x_1, \cdots, x_6)$, a large number of noisy variables $(x_7, \cdots, x_{100})$ are also introduced, which have no contribution to the response. These explanatory variables are independently generated within the domain $[0, 1]$, with 3 different distributions, i.e, uniform distribution $U(0, 1)$, normal distribution $N(0.5, 0.2^2)$ truncated within $[0, 1]$, and exponential distribution $Exp(0.5)$ truncated within $[0, 1]$. For each of these three distributions, four different sample sizes are tested, i.e., $n = \{1000, 2000, 5000, 10000\}$.

Table 1 reports the averaged test set RMSE and standard deviation (over 10 repetitions) of different models on this synthetic dataset. For each setting, the best interpretable and black-box models are both highlighted in bold, respectively. It can be observed that the proposed GAMI-Net outperforms all the compared models, including both interpretable and black-box models. In all the tested cases, GAMI-Net outperforms the black-box models, including MLP and RF.

The training and validation losses of GAMI-Net are presented in Fig. 2, for the case with uniform distribution and $n = 10000$. It can be observed that the losses decrease significantly as pairwise interactions are added to the network, which shows the necessity of adding pairwise interactions to GAM. At the beginning of the fine-tuning stage, there exists a sudden jump of training loss (increase) and validation loss (decrease), which corresponds to the pruning of trivial pairwise interactions. Besides, the validation loss for determining the optimal number of main effects and pairwise interactions are visualized in Fig. 3. The left and right x-axises denote the number of included main effects and pairwise interactions, respectively. Red star symbols mark the optimal number of main effects / pairwise interactions. The results show that $s_1 = 6$ main effects and $s_2 = 2$ pairwise interactions are included in GAMI-Net. The marginal benefits of adding more effects could be extremely small and may even lead to the overfitting problem.

Figs. 4 (a) and 4(b) draw the ground truth and global interpretation of GAMI-Net (with uniform distribution and $n = 10000$). Note

Table 1
Testing RMSE comparison of the synthetic function.

Distribution	n	GAMI-Net	EBM	pyGAM	GLM	MLP	RF	XGBoost
uniform	1000	**1.805** ±0.434	2.587±0.000	2.998±0.138	2.999±0.000	3.201±0.000	2.458±0.160	**2.383** ±0.172
uniform	2000	**1.215** ±0.172	2.529±0.000	2.889±0.069	3.124±0.000	3.054±0.000	2.127±0.056	**2.077** ±0.099
uniform	5000	**1.071** ±0.029	2.158±0.000	2.456±0.058	3.003±0.000	2.705±0.000	1.895±0.058	**1.790** ±0.045
uniform	10000	**1.044** ±0.020	1.799±0.132	2.450±0.046	3.103±0.065	2.615±0.057	1.822±0.043	**1.634** ±0.013
normal	1000	**1.858** ±0.360	2.029±0.000	2.528±0.120	2.399±0.000	2.484±0.000	1.914±0.104	**1.882** ±0.121
normal	2000	**1.340** ±0.221	1.914±0.000	2.526±0.081	2.579±0.000	2.474±0.000	1.702±0.052	**1.660** ±0.066
normal	5000	**1.043** ±0.020	1.854±0.000	2.105±0.050	2.505±0.000	2.148±0.000	1.541±0.059	**1.487** ±0.042
normal	10000	**1.050** ±0.021	1.647±0.124	2.010±0.032	2.560±0.055	2.110±0.024	1.485±0.027	**1.358** ±0.026
exponential	1000	**1.417** ±0.099	2.066±0.000	2.357±0.137	2.336±0.000	2.519±0.000	2.010±0.107	**1.960** ±0.098
exponential	2000	**1.273** ±0.083	1.965±0.000	2.221±0.083	2.360±0.000	2.325±0.000	1.890±0.101	**1.815** ±0.069
exponential	5000	**1.031** ±0.016	1.833±0.000	1.954±0.045	2.508±0.000	2.303±0.000	1.767±0.056	**1.624** ±0.040
exponential	10000	**1.025** ±0.013	1.614±0.104	1.893±0.023	2.523±0.043	2.078±0.030	1.659±0.037	**1.520** ±0.034

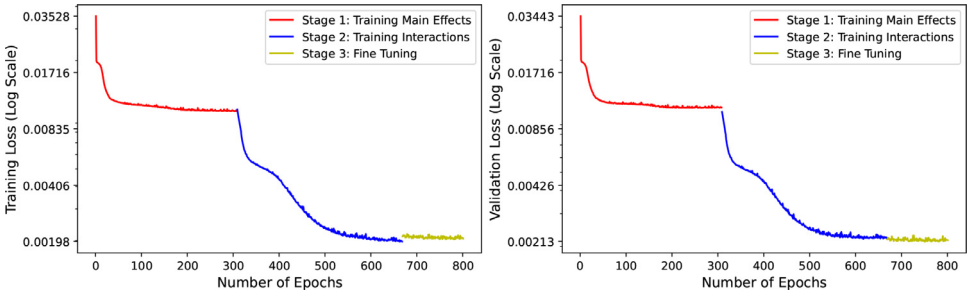


Fig. 2. The training and validation trajectories of GAMI-Net for the synthetic function (uniform distribution; $n = 10000$).

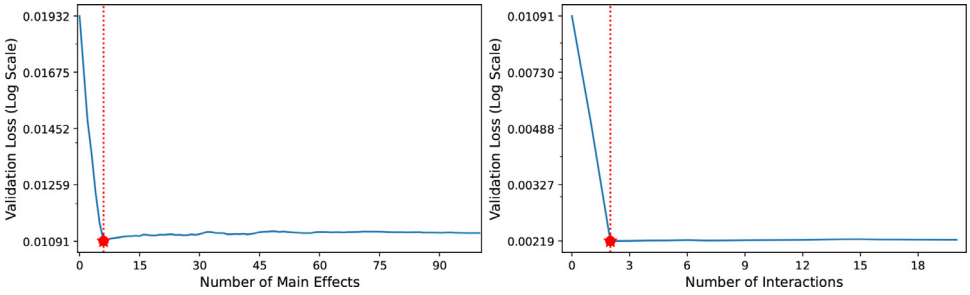


Fig. 3. The validation loss for determining s_1, s_2 for the synthetic function (uniform distribution; $n = 10000$).

that the original formulation (12) has only 2 active main effects (x_1, x_2) and 2 active interaction effects $\{(x_3, x_4), (x_5, x_6)\}$. But according to the functional ANOVA decomposition, this formula can be rewritten such that the marginal main effects are extracted from the interactions. Therefore, the active main effects also include x_3, x_4, x_5, x_6. Each main effect / pairwise interaction is ranked in the descending order of IR, and the pairwise interactions are all presented behind the main effects. It can be observed that all the 6 main effects and 2 pairwise interactions are successfully captured by GAMI-Net, which is close to that of the ground truth.

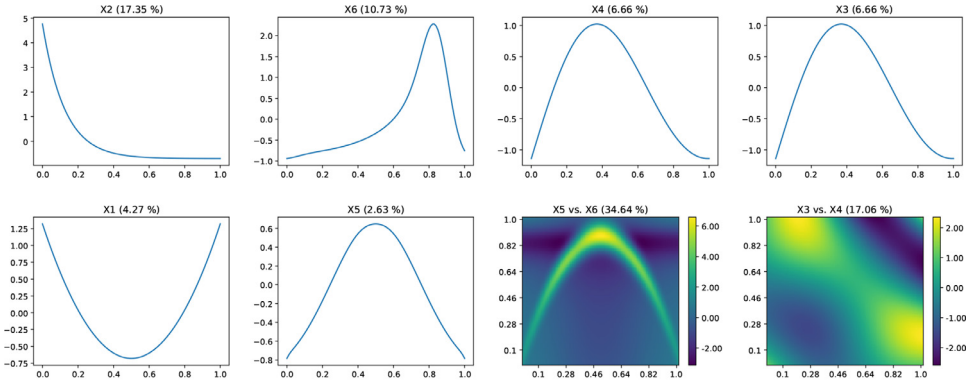
Since EBM does not have a pruning procedure, the final model of EBM includes 100 main effects and 20 pairwise interactions. To make a valid comparison, we also draw its first 6 main effects and first 2 pairwise interactions in Fig. 4c. The results indicate that EBM can also approximately capture the shape of these important effects. However, due to the use of gradient boosting trees, the estimated shape functions are all piecewise constant, and the existence of sudden jumps makes it hard to interpret. Second, we also calculate IR for each effect in EBM using the same method as in GAMI-Net. The result of EBM is shown to have a larger bias as compared to the actual model. For example, the interaction (x_5, x_6) is underestimated, and the overall IR captured by these true effects is just around 80%. That means the noise effects take more than 10% of the contribution.

The benefits of introducing the sparsity constraint in GAMI-Net are already demonstrated in Fig. 3. Moreover, ablation studies are conducted to justify the use of heredity and marginal clarity constraints. In Table 2, we report the training, validation, and test RMSE of the benchmark models. There exist several reasons that EBM fails in this task. First, as the ground truth function is continuous, the piecewise constant fits cannot well capture the ground truth. Second, due to the lack of sparsity consideration, EBM suffers from the overfitting problem. Third, as the main effects are not well captured, the correct pairwise interactions may not always be detected.

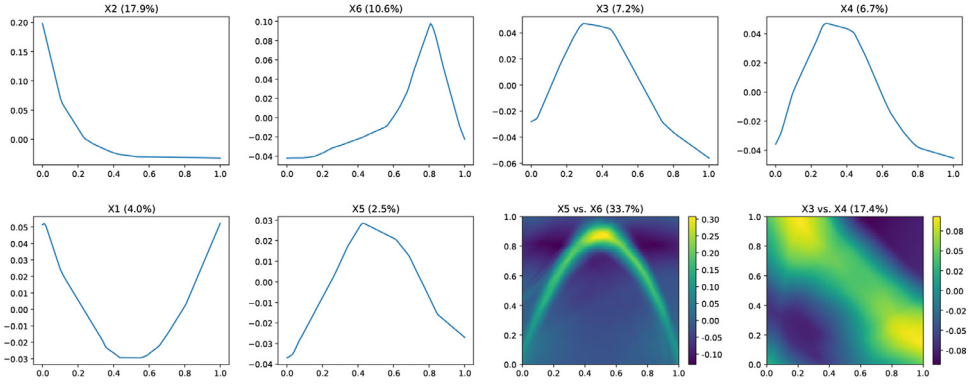
Then, the results of GAMI-Net with different marginal clarity regularization strengths are also reported in Table 2. The last row denotes the results of GAMI-Net without heredity constraint and $\lambda = 10^0$. The marginal clarity losses are calculated for both EBM and GAMI-Net, via

$$\sum_{j \in S_1} \sum_{(j,k) \in S_2} \Omega(h_j, f_{jk}), \tag{13}$$

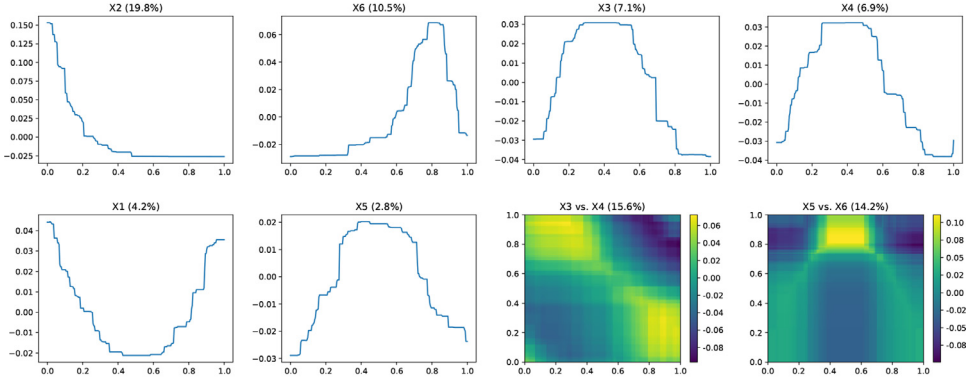
for which the smaller, the better. The results show that the increase of λ can help a) prevent overfitting, according to the validation RMSE; b) make the model more identifiable, see the decreasing trend of marginal clarity loss. As is discussed in the previous sections, the purpose of using heredity constraint is to help reduce



(a) Ground Truth



(b) GAMI-Net

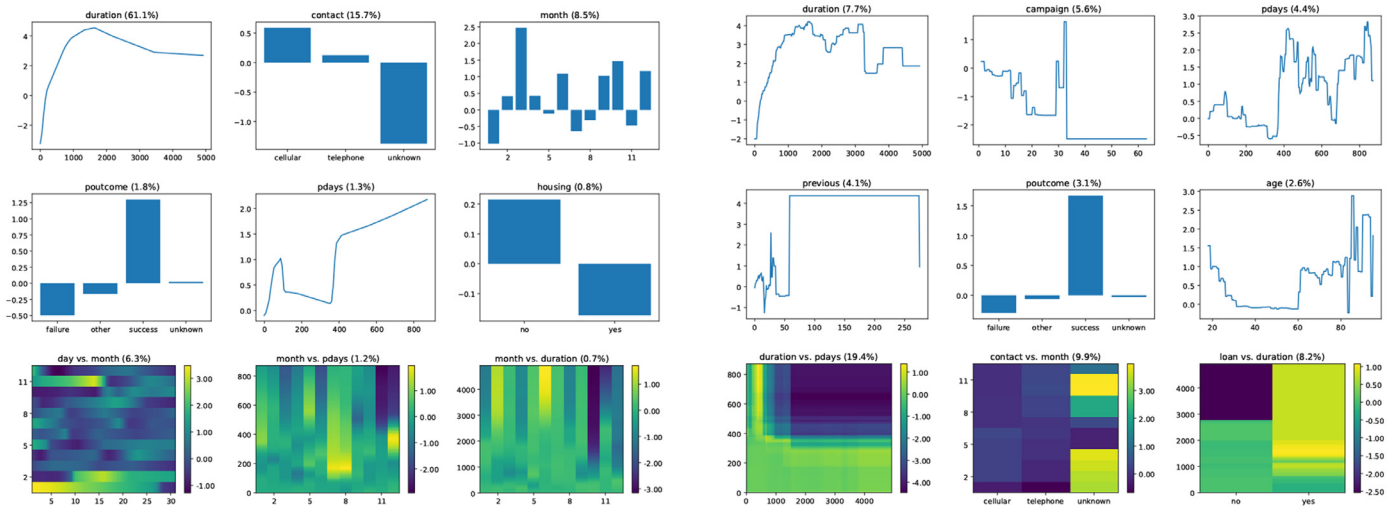


(c) EBM

Fig. 4. The fitted results of GAMI-Net and EBM vs. the ground truth of the synthetic function (uniform distribution; $n = 10000$).

Table 2
Comparison results of the synthetic function (uniform distribution; $n = 10000$).

Model		Train RMSE	Val-RMSE	Test RMSE	Clarity Loss
XGBoost		0.135±0.103	1.630±0.046	1.634±0.013	-
RF		1.500±0.019	1.840±0.060	1.822±0.043	-
MLP		2.320±0.076	2.372±0.119	2.615±0.057	-
GLM		3.062±0.039	3.123±0.094	3.103±0.065	-
pyGAM		2.253±0.025	2.459±0.057	2.450±0.046	-
EBM		1.155±0.157	1.211±0.165	1.799±0.132	0.0007±0.0001
GAMI-Net	$\lambda = 10^0$	1.004±0.009	1.054±0.020	1.044±0.020	0.0002±0.0001
	$\lambda = 10^{-1}$	0.977±0.023	1.056±0.018	1.051±0.019	0.0003±0.0001
	$\lambda = 10^{-2}$	0.976±0.013	1.050±0.023	1.046±0.024	0.0006±0.0004
	$\lambda = 10^{-3}$	0.967±0.016	1.068±0.020	1.061±0.018	0.0031±0.0022
	$\lambda = 10^{-4}$	0.953±0.010	1.063±0.019	1.060±0.019	0.0042±0.0026
	No Heredity $\lambda = 10^0$	1.004±0.013	1.051±0.023	1.045±0.019	0.0002±0.0001



(a) GAMI-Net (b) EBM

Fig. 5. The GAMI-Net and EBM global interpretation for the bank marketing dataset.

the search space of interactions and make the model structurally more interpretable. The possible drawback of heredity lies in that it may reduce the predictive performance. However, it is observed that the inclusion of heredity constraint does not have a significant influence on predictive performance. Therefore, the use of heredity constraint can be viewed as a bonus term for GAMI-Net.

3.3. Bank marketing dataset

This dataset is typically used in a binary classification setting (https://archive.ics.uci.edu/ml/datasets/Bank+Marketing). It has 45211 samples with 9 categorical variables and 7 numerical variables, denoting a client's age, education, job, and other related information. The goal is to predict whether a client would subscribe to the term deposit. Note that the variable "duration" is included just for benchmarking purpose, but not for realistic predictive modeling.

Due to the limit of page size, we only present the top 6 main effects and top 3 pairwise interactions, which is a subset of the selected 12 (out of 16) main effects and 4 (out of 20) pairwise interactions. The top 3 important variables are “duration” (last contact duration, in seconds), “contact” (contact communication type), and “month” (last contact month of year). The most significant pairwise interaction is “day vs. month”. The fitted results of EBM in Fig. 5b are rather difficult to interpret. For instance, there exist significant fluctuations as the variable “age” is greater than 80. It is hard to explain why the predicted outcomes of a client change greatly when his age increases from 80 to 90. In contrast, GAMI-Net is free from these problems as its fitted model is continuous and even smooth. Finally, the local interpretability of GAMI-Net is demonstrated in Fig. 6, which shows the prediction diagnosis of one sample point.

The comparison results of different methods are shown in Table 3, together with the GAMI-Net results under different settings. From the results, we can obtain similar conclusions to that of the simulation study. With the increase of λ, the marginal clarity losses get decreased while the predictive performance does not change too much until $\lambda = 10^{-1}$. Therefore, λ is set to 10^{-1} for this dataset to pursue a good balance between model interpretability and predictive performance. In addition, it is observed that GAMI-Net with and without the heredity constraint achieve almost the same accuracy. It indicates that the use of heredity constraint can help reduce the search space during model estimation, without sacrificing the predictive performance.

3.4. Bike sharing hour dataset

The bike sharing hour dataset is a regression task with 17379 samples and 12 explanatory variables (https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset). Each sample records the basic environmental information, including 8 categorical variables, e.g., the season and the weather situation; and 4 numerical variables, e.g., the temperature and wind speed. The target is to predict the hourly count of rental bikes in the capital bike share system between 2011 and 2012.

Similarly, the global interpretation of the bike sharing hour dataset is shown in Fig. 7. In total, 9 (out of 12) main effects and 17 (out of 20) pairwise interactions are shown to be non-trivial. The most important variable is “hr” (hour, ranges from 0 to 23) with IR equals to 57.1%. It can be observed that there exist 2 peaks of bike sharing around 8 AM and 5 PM, which correspond to the rush hour in a day. The categorical variable “yr” (year, 0 denotes 2011 and 1 means 2012) is the second important one, and the results show that there exists an increasing trend of bike sharing over time. The third important variable is “atemp” (normalized feeling temperature in Celsius divided to 50), and the most appealing temperature is around 30 degrees Celsius. Regarding pairwise interactions, the “hr vs. workingday” is shown to be the most important.

The comparison results of the bike sharing hour dataset are shown in Table 4, in which the consistent conclusions can be drawn. The default marginal clarity regularization strength is also set to 10^{-1} considering the balance between predictive performance and model interpretability. Note that XGBoost achieves extremely small RMSE in the training data, and it also ranks the best regarding test performance. That is, although XGBoost overfits the training data, it still has better generalization performance than other compared models.

3.5. More real-world datasets

In addition to the simulation study and 2 real-world applications, we test the predictive performance of the proposed GAMI-Net on another 20 regression datasets, collected from different domains. These datasets are available in the UCI machine learning

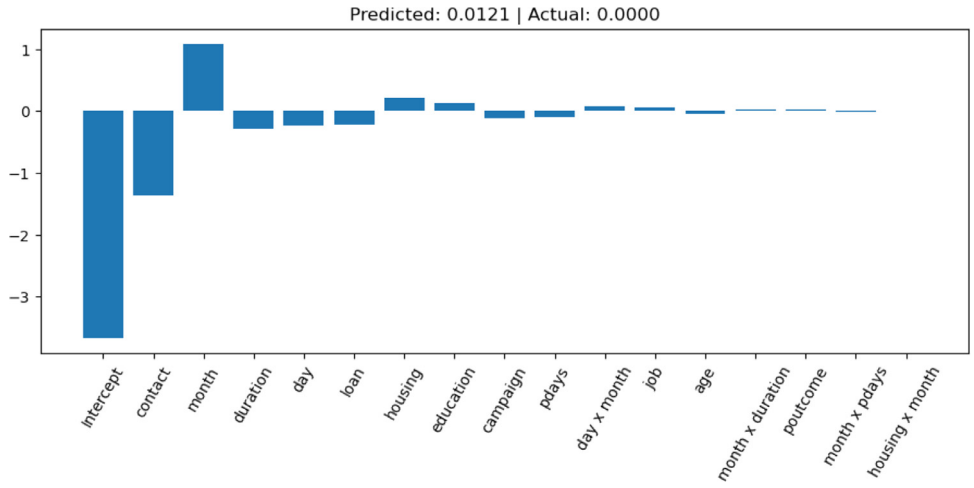
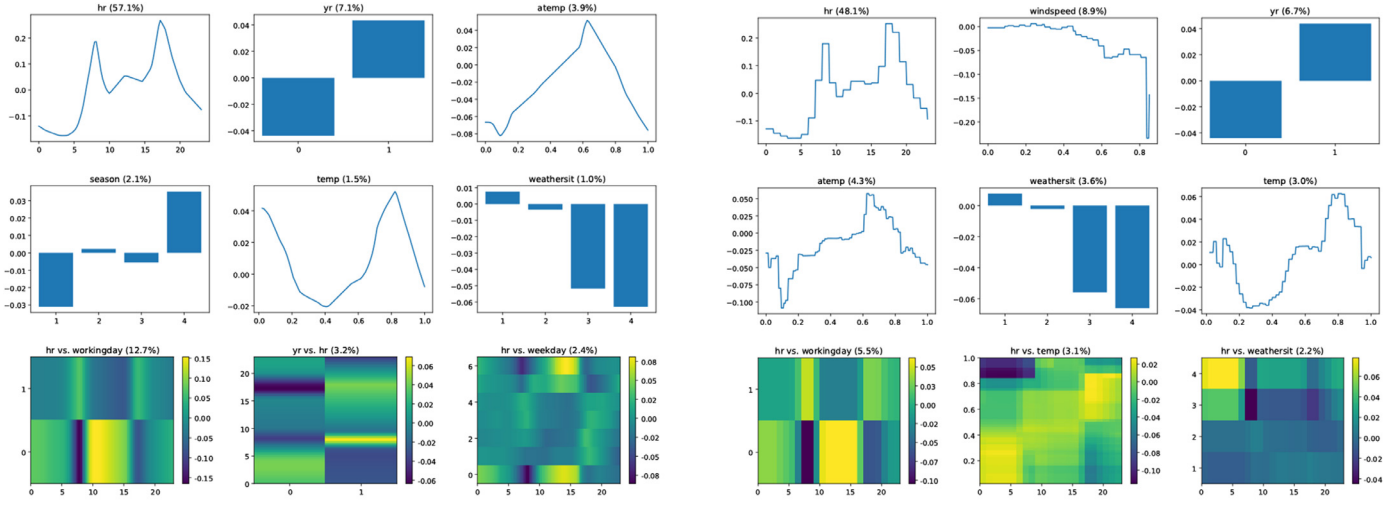


Fig. 6. The GAMI-Net local interpretation of one sample point in the bank marketing dataset.

Table 3
Comparison results of the bank marketing dataset.

Model		Train AUC (%)	Val-AUC (%)	Test AUC (%)	Clarity Loss
XGBoost		97.38±0.09	93.24±0.52	93.01±0.26	-
RF		93.59±0.10	92.01±0.62	91.98±0.35	-
MLP		93.15±0.65	93.12±0.90	92.29±0.41	-
GLM		90.81±0.21	90.77±0.47	90.60±0.30	-
pyGAM		91.90±0.17	91.68±0.39	91.53±0.36	-
EBM		94.59±0.16	94.63±0.40	93.16±0.31	1.0228±0.0755
GAMI-Net	$\lambda = 10^{0}$	92.21±0.37	92.13±0.60	91.89±0.45	0.0012±0.0010
	$\lambda = 8 \times 10^{-1}$	92.47±0.35	92.40±0.55	92.15±0.47	0.0018±0.0010
	$\lambda = 6 \times 10^{-1}$	92.76±0.18	92.68±0.43	92.42±0.29	0.0034±0.0022
	$\lambda = 4 \times 10^{-1}$	93.01±0.15	92.87±0.47	92.65±0.36	0.0059±0.0018
	$\lambda = 2 \times 10^{-1}$	93.43±0.17	93.19±0.47	93.00±0.34	0.0092±0.0044
	$\lambda = 10^{-1}$	93.80±0.13	93.39±0.47	93.21±0.41	0.0258±0.0141
	$\lambda = 10^{-2}$	94.12±0.15	93.44±0.41	93.28±0.38	0.0874±0.0212
	$\lambda = 10^{-3}$	94.12±0.14	93.41±0.46	93.26±0.36	0.4293±0.1094
	$\lambda = 10^{-4}$	94.04±0.10	93.36±0.45	93.22±0.37	1.3240±0.2679
	No Heredity $\lambda = 10^{-1}$	93.81±0.12	93.41±0.47	93.23±0.39	0.0235±0.0109



(a) GAMI-Net (b) EBM

Fig. 7. The GAMI-Net and EBM global interpretation for the bike sharing hour dataset.

Table 4
Comparison results of the bike sharing hour dataset.

Model		Train RMSE	Val-RMSE	Test RMSE	Clarity Loss
	XGBoost	5.64±4.52	40.96±1.17	42.33±0.72	-
	RF	62.00±0.72	65.12±1.47	65.69±2.10	-
	MLP	35.54±1.31	37.09±1.52	43.35±1.32	-
	GLM	158.98±0.52	158.54±2.49	159.09±1.57	-
	pyGAM	99.52±0.60	100.16±1.61	100.22±1.12	-
	EBM	54.32±0.55	54.27±0.99	57.48±1.20	0.0026±0.0005
GAMI-Net	$\lambda = 10^{0}$	52.97±0.45	55.26±1.34	55.90±1.19	0.0006±0.0002
	$\lambda = 8 \times 10^{-1}$	51.95±0.79	54.54±1.23	55.10±1.22	0.0006±0.0002
	$\lambda = 6 \times 10^{-1}$	51.59±1.01	54.18±1.49	54.62±1.28	0.0007±0.0002
	$\lambda = 4 \times 10^{-1}$	50.54±0.79	53.35±1.05	53.84±1.01	0.0006±0.0002
	$\lambda = 2 \times 10^{-1}$	50.39±0.61	53.27±1.21	53.75±0.90	0.0008±0.0002
	$\lambda = 10^{-1}$	50.13±1.16	53.18±1.36	53.63±0.83	0.0008±0.0002
	$\lambda = 10^{-2}$	49.83±0.80	52.95±1.39	53.24±0.81	0.0014±0.0003
	$\lambda = 10^{-3}$	49.88±0.59	53.00±1.26	53.23±0.94	0.0047±0.0017
	$\lambda = 10^{-4}$	49.83±0.78	53.00±1.19	53.37±0.96	0.0142±0.0040
	No Heredity $\lambda = 10^{-1}$	50.13±1.09	53.18±1.33	53.66±0.83	0.0008±0.0002

Table 5
Test set RMSE on 20 real-world regression datasets.

Dataset	*n*	*p*	GAMI-Net	EBM	pyGAM	GLM	MLP	RF	XGBoost	Scale
no2	500	7	4.992±0.484	**4.681** ±0.396	4.971±0.514	6.508±0.626	5.201±0.466	**4.688** ±0.422	4.911±0.342	×0.1
sensory	576	11	7.284±0.437	**7.054** ±0.475	7.923±0.279	8.066±0.226	7.455±0.339	**7.318** ±0.497	8.205±0.549	×0.1
disclosure z	662	3	2.432±0.261	2.438±0.275	**2.429** ±0.277	2.438±0.267	**2.442** ±0.270	2.445±0.260	2.856±0.184	×10000
bike share day	731	11	0.688±0.028	**0.663** ±0.043	0.710±0.036	1.144±0.045	0.827±0.042	0.727±0.051	**0.711** ±0.067	×1000
era	1000	4	**1.566** ±0.036	1.566±0.038	1.568±0.039	1.684±0.041	1.583±0.041	**1.573** ±0.041	1.596±0.045	×1
treasury	1049	15	2.197±0.269	2.513±0.400	**2.114** ±0.260	8.971±0.739	**2.367** ±0.282	2.416±0.324	2.489±0.247	×0.1
weather izmir	1461	9	**1.116** ±0.133	1.322±0.073	1.150±0.131	3.231±0.133	**1.289** ±0.129	1.303±0.103	1.337±0.095	×1
airfoil	1503	5	**2.101** ±0.149	2.169±0.100	4.563±0.194	6.357±0.205	2.607±0.269	2.440±0.126	**1.742** ±0.161	×1
wine red	1599	11	6.225±0.153	**5.991** ±0.231	6.252±0.129	7.473±0.180	6.201±0.165	**5.918** ±0.210	6.135±0.251	×0.1
skill craft	3395	18	0.969±0.025	**0.920** ±0.026	1.154±0.531	1.200±0.022	1.019±0.075	**0.927** ±0.025	0.997±0.021	×1
abalone	4177	8	**2.133** ±0.066	2.240±0.052	2.168±0.091	2.975±0.126	**2.142** ±0.078	2.186±0.074	2.372±0.066	×1
parkinsons tele	5875	19	**0.377** ±0.038	0.412±0.008	0.771±0.035	1.061±0.018	0.580±0.026	0.321±0.014	**0.198** ±0.009	×10
wind	6574	14	**3.050** ±0.068	3.085±0.064	3.071±0.062	4.590±0.064	**3.046** ±0.071	3.258±0.069	3.205±0.095	×1
cpu small	8192	12	0.288±0.008	**0.286** ±0.010	0.327±0.011	1.464±0.049	0.310±0.006	0.314±0.011	**0.294** ±0.023	×10
topo 2 1	8885	266	2.884±0.320	**2.873** ±0.312	3.054±0.337	2.940±0.318	2.892±0.314	**2.864** ±0.318	3.067±0.301	×0.01
ccpp	9568	4	3.866±0.070	**3.673** ±0.069	4.087±0.084	6.143±0.053	4.157±0.084	3.705±0.073	**3.121** ±0.090	×1
electrical grid	10000	11	**0.935** ±0.023	0.951±0.016	1.718±0.020	2.885±0.049	**0.658** ±0.019	1.448±0.028	0.993±0.016	×0.01
ailerons	13750	40	**1.660** ±0.049	**1.660** ±0.049	1.690±0.054	3.380±0.087	**1.590** ±0.030	1.680±0.040	1.650±0.050	×0.0001
elevators	16599	18	**2.218** ±0.042	2.284±0.087	2.389±0.053	6.710±0.167	**2.106** ±0.070	3.147±0.064	2.167±0.041	×0.001
california housing	20640	8	**5.118** ±0.119	5.235±0.084	6.504±0.327	9.160±0.091	5.777±0.169	5.796±0.095	**4.708** ±0.087	×0.1

Table 6
Pairwise comparison of test set RMSE for different models.

	GLM	pyGAM	MLP	RF	XGBoost	EBM	GAMI-Net
GLM	-	1 (1)	1 (0)	1 (0)	3 (2)	0 (0)	0 (0)
pyGAM	19 (17)	-	7 (7)	8 (5)	9 (7)	5 (3)	3 (1)
MLP	19 (19)	13 (11)	-	10 (5)	11 (9)	7 (5)	5 (3)
RF	19 (19)	12 (10)	10 (8)	-	10 (9)	6 (4)	6 (6)
XGBoost	17 (17)	11 (8)	9 (7)	10 (10)	-	7 (5)	8 (5)
EBM	20 (19)	15 (13)	13 (12)	14 (10)	13 (11)	-	8 (5)
GAMI-Net	20 (19)	17 (14)	15 (11)	14 (10)	12 (10)	11 (8)	-

repository or the OpenML platform. The sample sizes range from 500 (no2) to 20640 (california housing), and the number of variables varies from 3 (disclosure z) to 266 (topo 2 1). The detailed information of each data and numerical results is presented in Table 5, where the test set performance of each compared method is reported. Note that the reported results are all rounded to a certain precision, while the best performer (based on the full precision) is highlighted in bold. Besides, the listed results should be multiplied by the corresponding scaling factors in the last column.

Table 6 presents the pairwise comparison results of the compared methods. The numbers indicate how often the methods in row (significantly; by paired *t*-test over the 10 repetitions, with a *p*-value of 0.05) outperform the methods in column, which are counted using the 20 real-world datasets. For instance, GAMI-Net outperforms EBM in 11 out of the 20 datasets, in which 8 are tested to be significant; in contrast, EBM beats GAMI-Net in only 8 datasets, and they show the same performance on the ailerons dataset.

Generally speaking, GAMI-Net shows comparative predictive performance to EBM and other benchmark models. GAMI-Net is more likely to have better predictive performance when the actual shape functions are continuous and smooth, while EBM is more likely to perform better when the shape functions are piece-wise constant. Both of them are competitive regarding predictive power, while the proposed GAMI-Net is designed with more interpretability considerations. Therefore, GAMI-Net is a promising tool in the area of interpretable machine learning.

4. Conclusion

In this paper, an intrinsically explainable neural network called GAMI-Net is proposed. It approximates the complex functional re-

lationship using subnetwork-represented main effects and pairwise interactions, which can be easily interpreted using 1D line plots / bar charts and 2D heatmaps. Several statistically meaningful constraints are considered to enhance the model interpretability, including the heredity constraint for enforcing structural pairwise interactions, the sparsity constraint for promoting model parsimony, and the marginal clarity constraint for avoiding the effects mixing problem. The experimental results show that the proposed model has competitive predictive performance to black-box machine learning models. Meanwhile, the model estimated by GAMI-Net is highly interpretable and easily visualizable.

To extend GAMI-Net, the following topics are promising. One direction is to consider additional shape constraints for each component function, e.g., monotonic increasing / decreasing, convex or concave, according to prior experience or domain knowledge. Another direction is to consider higher-order interactions for more sophisticated developments.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. Tang, X. Huang, M. Chen, C. Sun, J. Yang, Adversarial attack type I: cheat classifiers by significant changes, IEEE Trans. Pattern Anal. Mach. Intell. 43 (3) (2021) 1100–1109, doi:10.1109/TPAMI.2019.2936378.

[2] Y. Shi, Y. Han, Q. Zhang, X. Kuang, Adaptive iterative attack towards explainable adversarial robustness, Pattern Recognit. 105 (2020) 107309.

[3] W.J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, B. Yu, Definitions, methods, and applications in interpretable machine learning, Proc. Natl. Acad. Sci. U. S. A. 116 (44) (2019) 22071–22080.

[4] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. 29 (5) (2001) 1189–1232.

[5] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 1135–1144.

[6] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, 2017, pp. 4765–4774.

[7] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, K.-R. Müller, Explaining nonlinear classification decisions with deep taylor decomposition, Pattern Recognit. 65 (2017) 211–222.

[8] D. Liu, L. Zhang, T. Luo, L. Tao, Y. Wu, Towards interpretable and robust hand detection via pixel-wise prediction, Pattern Recognit. 105 (2020) 107202.

[9] J.H. Friedman, W. Stuetzle, Projection pursuit regression, J. Am. Stat. Assoc. 76 (376) (1981) 817–823.

[10] J.-N. Hwang, S.-R. Lay, M. Maechler, R.D. Martin, J. Schimert, Regression modeling in back-propagation and projection pursuit learning, IEEE Trans. Neural Netw. 5 (3) (1994) 342–353.

[11] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, New York, NY: Springer, 2009.

[12] J. Vaughan, A. Sudjianto, E. Brahimi, J. Chen, V.N. Nair, Explainable neural networks based on additive index models, The RMA Journal (2018) 40–49.

[13] Z. Yang, A. Zhang, A. Sudjianto, Enhancing explainability of neural networks through architecture constraints, IEEE Trans. Neural Netw. Learn. Syst. 32 (6) (2021) 2610–2621, doi:10.1109/TNNLS.2020.3007259.

[14] T. Hastie, R. Tibshirani, Generalized Additive Models, Chapman & Hall, London, 1990.

[15] Y. Lou, R. Caruana, J. Gehrke, Intelligible models for classification and regression, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2012, pp. 150–158.

[16] R. Agarwal, N. Frosst, X. Zhang, R. Caruana, G.E. Hinton, Neural additive models: interpretable machine learning with neural nets. arXiv preprint arXiv: 2004.13912.

[17] Y. Lou, R. Caruana, J. Gehrke, G. Hooker, Accurate intelligible models with pairwise interactions, in: ACM SIGKDD international conference on Knowledge Discovery and Data Mining, ACM, 2013, pp. 623–631.

[18] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2015, pp. 1721–1730.

[19] H. Nori, S. Jenkins, P. Koch, R. Caruana, InterpretML: a unified framework for machine learning interpretability. arXiv preprint arXiv:1909.09223.

[20] P. Ravikumar, J. Lafferty, H. Liu, L. Wasserman, Sparse additive models, J. R. Stat. Soc. Ser. B Methodol. 71 (5) (2009) 1009–1030.

[21] J.A. Nelder, The selection of terms in response-surface models how strong is the weak-heredity principle? Am. Stat. 52 (4) (1998) 315–318.

[22] P. McCullagh, J.A. Nelder, Generalized Linear Models, Chapman & Hall, London, 1983.

[23] J. Bien, J. Taylor, R. Tibshirani, A lasso for hierarchical interactions, Ann. Stat. 41 (3) (2013) 1111–1141.

[24] N.H. Choi, W. Li, J. Zhu, Variable selection with the strong heredity constraint and its oracle property, J. Am. Stat. Assoc. 105 (489) (2010) 354–364.

[25] M. Yuan, V.R. Joseph, H. Zou, Structured variable selection and estimation, Ann. Appl. Stat. 3 (4) (2009) 1738–1757.

[26] J.L. Peixoto, Hierarchical variable selection in polynomial regression models, Am. Stat. 41 (4) (1987) 311–313.

[27] X. Li, N. Sudarsanam, D.D. Frey, Regularities in data from factorial experiments, Complexity 11 (5) (2006) 32–45.

[28] G. Hooker, Generalized functional anova diagnostics for high-dimensional functions of dependent variables, J. Comput. Graph. Stat. 16 (3) (2007) 709–732.

[29] B. Lengerich, S. Tan, C.-H. Chang, G. Hooker, R. Caruana, Purifying interaction effects with the functional ANOVA: An efficient algorithm for recovering identifiable additive models, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 2402–2412.

[30] M. Tsang, D. Cheng, Y. Liu, Detecting statistical interactions from neural network weights, in: International Conference on Learning Representations, 2018.

[31] I.M. Sobol, Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates, Math. Comput. Simul. 55 (1–3) (2001) 271–280.

[32] D. Servén, C. Brummitt, pyGAM: Generalized additive models in python, 2018, 10.5281/zenodo.1208723

Zebin Yang is currently pursuing the Ph.D. degree in the Department of Statistics and Actuarial Science, The University of Hong Kong. His research interests include machine learning and its application in decision making.

Aijun Zhang leads self-explanatory machine learning in Corporate Model Risk of Wells Fargo. His research interests include experimental design, machine learning, and explainable artificial intelligence. He holds Ph.D. degree in Statistics from the University of Michigan at Ann Arbor and has published about 30 papers in professional conferences and journals.

Agus Sudjianto is Executive Vice President and Head of Corporate Model Risk at Wells Fargo, where he is responsible to lead Enterprise-wide Model Risk Management. His research interests include quantitative risk, statistical and machine learning in finance, and statistical computing. He has published in various journals and co-authored a monograph in design and modeling for computer experiments. He holds M.S. and Ph.D. in Engineering from Wayne State University and Massachusetts Institute of Technology.