Scientific Reports

https://doi.org/10.1038/s41598-026-43928-4

Article in Press

The most important features in generalized additive models might be groups of features

Received: 24 June 2025

Accepted: 9 March 2026

Published online: 21 March 2026

Cite this article as: Bosschieter T., França L., Wolk J. *et al.* The most important features in generalized additive models might be groups of features. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-43928-4

Tomas Bosschieter, Luis França, Jessica Wolk, Yiyuan Wu, Bella Mehta, Joseph Dehoney, Orsolya Kiss, Fiona C. Baker, Qingyu Zhao, Rich Caruana & Kilian M. Pohl

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

ARTICLE IN PRESS

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

The Most Important Features in Generalized Additive Models Might Be Groups of Features

Tomas Bosschieter[1,*], **Luis França**[2], **Jessica Wolk**[2], **Yiyuan Wu**[3], **Bella Mehta, M.D.**[3], **Joseph Dehoney**[4], **Orsolya Kiss**[5], **Fiona C. Baker**[5], **Qingyu Zhao**[6], **Rich Caruana**[7], **and Kilian M. Pohl**[4]

[1]Stanford University, Institute for Computational and Mathematical Engineering, Stanford, CA 94305, USA
[2]Microsoft Research, Redmond, WA 98052, USA
[3]Hospital for Special Surgery, New York City, NY 10021, USA
[4]Stanford University, Department of Psychiatry & Behavioral Sciences, Stanford, CA 94305, USA
[5]SRI International, Menlo Park, CA 94025, USA
[6]Cornell University, Weill Cornell Medicine, New York City, NY 10065, USA
[7]Intelligible, Inc., Seattle, WA 98700, USA
[*]Corresponding author: tomasbos@alumni.stanford.edu

ABSTRACT

While analyzing the importance of features has become ubiquitous in interpretable machine learning, the joint signal from a group of related features is sometimes overlooked or inadvertently excluded. Neglecting the joint signal could bypass a critical insight: in many instances, the most significant predictors are not isolated features, but rather the combined effect of groups of features. This can be especially problematic for datasets that contain natural groupings of features, including multimodal datasets. This paper introduces a novel approach to determine the importance of a group of features for Generalized Additive Models (GAMs) that is efficient, requires no model retraining, allows defining groups posthoc, permits overlapping groups, and remains meaningful in high-dimensional settings. We showcase properties of our method on three synthetic experiments that illustrate the behavior of group importance across various data regimes. We then demonstrate the importance of groups of features in identifying depressive symptoms from a multimodal neuroscience dataset, and study the importance of social determinants of health after total hip arthroplasty. These two case studies reveal that analyzing group importance offers a more accurate, holistic view of medical issues compared to a single-feature analysis.

1 Introduction

As machine learning models become ever-more ubiquitous in healthcare, there is a growing necessity for model interpretability. While it has proven difficult to reach a consensus on the meaning of interpretability[1], one particularly common instrument is determining the importance of individual features (aka feature importance or feature attribution), which in and of itself also does not have one single definition. Common definitions of feature importance include the average contribution of a feature to prediction in a separable model[2–4], average impurity decrease in tree-based models[5,6], and the average increase in the model's error or loss after shuffling the values of a feature (i.e., permutation feature importance)[7,8].

Providing an interpretation based on the importance of individual features is often challenging due to feature correlation and plausibility issues[9]. For a group of correlated features, their joint signal could be crucial for prediction, while the individual features in the group might be attributed low feature importance[10]. As a result, there could be signal in groups of correlated features that might be lost when analyzing individual features by themselves. Even worse, these features could be filtered out during feature selection and thereby diminish the signal for inference[6]. This could be particularly troublesome on data containing natural groupings of features, such as modalities in multimodal data. Furthermore, interactions between features (in a group) are not captured through feature importance[11]. One way to address these issues is via group importance (aka grouped feature importance), i.e., by grouping features together and determining the importance of the group[12,13].

Group importance has previously been proposed using LASSO[14], the Penalty Decomposition[15], Shapley values[16], and permutation testing[8,17]. Unfortunately, most of these methods come with computational challenges, mainly originating from the need to retrain the underlying machine learning model or compute expensive gradients. In the case of groupSHAP[16], for example, group-level attributions require estimating Shapley values over subsets of features or feature groups, which involves evaluating the model on a large number of masked feature combinations. Exact Shapley values require an exponential number of masked combinations in terms of the number of features. While efficient approximations of Shapley values exist for tree-based models, e.g. TreeSHAP, Shapley values remain computationally expensive for general models and large datasets.

To address these challenges, we propose a robust, versatile, and computationally efficient method for computing group importance for Generalized Additive Models (GAMs), a very broad class of interpretable models that includes linear, logistic, and generalized linear models. GAMs with interactions[2,3] are typically of the form $g(\mathbb{E}(y)) = f_1(x_1) + \cdots + f_n(x_n) + \sum_{i \neq j} f_{i,j}(x_i, x_j)$, where g is a link function, y the target, x_i are features, β_0 is an intercept, and $f_i(x_i)$ and $f_{i,j}(x_i, x_j)$ are shape functions (aka component functions). In this setting, the (individual) importance I_{x_j} of a feature x_j is typically computed as the average absolute contribution of the corresponding shape function $f_j(\cdot)$[4], i.e.,

$$I_{x_j} = \frac{1}{|\mathscr{T}|} \sum_{t \in \mathscr{T}} |f_j(t_j)|, \tag{1}$$

where $t \in \mathscr{T}$ represents samples t in training data $\mathscr{T}$. For a group $G = \{x_{i_1}, \ldots, x_{i_k}, (x_{\ell_1}, x_{\ell_2}), \ldots, (x_{\ell_{k'-1}}, x_{\ell_{k'}})\}$ consisting of both individual features x_{i_j} as well as interaction terms $(x_{\ell_i}, x_{\ell_{i+1}})$, we extend this definition and propose

$$I_G := \frac{1}{|\mathscr{T}|} \sum_{t \in \mathscr{T}} |f_{i_1}(t_{i_1}) + \cdots + f_{i_k}(t_{i_k}) + f_{\ell_1,\ell_2}(t_{\ell_1}, t_{\ell_2}) + \cdots + f_{\ell_{k'-1},\ell_{k'}}(t_{\ell_{k'-1}}, t_{\ell_{k'}})|, \tag{2}$$

with mathematical justification given in Section 2.3. This definition of group importance has several properties we consider valuable, including:

- groups can be defined post hoc (and thus do not affect the model) and may be overlapping,
- computing group importance is computationally inexpensive due to the separability of the GAM,
- it remains meaningful when there are many correlated features, unlike, e.g., group permutations,
- for a group consisting of one feature, its group importance equals the feature's individual importance,
- adding useless features, i.e. random noise, to a group does not change its group importance (up to small error ε),
- both individual features and interaction terms can be included in the group G.

We implement our method for an Explainable Boosting Machine (EBM), a GAM-based model that achieves state-of-the-art accuracy while providing interpretability[2,4]. EBMs have recently gained popularity for their application in fields such as concrete compressive strength[18], brain tumors[19], Parkinson's disease[20], and revealing healthcare biases that result in significant loss of life[21]. We evaluate the properties of our definition of group importance on three synthetic experiments and two real-world datasets that contain natural groupings of features. In the first case study, we identify the depressive symptom of negative valence[8,22] on a neuroscience dataset consisting of measurements from functional brain networks and groups (aka domains) such as sleep, traumatic experiences, biographic data, and social support. In the second case study, we focus on the influence of community-level social determinants of health in predicting mortality for total hip arthroplasty using a large-scale healthcare dataset ($>$ 100k samples). These determinants, encompassing environmental, economic, and social conditions, are paramount in understanding health outcomes. Our main contributions are as follows:

- Proposing a definition for group importance for Generalized Additive Models that is computationally inexpensive.
- Showing that group importance offers enhanced interpretability in addition to individual feature importance, especially for correlated features, as demonstrated on two real-world datasets.
- Demonstrating group importance can be used as a low-cost, yet effective method that achieves similar importances as groupSHAP in a real-world dataset[16]. It is also a low-cost method for feature selection.

2 Methods

We first describe Explainable Boosting Machines (EBMs) as a specific implementation of a Generalized Additive Model (GAM), and how feature importance is typically computed in this setting. Then, we extend this approach to define group importance. Subsequently, we turn to feature selection and briefly discuss the parallel between group importance and explained variation in statistics.

2.1 Explainable Boosting Machine

Explainable Boosting Machines (EBMs)[2] are Generalized Additive Models (GAMs)[3] with interaction terms, i.e.,

$$g(\mathbb{E}[y]) = \beta_0 + f_1(x_1) + \cdots + f_n(x_n) + \sum_{i \neq j} f_{i,j}(x_i, x_j). \tag{3}$$

The shape functions $f_i(\cdot)$ and $f_{i,j}(\cdot,\cdot)$ are trained using cyclic gradient boosting with shallow learners[2]. During training, a *purification process*[4,23] ensures that the contribution of a feature x_k for prediction is not moved between $f_k(x_k)$ and its interaction terms $f_{k,\ell}(x_k,x_\ell)$ for $\ell \in \{1,\ldots,n\}\backslash\{k\}$. That is, $f_k(x_k)$ encodes the sole and total contribution of x_k to prediction (i.e., the main effect), while $f_{k,\ell}(x_k,x_\ell)$ encodes the interaction term adjusted for all main effects.

Thus, the shape functions form a functional ANOVA decomposition[23,24] through a functional ANOVA projection so that all lower-order effects integrate to zero, ensuring each main effect and interaction captures only its unique contribution. For enhanced interpretability, these functions are zero-centered and can be visualized to show the contribution of each feature or interaction term to prediction. Assuming that the GAM satisfies the functional ANOVA decomposition (e.g., after purification), there is no projection or correlation between higher-order interaction terms and lower-order interactions or univariate terms. For example, 3-bit parity is uncorrelated with and has no projection on 2-bit parity or the constituent univariate features. Thus if the 3-bit parity interaction is not included in the model, the contribution of 3-bit parity to the true generating function is left unmodeled and does not “pollute" the lower-order terms. This is part of what makes functional ANOVA useful for the interpretation of GAMs and why the EBM algorithm is designed to approximate functional ANOVA. It is also why purification is important when the approximation to the functional ANOVA is poor[4,24]. In the absence of purification (and functional ANOVA), the signal of important higher-order interactions that are not included in the additive structure of the GAM may be distributed across lower-order terms of the corresponding features. This potentially affects the interpretation of both individual and group-level importances, which is a limitation if the functional ANOVA (e.g., purification) is not employed[2,4,24].

EBMs have been shown to be stable even under a low prevalence regime of the target, e.g. in credit card fraud (0.17%)[25], various adverse pregnancies outcomes (0.4% – 3.4%)[26], and intensive care unit readmission (8.5%)[27]. However, if the sample size is so low relative to the number of features and interaction terms included in the model that the shape functions themselves become unreliable, then the reliability of the feature importances also will be reduced.

The additivity of the GAM makes it easy to define an importance metric. That is, individual, ‘global’ feature importance is typically defined as the average *absolute* contribution of a feature to prediction over all samples in the training dataset[4]. Notably, feature importance can sometimes refer to *local* feature importance, which attributes importance to features for one sample prediction specifically[4], but we only consider global feature importance in this paper. That is, for feature x_j and samples $t \in \mathscr{T}$ taking on values $(t_1,\ldots,t_n)$, the feature importance I_{x_j}[4] of x_j is typically defined as

$$I_{x_j} = \frac{1}{|\mathscr{T}|}\sum_{t\in\mathscr{T}} |f_j(t_j)|. \tag{4}$$

Note, each contribution $|f_j(t_j)|$ can include weights defined by the user[4]. This approach not only provides insights into the internal mechanisms of the model, but also ensures a symmetric attribution of importance, i.e., features signaling very high or low risk receive similar importance (a direct consequence of the zero-centering in the shape functions). Notably, the importance of a feature x_1 is not the cumulative contribution of all model terms that contain x_1; instead, its feature importance is defined solely as the average absolute contribution to prediction of the main-effect shape function $f_1(x_1)$, which is adjusted for all other model terms. Feature importance for GAMs under the functional ANOVA assumption is different from feature importance for alternate methods such as SHAP.

2.1.1 Correlated features

Note that, by design, EBMs split the shared signal of highly correlated features evenly. This ensures that each feature receives a proportionate amount of signal and importance. This is a property worth highlighting, as there also exist alternate approaches. For example, other GAM implementations might explicitly aim to assign all importance to one of the features rather than splitting it evenly, such as in GRAND-SLAMIN’[28].

2.2 Group importance

Similar to individual feature importance, group importance has been well-studied and has multiple definitions, including through Shapley values[16] and permutation testing[8,17]. Group importance is understood differently across various domains, each with its own set of definitions and accompanying challenges[12,13]. This lack of consensus might lead to arbitrary selection of definitions rather than standardization. We aim to bridge this gap for GAMs specifically and propose a definition that naturally extends the notion of individual feature importance.

One example of a definition of total importance could be a simple sum of individual importances, but this leads to a misleading scale when features exhibit substantial correlation, see Section 3. Another common definition is groupSHAP[16]. While (group)SHAP is very popular, such group-level attributions require estimating Shapley values over exponentially many subsets of features, which involves evaluating the model on a large number of masked feature combinations, resulting in potentially severe computational challenges.

2.3 Group importance for GAMs

To post hoc evaluate the importance of a group of features after training the GAM, we average the absolute values of the sum of feature contributions in a group over all samples in the training dataset $\mathcal{T}$. That is, let group $G = \{x_{i_1}, \ldots, x_{i_k}, (x_{\ell_1}, x_{\ell_2}), \ldots, (x_{\ell_{k'-1}}, x_{\ell_{k'}})\}$ be a set of features whose group importance I_G we wish to evaluate. Then, we define group importance as

$$I_G := \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} |f_{i_1}(t_{i_1}) + \cdots + f_{i_k}(t_{i_k}) + f_{\ell_1,\ell_2}(t_{\ell_1}, t_{\ell_2}) + \cdots + f_{\ell_{k'-1},\ell_{k'}}(t_{\ell_{k'-1}}, t_{\ell_{k'}})|, \tag{5}$$

This is a canonical choice given that

1. it is a direct extension of the $k = 1$ case, with the property that a feature's individual importance equals the importance of a group consisting of only that feature,
2. the *null* group (i.e. group consisting of no features) has 0 importance,
3. group importance satisfies $0 \leq I_{\{x_{i_1}, \ldots, x_{i_k}\}} \leq I_{x_{i_1}} + \cdots + I_{x_{i_k}}$.

For (ii), the null group has zero importance as the summand in Eq. (5) is always zero. Observe that (iii) is a particularly important and useful property, as it represents the triangle inequality for grouped feature importance. As such, (iii) naturally extends to groups: if $G_1, \ldots, G_\ell$ are groups of features and we let $G = \{G_1, \ldots, G_\ell\}$, then $0 \leq I_G \leq I_{G_1} + \cdots I_{G_\ell}$. This is a direct result of the definition in Eq. (5). Note, another natural extension of group importance would have been $|f_{i_1}(t_{i_1})| + \cdots + |f_{i_k}(t_{i_k})|$ as the summand. However, this would result in group importance simply being the sum of feature importances, as

$$\begin{aligned} \tilde{I}_{\{x_{i_1}, \ldots, x_{i_k}\}} &= \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} (|f_{i_1}(t_{i_1})| + \cdots + |f_{i_k}(t_{i_k})|) \\ &= \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} |f_{i_1}(t_{i_1})| + \cdots + \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} |f_{i_k}(t_{i_k})| \\ &= I_{x_{i_1}} + \cdots + I_{x_{i_k}}. \end{aligned}$$

A benefit of this alternative definition is that it is perfectly additive for non-overlapping groups. In our proposed definition, group importance scores are not additive: summing group importances as we propose them corresponds to summing up absolute values (averaged over training samples), whereas the proposed definition of group importance is defined as the absolute value of a summed contribution (also averaged over training samples).

Similar to other definitions of group importance[8,14–17], our definition is permutation invariant. That is, for any permutation $\sigma \in \mathrm{Sym}(\{1, 2, \ldots, k\})$, the symmetry group, it immediately follows from the definition that $I_{\sigma(G)} = I_G$. Besides permutation invariance, our notion of group importance is also invariant under linear feature scaling: linearly scaling the values of a feature correspondingly stretches the shape function, thus leaving feature and group importance invariant. However, under nonlinear transformations (e.g. the log transform), feature importances may be affected.

In the same vein, having a different number of features between groups does not artificially inflate the larger group's importance, since all shape functions are zero-centered and we adopt the definition in Eq. (5) rather than the (flawed) alternative definition above.

2.3.1 Including higher order interaction terms

Since the interaction terms learned by a GAM are adjusted for the individual effects (assuming purification[23]), group importance can be computed over any set of individual features and interaction terms. This holds true for both second order interactions, and for higher order terms if they were included in the GAM, which is not traditionally the case[2,3]. For any GAM with higher-order interactions,

$$g(\mathbb{E}[y]) = \beta_0 + f_1(x_1) + \cdots + f_n(x_n) + \sum_{i \neq j} f_{i,j}(x_i, x_j) + \sum_{\substack{i,j,k \\ \text{different}}} f_{i,j,k}(x_i, x_j, x_k) + \ldots, \tag{6}$$

we can define any group G of features and interactions to compute the feature importance over. For example, $G = \{(x_1), (x_2), (x_2, x_3), (x_1, x_2, x_3)\}$ with indices $\mathrm{Ind}_G = \{(1), (2), (2,3), (1,2,3)\}$, where each group of features $(x_{i_1}, x_{i_2}, \ldots, x_{i_k})$ corresponds to the shape function $f_{i_1, i_2, \ldots, i_k}(x_{i_1}, x_{i_2}, \ldots, x_{i_k})$ in the GAM. Notably, if $n = 1$, this term reduces to the individual shape function $f_{i_1}(x_{i_1})$. For any group G and its feature indices Ind_G, our notion of group importance naturally extends to

$$I_G := \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \left| \sum_{(i_1, \ldots, i_k) \in \mathrm{Ind}_G} f_{i_1, \ldots, i_k}(t_{i_1}, \ldots, t_{i_k}) \right|. \tag{7}$$

In the example of $G = \{(x_1),(x_2),(x_2,x_3),(x_1,x_2,x_3)\}$ and $\text{Ind}_G = \{(1),(2),(2,3),(1,2,3)\}$, the group importance reduces to

$$I_G = \frac{1}{|\mathscr{T}|}\sum_{t\in\mathscr{T}} |f_1(t_1)+f_2(t_2)+f_{2,3}(t_2,t_3)+f_{1,2,3}(t_1,t_2,t_3)|\,. \tag{8}$$

As a concrete example, consider individual features x_1, x_2, their feature interaction (x_1,x_2), and a target $y = x_1$ XOR x_2. The main effect of x_1, adjusted for all other individual features and feature interactions, contains no information about the XOR target. Consequently, $f_{x_1}(x_1)$ is flat and the importance of x_1 is zero, i.e. $I_{\{x_1\}} = 0$. The same holds true for x_2: after adjusting for other terms, it carries no predictive signal for the XOR target, yielding a flat shape function $f_{x_2}(x_2)$ and $I_{\{x_2\}} = 0$. Since both main effects are identically zero, the group consisting of the two individual features has importance

$$I_{\{x_1,x_2\}} = \frac{1}{|\mathscr{T}|}\sum_{t\in\mathscr{T}} |f_1(t_1)+f_2(t_2)| = \frac{1}{|\mathscr{T}|}\sum_{t\in\mathscr{T}} |0+0| = 0,$$

as the two flat shape functions contribute no signal jointly. In contrast, the feature interaction (x_1,x_2) can perfectly encode the target, because jointly observing x_1 and x_2 fully determines x_1 XOR x_2. Its corresponding shape function is therefore non-flat and yields $I_{\{(x_1,x_2)\}} > 0$. The power set of the features has group importance $I_{\{x_1,x_2,(x_1,x_2)\}} = I_{\{(x_1,x_2)\}} > 0$. Thus, $I_{\{x_1\}} = I_{\{x_2\}} = I_{\{x_1,x_2\}} = 0$, while $I_{\{x_1,x_2,(x_1,x_2)\}} = I_{\{(x_1,x_2)\}} > 0$. These importance scores align with the ANOVA interpretation that XOR is captured solely by the interaction term.

Because the mechanics of including an individual feature in a group importance computation are the same as including a higher-order interaction term, the remainder of the manuscript continues to focus without loss of generality on groups of individual features.

2.4 Explained variation and feature importance

We contrast feature importances with an alternate measure of importance in statistics: explained variation (aka explained variance). Explained variation represents to what extent a set of variables in a model accounts for the variation (often variance) of a target. In regression problems, explained variance is typically defined as R^2, the coefficient of determination[29]. For classification problems, however, it depends on the setting, but is usually also a measure of 'goodness of fit'[30].

In Principal Component Analysis (PCA)[31], the scale for explained variance is the sum of the individual component variances $\text{Var}(x_1)+\cdots+\text{Var}(x_n)$. For independent variables (e.g., principal components) $x_1,\ldots,x_n$, the variance explained by x_i is then defined as $\frac{\text{Var}(x_i)}{\text{Var}(x_1)+\cdots+\text{Var}(x_n)}$[32]. More generally, if the x_i are independent or at least weakly correlated, $\text{Var}(x_1)+\cdots+\text{Var}(x_n)$ can be approximated by the total variance $\text{Var}(x_1+\cdots+x_n)$. While one might loosely view total variance as analogous to the group importance on the full feature set, there exist pathological settings in which there exists a group G that is a strict subset of all features such that $I_G > I_{x_1,\ldots,x_n,\text{all interactions}}$. Even the maximum group importance attainable over all feature subsets is not analogous to total variation, as this would imply comparing a group of features to a potentially different set of features, which is nonsensical. The sum of individual importances $I_{x_1}+\cdots+I_{x_n}$ (plus importance of interactions) is also not appropriate as this scale is generally not achieved unless all feature importance contributions are perfectly additive.

2.4.1 Feature selection using group importances

Once relative group importances have been computed, or a ranking has been obtained more generally, this can also be used for feature selection. That is, by picking the top k groups of features that are assigned most importance by the model. This is particularly useful when feature selection has to be performed at a group-level for predefined groups. This is important because in many practical applications individual features cannot be selected independently.

For example, in the case of a budget-constrained finance company choosing which datasets to acquire from a data vendor, they often get to trial and evaluate each dataset before choosing which to acquire. Once you determine which ones to buy, you can only select the features available in said datasets and not in the others. Thus, features are selected only at a group level, and it is not possible to select the top individual features from each dataset. Other examples where features can only be selected in predefined groups include medical lab values, where each lab test results in multiple lab values, or choosing what questionnaires to give patients during a psychology study.

3 Synthetic Experiments

To demonstrate the behavioral properties of our proposed definition of group importance for GAMs, we generate three synthetic datasets. We examine the behavior exhibited when features contain signal that is (1) additive and perfectly correlated, (2) opposing and independent, and (3) opposing and highly correlated. One real-world healthcare example of opposing (and correlated) signal can be found in age-related diseases: advanced age indicates increased risk, while a sufficient amount of medication lowers risk, therefore containing opposite signal. At the same time, age and the amount of medication are likely correlated.

Generalizing from these three extreme, synthetic cases, we then study group importance as a function of correlation in both the additive and opposing signal case. We also review various edge cases in Supplementary Information S1. Note, the EBMs achieve near-perfect prediction accuracy in each synthetic experiment due to the simplistic setup of the synthetic data. The synthetic datasets consist of samples $\{(x_i, z_i, y_i)\}_{i=1}^{n=10^6}$, where $x_i \overset{\text{iid}}{\sim} \mathscr{U}[0,10]$ is sampled from the uniform distribution; we define z_i and y_i in each subsection. We denote $x = (x_1, \ldots, x_n)^T$ and $[n] := \{1, 2, \ldots, n\}$.

3.1 Additive signals

To create synthetic data consisting of additive signals, let z be a copy of x. Furthermore, let $\varepsilon_i \overset{\text{iid}}{\sim} \mathscr{N}(0,1)$ be noise sampled from the normal distribution in order to construct label $y_i = 1$ if $x_i + \varepsilon_i > 5$ and 0 else. No matter whether $z_i = x_i$ or $z_i = -x_i$ $\forall i \in [n]$, z and x are equally indicative of any target y.

Then, due to perfect correlation, $f_x(\cdot)$ and $f_z(\cdot)$ each learn an identical signal that is half of the total signal in the data (see Figure 1(a)), as described in Section 2.1.1. Thus, the group importance of x and z is equal to the sum of their individual importances, i.e., $I_{\{x,z\}} \approx 2I_x$ (see Figure 1(a)). Similarly, if a feature x is duplicated and negated to $-x$, the corresponding shape functions $f_x(\cdot)$ and $f_{-x}(\cdot)$ contain negated, yet equivalent, signal and therefore also have equal feature importance. This is a direct consequence of the scaling property discussed previously. This example is shown in Supplementary Information S1.1.

3.2 Independent, conflicting signals

To obtain a feature z that is independent of x, sample $z_i \overset{\text{iid}}{\sim} \mathscr{U}[0,10]$. Then, to create opposing signal of z with respect to x, let $y_i = 1$ if $x_i > z_i$ and 0 else. We visualize the shape functions $f_x(t_x)$ and $f_z(t_z)$ and show the individual and group importances in Figure 1(b).

By symmetry in the construction of y_i, an increase in x_i increases the probability that $y_i = 1$ (for random $z_i \sim \mathscr{U}[0,10]$), while an increase in z_i *decreases* that probability (for random $x_i \sim \mathscr{U}[0,10]$). As a result, the shape functions $f_x(\cdot)$ and $f_z(\cdot)$ in the EBM are each other's additive inverses, and their group importance is less than the sum of individual importances.

To analyze the behavior shown in the group and feature importances in a more detailed fashion, recall that the group importance of features x and z is formulated as

$$I_{\{x,z\}} = \frac{1}{|\mathscr{T}|} \sum_{t \in \mathscr{T}} |f_x(t_x) + f_z(t_z)|. \tag{9}$$

In this synthetic experiment, features x and z contain exactly opposite signal, i.e., $f_x(\cdot) \approx -f_z(\cdot)$. Because feature values $t = (t_x, t_z) \in \mathscr{T}$ are independently sampled, there might be many samples t where $f_x(t_x) \gg 0$ and $f_z(t_z) \ll 0$ (and vice versa) such that these terms do not cancel each other out and, as a result, contribute to the group importance. Similarly, there are roughly equally many samples such that $f_x(t_x) \gg 0$ and $f_z(t_z) \gg 0$, as well as $f_x(t_x) \ll 0$ and $f_z(t_z) \ll 0$, assuming large n. Thus, the group importance $I_{x,z}$ is slightly larger than I_x and I_z, but far smaller than their sum $I_x + I_z$.

3.3 Correlated, conflicting signals

We now examine how our proposed definition of group importance behaves for correlated features that contain opposing signal. This is not an uncommon scenario in practice as, in many instances, features tend to be correlated and samples be drawn non-iid.

To have z be correlated with x, we let $z_i = x_i + \delta_i$, where we sample $\delta_i \sim \mathscr{U}[-2,2]$ (and scale z_i to $[0,10]$), so that x and z are highly correlated. To retain conflicting signal, we again define $y_i = 1$ if $x_i > z_i$ and 0 else.

The shape functions remain mostly unchanged with respect to the previous synthetic experiment, and the individual feature importances also remain balanced. However, in Figure 1(c) we see that the group importance $I_{\{x,z\}} < I_x \approx I_z < I_x + I_z$. This is a direct result of the correlation and, in particular, $t_x \approx t_z$ for all samples $(t_x, t_z) \in \mathscr{T}$ and $f_x(t_x) \approx -f_z(t_z)$, meaning the shape functions nearly cancel each other out for most samples. That is, for perfectly opposing signal: the more correlation, the lower the group importance compared to the individual feature importances. In general, if feature x_{i_1} is such that $I_{\{x_{i_1}, x_{i_2}, \ldots, x_{i_k}\}} < I_{x_{i_1}}$, then the features $x_{i_2}, \ldots, x_{i_n}$ must contain opposing signal as a whole compared to x_{i_1}, since they lower the mean absolute contribution to the target.

3.4 Group importance as function of correlation

In the three synthetic examples above, we considered the extreme cases of perfectly additive or conflicting signal with either high or no correlation between feature values. We now generalize these experimental setups by computing group importance as a function of the Pearson correlation between x and z. This is achieved by varying boundary $b \in [0, 10^3]$ such that $\delta_i \sim \mathscr{U}[-b, b]$ yields a nearly uniformly distributed $z_i = x_i + \delta_i$ on one end, or a z perfectly correlated with x on the other end of the spectrum.

(We again scale z_i to $[0, 10]$ for convenience in constructing y.) In other words, we vary the correlation between x and z and study the effects on their group importance, in both the additive and conflicting setup.

3.4.1 Additive signal

To generalize the experiment on additive signals in Section 3.1, we keep the same target y and compute the group importance as a function of the Pearson correlation ρ between x and z (see Figure 1(d)). We also show the individual importances of x and z given the asymmetry in how y is constructed. Observe that the total group importance is constant (up to random noise) since target y is determined using only feature x and not z (see Figure 1(d)). For perfect correlation (i.e., $\rho = 1$ as in Section 3.1), we see that x and z are both assigned half the total, group importance, while y has near-zero importance when it is near-uniformly distributed and uncorrelated with x.

3.4.2 Conflicting signal

Generalizing the conflicting signals experiment of Section 3.2, we visualize the group importance as a function of the correlation in Figure 1(e). We observe a roughly linear decline in the group importance of x and z as their correlation increases, transitioning to a more exponential decrease at higher correlation levels. This pattern likely arises from measuring group importance as the mean *absolute* contribution, which inherently slows the convergence to zero. However, the group importance does converge to zero as the correlation goes to one.

4 Real-World Experiments

We evaluate the computational efficiency of group importance and demonstrate the utility and relevance of group importance in two real-world medical datasets. In our first case study, we identify the depressive symptom *negative valence* based on groups of correlated features drawn from functional brain networks and various life and behavioral domains. As a second case study, we investigate the importance of social determinants of health in the outcome of hip replacements. For each we analyze the group importance ranking, and use this as a guide for applying group importance for low-cost feature selection purposes.

4.1 Computational Efficiency

Unlike other group importance methods, our method is computationally low-cost. It requires only k function evaluations for each sample $t \in \mathscr{T}$, resulting in a complexity of $\mathscr{O}(k|\mathscr{T}|)$, or $\mathscr{O}(|\text{no. features in group}| \cdot |\text{no. samples}|)$ in terms of function evaluations. This is an order of magnitude faster than the popular grouped permutation importance (GPI) method[12] and groupSHAP[16] across different experimental setups, even if the number of features or samples is very high. For various OpenML datasets with such different data regimes[33], there is an order of magnitude difference in terms of runtime (on an Intel(R) Xeon(R) CPU @ 2.20GHz with 12GB RAM) of group importances implemented for GAMs versus GPI and groupSHAP (see Figure 2).

4.2 Neuroscience Experiments

While better understanding the underlying mechanism of depressive symptoms could hugely improve mental health care, this is a difficult task as the manifestation of the symptoms is quite heterogeneous. That is, they are not confined to a single, easily identifiable brain region or network, nor clear changes in behavior[34–36].

4.2.1 Data

We conduct our analysis on publicly available data acquired by the National Consortium on Alcohol and Neurodevelopment in Adolescence (NCANDA)[37] (data release `NCANDA_PUBLIC_6Y_REDCAP_V04`[38] and, for fMRIs, `NCANDA_PUBLIC_6Y_RESTINGSTATE_V01`[39]). The NCANDA dataset contains multimodal and longitudinal data comprised of biographical, self-reported behavioral, and fMRI data of 1396 observations (aka visits) from 521 participants (ages 12 to 18 years).

Predictors There are 23 resting state fMRI (rs-fMRI) features as well as 131 demographic & behavioral measurements that are categorized into eight disjoint groups[8,40]:

1. **Life & Trauma Events** (full: *Life events and childhood trauma*): encompasses trauma, emotional neglect, and (parental) marital separation
2. **Personality Traits**: include agreeableness, emotional stability, extraversion, positive thinking, and cognitive restructuring.
3. **Neuropsychological Battery**: includes problem-solving and reasoning skills, attention, memory, and language[41].
4. **Executive Function Spectrum** (full: *Behavior Rating Inventory of Executive Function* (BRIEF)): assesses executive functions inhibition and shifting, measuring the ability to resist or delay impulses and to tolerate change[42].

5. **Social Support Features**: records relationships with friends and family, as well as active participation in social clubs and events.
6. **Sleep Patterns**: tracks wake-up time and bedtime for weekdays and weekends, as well as sleepiness and circadian preferences.
7. **Demographics** (full: *Demographics and pubertal development*): includes age, sex, BMI, ethnicity, pubertal development scale score.
8. **Alcohol & Drug Use Features**: records alcohol and substance use.

The 23 brain functional measurements were generated by processing each resting-state fMRI (acquired on a GE or Siemens scanner; voxels: 64x64x32, voxel size: 3.75mm x 3.75mm x 5mm; TR: 2.2s) using the NCANDA pipeline outlined in[43]. The pipeline consists of skull-stripping, motion correction, detrending, spatial smoothing, and anomaly detection through a manual quality check. All scans of this analysis had a minimum of 7.5 minutes of usable scan time and were registered to the SRI24 atlas[44]. 23 intrinsic function networks (IFN) were extracted via longitudinal independent component analysis (ICA), and each IFN is a brain network that regulates different aspects of our cognition and behavior, such as memory, attention, emotions, and motor skills. Each of the 23 IFNs is assigned a score representing its strength, i.e. the average functional connectivity within the brain network, sometimes referred to as the functional network's *efficiency score*[45]. To account for scanner differences in the acquisition (GE or Siemens), we regress out scanner type from the efficiency scores by performing scanner-level standardization.

The eight disjoint, non-fMRI groups defined above, and its features, are significantly correlated[8,40,46]. This may affect the individual importance scores, as such features share correlated signal. However, the correlation does not hinder the interpretation of the group importances; in fact, our proposed group importance is able to account for correlation (see Section 3).

Target variable We identify *negative valence*[8,22], an aggregate statistic for feeling fear, anxiety, sadness, loss, and threat, as a measure of depressive symptoms. The negative valence target has been extracted from the National Institute of Mental Health (NIMH) Research Domain Criteria (RDoC) initiative, which is a research framework for approaches to investigating mental disorders[47,48]. It has been shown that negative valence's subdomains are highly correlated with Major Depressive Disorder (MDD) and that MDD reflects dysfunctions in negative valence systems[49]. Negative valence has a prevalence of 6.02% across all visits.

Model (EBM) configuration The EBM was trained using 100 outer bags, 50 inner bags, 25 min samples leaf, 64 max bins, and 500 smoothing rounds (to prevent overfitting).

4.2.2 Group Importance Ranking

To assess the accuracy achieved by the EBM, we conducted stratified 10-fold cross validation. All visits by a subject were consistently in either the training or test set in order to avoid information leakage to the test set. This is necessary since the data is longitudinal in nature. The EBM model attains a mean test AUC of 0.874 ± 0.011, balanced accuracy of 0.735 ± 0.016, and Brier score of 0.061 ± 0.002 across visits, where the errors show the standard error of the mean (SEM). This is on par with models such as random forests, recurrent neural networks, and logistic regression trained on this dataset[8]. The strong predictive performance of the EBM model shows that the learned shape functions are discriminative. The results in Table 1 show the stability of the derived statistics, such as feature importances. As with most machine learning models, an EBM is not intended to recover causal relationships or the true data-generating process; rather, its value lies in providing transparent predictions from the available data and features.

The most important individual feature was emotional stability, a feature belonging to the Personality Traits group. Since there are significant correlations between features and groups in NCANDA[8,40,46], we aim to gain a deeper insight into which *groups of features* were driving prediction. To that end, we combined all fMRI features into one group and divided the rest of the measurements into the eight groups described in Section 4.2.1. This yields the group importances shown in Figure 5, revealing that five of the nine groups are more important than the most important individual feature.

Interestingly, the most important group is Life & Trauma Events while only five of its features are among the 20 most important features. The group Life & Trauma Events contains mostly highly correlated features, namely chronic/non-chronic life aspects that are controllable/uncontrollable and have a positive/negative/ambiguous outcome (each of the $2 \cdot 2 \cdot 3 = 12$ combinations is a feature). Naturally, these features are highly correlated and are thus each attributed relatively low feature importance. Yet, when considered collectively, these features form the most critical group for prediction, highlighting the significance of group importances in our analysis. This phenomenon of high group importance for correlated, low-importance features was also observed in the additive synthetic experiment (Section 3.1) and its generalization (Section 3.4.1).

Table 1 presents the rankings of the NCANDA groups based on their group importance, as determined by several methods. Our method's ranking is compared to those obtained through groupSHAP[16], Grouped Permutation Importance (GPI)[12] and a previously reported ranking on the NCANDA dataset, derived using another permutation-based method[8]. While there is significant overlap in rankings, our method appears to attribute much more importance to neuropsychology battery than the permutations-based ranking[8], ranking it as the third most important group instead of the seventh. GPI ranks neuropsychology as the fifth most important group. Our higher ranking seems more in line with the literature, which extensively links depressive symptoms to measurements of the neuropsychological battery group, including reasoning skills[50,51], memory[52], and intellectual functioning[53,54]. Notably, our ranking is most similar to that of groupSHAP. All four methods agree on the two most important groups, namely Life & Trauma Events and Personality Traits, while Executive Function is also found to be an important group. On the other end of the spectrum, the group Demographics is attributed low group importance across the board, with its individual features also being attributed low individual feature importance. Similarly, our method, groupSHAP, and GPI rank the group Alcohol & Drug Use as the least important group, while it is the 5th most important group according to the permutations-based ranking[8].

One group omitted by Paschali et al. (2022)[8] was Brain Networks. Given that functional interactivity between brain networks tend to be correlated[55], just like in the case of Life & Trauma Events, this decision might have been based on analyzing low feature importances. However, our analysis reveals that the importance of the group of fMRI features is substantial and is larger than all individual features but emotional stability, while also outranking the two groups Demographics and Alcohol & Drug Use.

To evaluate the significance of the rankings, we evaluated ranking separability and stability through statistical tests across 10 folds. This is done for our proposed group importance metric, groupSHAP and GPI; we were unfortunately unable to reproduce the Permutations baseline[8]. Using a Friedman test on the full group rankings, we obtain statistically significant results ($p < 0.001$) for the proposed method, the GPI baseline, and groupSHAP[16], indicating consistent rank separation among feature groups. Focusing on the top three groups, the Friedman test yields $p < 0.05$ for all three methods. The Wilcoxon signed-rank test further indicates that the relative ordering of the top two groups is not statistically distinguishable for any of the methods considered. The exact p-values can be found in Table 1.

4.2.3 Feature selection

Following the analysis of the group importance ranking, our next step is to evaluate the practicality of using this ranking for feature selection. We conduct a series of experiments where we first train the EBM using the top $1,\ldots,9$ groups, assessing model accuracy with test AUC and the Brier score through 5-fold cross-validation. The cumulative importance of these groups as they are sequentially added in order of their importance can be found in Figure 3.

Accuracy for top groups: Based on the ranking of groups as defined in Table 1 by our method, we compute the average test AUC and Brier score when trained on the top $1,\ldots,9$ groups respectively (see Figure 4). Training with just the top three groups achieves a test AUC and Brier score almost comparable to using all nine groups. Figure 4 shows that additional groups of features did not appear to significantly increase predictive power once (at least) the first three groups of features are used for training. However, Figure 5 shows that the remaining six groups are assigned non-trivial group importance each. This implies these groups make non-negligible contributions to the model output through the learned shape functions, even though they are, as a whole, significantly correlated with the groups of Life & Trauma Events, Personality Traits, and Neuropsychological battery. Figure 4 shows there is barely an improvement in the AUC and Brier score after including these three groups.

4.3 Healthcare Experiments

In our healthcare case study, we evaluate the importance of community-level social determinants of health variables in predicting mortality after total hip arthroplasty. It is debated in the literature how important patient demographics, such as race, sex, and age are compared to broader, community-level social determinants of health (SDOH), like healthcare access and socioeconomic status[56–58]. In this section we compare an aggregate measure of SDOH, i.e. 'community', to various individual factors related to demographics as well as the Elixhauser Comorbidity Index.

4.3.1 Data

Our healthcare study utilizes data extracted from the Pennsylvania Health Care Cost Containment Council (PHC4) Database, covering the period from 2012 to 2018. This comprehensive dataset encompasses de-identified patient information, diagnostic and procedural codes, and financial data from 170 non-governmental acute care hospitals in Pennsylvania. Specifically, we focused on 140,092 patients who underwent elective primary Total Hip Arthroplasty (THA) according to validated International Classification of Diseases codes. After applying exclusion criteria such as inflammatory arthritis, non-elective admissions, and demographic omissions, our final cohort consisted of 105,336 patients. Key patient- and facility-level variables were extracted, including demographics and medical comorbidities indexed by the Quan adaptation of the Elixhauser Comorbidity Index. Unfortunately, certain granular, patient-level information such as a patient's income was unavailable, although it has been shown that socioeconomic status is very correlated with income and poses a reasonable proxy[59].

Community-level variables were an integral part of our analysis. These were extracted by geocoding the residential 5-digit zip codes of each patient to their corresponding census tract variables from the American Community Survey (ACS). We focused on variables known to influence THA outcomes, including social support indicators (percent living alone), immigration and acculturation factors (percent foreign-born, percent non-English speakers), digital literacy (percent with computer and internet access), socioeconomic status (median household income, percent uninsured), and education levels. Additionally, we included the National Walkability Index, given its relevance to osteoarthritis and post-arthroplasty outcomes. These variables collectively formed the 'community factors' group in our analysis, allowing us to assess their aggregated impact alongside individual patient factors such as race.

Target variable Our target variable is *90-day mortality*, a critical measure in hospital settings due to its direct relevance to patient outcomes and healthcare quality. Mortality occurred in 0.3% of cases.

Model (EBM) configuration The EBM was trained using default parameters.

4.3.2 Results

Our EBM obtains an average test AUC of 0.76. Similar to the neuroscience case study, we find that it is in fact a group of individually weak features that is driving the prediction of 90-day mortality, not necessarily individual features. We compute the feature group importance of the group of community-based features, and contrast this in Figure 6 with the individual feature importance of race, sex, age, discharge location, and the Elixhauser Comorbidity Index.

Remarkably, the most important feature for predicting mortality in hip arthroplasty is the group of community-based measurements, carrying $\sim$ 8x the importance of Race, a well-recognized risk factor[56, 60, 61]. Moreover, the community-level SDOH features are assigned more importance than sex and age combined, while these are also considered important risk factors for adverse outcomes after undergoing total hip arthroplasty[62–64]. The second most important feature is discharge location, which is an important predictor that could be informative of the amount of care available to a patient. It has been shown in the literature that patients discharged to a nursing or rehabilitation facility are less likely to have severe adverse events (e.g., mortality) compared to those who attempt to self-manage at home[65].

These findings underscore the importance of considering social determinants of health in hip arthroplasty outcomes, and beyond in the healthcare system. Despite the tendency for social determinants to be highly correlated[66, 67], which often leads to low individual feature importance, their collective influence as a group is substantial. This reinforces the value of group importance analysis in revealing critical insights that might be obscured when focusing solely on individual feature importance.

4.3.3 Ethical considerations

The inclusion of sensitive features such as race, trauma-related measures, or community-level socioeconomic indicators warrants careful ethical consideration. While such features can be controversial, excluding them does not eliminate their influence on model predictions, as correlated features may still encode similar information implicitly. In this sense, transparent models such as EBMs allow the contribution of sensitive features to be examined explicitly, rather than obscured within complex interactions and correlations. It is important to reemphasize that an EBM is not intended to recover causal relationships or the true data-generating process, but instead provides transparent predictions from the available data and features. Thus, including sensitive features can help better understand the structure of predictive signals present in the data. When used responsibly, transparent group-level importance analyses can help surface and scrutinize potential sources of bias, rather than inadvertently hiding them.

5 Conclusions

We propose a novel method to measure the importance of groups of features within Generalized Additive Models (GAMs) by computing their mean absolute joint contribution. This approach naturally extends the concept of individual feature importance in GAMs, and remains meaningful when there are many correlated features. Our definition is computationally low-cost, requires no model retraining, and yields similar results as groupSHAP in a real-world case study. We demonstrate its properties by applying a GAM-based Explainable Boosting Machine (EBM) to two real-world datasets (in neuroscience and healthcare) and three synthetic datasets. Our experiments reveal that group importance is particularly important for understanding the total signal in a group of correlated features, which can be overlooked when individual features are attributed low importance due to correlation. As such, group importance offers additional interpretability beyond individual feature importance in a wide range of applications.

Future work might involve applying group importances in real-world settings where feature selection can only be done at a group level. Secondly, the proposed metric of group importance assumes predefined groups, while methodology could be developed to discover these groups automatically, e.g. through hierarchical clustering using a group importance metric.

Data Availability

The data used for the real-world neuroscience experiment is made publicly available by the National Consortium on Alcohol and Neurodevelopment in Adolescence (NCANDA)[37] (data release `NCANDA_PUBLIC_6Y_REDCAP_V04`[38] and, for fMRIs, `NCANDA_PUBLIC_6Y_RESTINGSTATE_V01`[39]). The data for the real-world healthcare experiment is made publicly available by Pennsylvania Health Care Cost Containment Council (PHC4), and as of June 24, 2025, can be accessed at `https://www.phc4.org/request-custom-data/available-data/`.

All data generated and analysed for the synthetic experiments are included in this published article and its Supplementary Information file.

Funding

This work was generously supported by funding from the National Institute of Health (R01-AA005965, U01-AA017347, U24-AA021697), Microsoft's AI for Good Lab, and Intelligible, Inc.

References

1. Doshi-Velez, F. & Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).
2. Lou, Y., Caruana, R., Gehrke, J. & Hooker, G. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 623–631 (ACM, 2013).
3. Hastie, T. J. Generalized additive models. In *Statistical models in S*, 249–307 (Routledge, 2017).
4. Nori, H., Jenkins, S., Koch, P. & Caruana, R. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223* (2019).
5. Perrier, A. Feature importance in random forests. *URl: https://alexisperrier.com/datascience/2015/08/27/feature-importancerandom-forests-gini-accuracy.html* (2015).
6. Louppe, G., Wehenkel, L., Sutera, A. & Geurts, P. Understanding variable importances in forests of randomized trees. *Adv. neural information processing systems* **26** (2013).
7. Breiman, L. Random forests. *Mach. learning* **45**, 5–32 (2001).
8. Paschali, M., Zhao, Q., Adeli, E. & Pohl, K. M. Bridging the gap between deep learning and hypothesis-driven analysis via permutation testing. In *International Workshop on PRedictive Intelligence In MEdicine*, 13–23 (Springer, 2022).
9. Bilodeau, B., Jaques, N., Koh, P. W. & Kim, B. Impossibility theorems for feature attribution. *Proc. Natl. Acad. Sci.* **121**, e2304406120 (2024).
10. Darst, B. F., Malecki, K. C. & Engelman, C. D. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC genetics* **19**, 1–6 (2018).
11. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable ai for trees. *Nat. machine intelligence* **2**, 56–67 (2020).
12. Plagwitz, L., Brenner, A., Fujarski, M. & Varghese, J. Supporting ai-explainability by analyzing feature subsets in a machine learning model. In *Challenges of Trustable AI and Added-Value on Health*, 109–113 (IOS Press, 2022).
13. Au, Q., Herbinger, J., Stachl, C., Bischl, B. & Casalicchio, G. Grouped feature importance and combined features effect plot. *Data Min. Knowl. Discov.* **36**, 1401–1450 (2022).
14. Yuan, M. & Lin, Y. Model selection and estimation in regression with grouped variables. *J. Royal Stat. Soc. Ser. B: Stat. Methodol.* **68**, 49–67 (2006).
15. Zhang, Y., Kwon, D. & Pohl, K. M. Computing group cardinality constraint solutions for logistic regression problems. *Med. image analysis* **35**, 58–69 (2017).
16. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. neural information processing systems* **30** (2017).
17. Gregorutti, B., Michel, B. & Saint-Pierre, P. Grouped variable importance with random forests and application to multiple functional data analysis. *Comput. Stat. & Data Analysis* **90**, 15–35 (2015).
18. Liu, G. & Sun, B. Concrete compressive strength prediction using an explainable boosting machine model. *Case Stud. Constr. Mater.* e01845 (2023).

19. Charlton, C. E., Poon, M. T., Brennan, P. & Fleuriot, J. D. Comparing the interpretability of machine learning classifiers for brain tumour survival prediction. *Available at SSRN 4164349* (2022).

20. Sarica, A., Quattrone, A. & Quattrone, A. Introducing the rank-biased overlap as similarity measure for feature importance in explainable machine learning: A case study on parkinson's disease. In *Brain Informatics: 15th International Conference, BI 2022, Padua, Italy, July 15–17, 2022, Proceedings*, 129–139 (Springer, 2022).

21. Lengerich, B. J., Caruana, R., Nunnally, M. E. & Kellis, M. Death by round numbers: Glass-box machine learning uncovers biases in medical practice. *medRxiv* 2022–04 (2022).

22. Woody, M. L. & Gibb, B. E. Integrating nimh research domain criteria (rdoc) into depression research. *Curr. opinion psychology* **4**, 6–12 (2015).

23. Lengerich, B., Tan, S., Chang, C.-H., Hooker, G. & Caruana, R. Purifying interaction effects with the functional anova: An efficient algorithm for recovering identifiable additive models. In *International Conference on Artificial Intelligence and Statistics*, 2402–2412 (PMLR, 2020).

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

25. Liyanage, I. & Thayasivam, U. Pushing the boundaries of interpretability: Incremental enhancements to the explainable boosting machine. *arXiv preprint arXiv:2512.00528* (2025).

26. Bosschieter, T. M. *et al.* Interpretable predictive models to understand risk factors for maternal and fetal outcomes. *J. Healthc. Informatics Res.* **8**, 65–87 (2024).

27. Hegselmann, S. *et al.* Development and validation of an interpretable 3 day intensive care unit readmission prediction model using explainable boosting machines. *Front. Medicine* **9**, 960296 (2022).

28. Ibrahim, S., Afriat, G. I., Behdin, K. & Mazumder, R. Grand-slamin'interpretable additive modeling with structural constraints. In *Thirty-seventh Conference on Neural Information Processing Systems* (2023).

29. O'Grady, K. E. Measures of explained variance: Cautions and limitations. *Psychol. Bull.* **92**, 766 (1982).

30. Ballabio, D. & Consonni, V. Classification tools in chemistry. part 1: linear models. pls-da. *Anal. methods* **5**, 3790–3798 (2013).

31. Wold, S., Esbensen, K. & Geladi, P. Principal component analysis. *Chemom. intelligent laboratory systems* **2**, 37–52 (1987).

32. Shen, H. & Huang, J. Z. Sparse principal component analysis via regularized low rank matrix approximation. *J. multivariate analysis* **99**, 1015–1034 (2008).

33. Vanschoren, J., van Rijn, J. N., Bischl, B. & Torgo, L. Openml: networked science in machine learning. *SIGKDD Explor.* **15**, 49–60, DOI: 10.1145/2641190.2641198 (2013).

34. Iordan, A. & Dolcos, F. Brain activity and network interactions linked to valence-related differences in the impact of emotional distraction. *Cereb. cortex* **27**, 731–749 (2017).

35. Lopresti, A. L., Hood, S. D. & Drummond, P. D. A review of lifestyle factors that contribute to important pathways associated with major depression: diet, sleep and exercise. *J. affective disorders* **148**, 12–27 (2013).

36. Binnewies, J. *et al.* Associations between depression, lifestyle and brain structure: A longitudinal mri study. *NeuroImage* **231**, 117834 (2021).

37. Brown, S. A. *et al.* The national consortium on alcohol and neurodevelopment in adolescence (ncanda): a multisite study of adolescent development and substance use (2015).

38. Pohl, K. *et al.* The 'ncanda_public_6y_redcap_v04' data release of the national consortium on alcohol and neurodevelopment in adolescence (ncanda). *Sage Bionetw. Synap.* (2022).

39. Pohl, K. *et al.* The 'ncanda_public_6y_restingstate_v01' data release of the national consortium on alcohol and neurodevelopment in adolescence (ncanda). *Sage Bionetworks Synap.* (2022).

40. Brown, S. A. *et al.* The national consortium on alcohol and neurodevelopment in adolescence (ncanda): a multisite study of adolescent development and substance use. *J. studies on alcohol drugs* **76**, 895–908 (2015).

41. Zgaljardic, D. J. & Temple, R. O. Neuropsychological assessment battery (nab): Performance in a sample of patients with moderate-to-severe traumatic brain injury. *Appl. Neuropsychol.* **17**, 283–288 (2010).

42. Gioia, G. A., Isquith, P. K., Guy, S. C. & Kenworthy, L. Test review behavior rating inventory of executive function. *Child Neuropsychol.* **6**, 235–238 (2000).

43. Zhao, Q. *et al.* Longitudinally consistent estimates of intrinsic functional networks. *Hum. brain mapping* **40**, 2511–2528 (2019).

44. Rohlfing, T., Zahr, N. M., Sullivan, E. V. & Pfefferbaum, A. The sri24 multichannel atlas of normal adult human brain structure. *Hum. brain mapping* **31**, 798–819 (2010).

45. Müller-Oehring, E. M. *et al.* Influences of age, sex, and moderate alcohol drinking on the intrinsic functional architecture of adolescent brains. *Cereb. Cortex* **28**, 1049–1063 (2018).

46. Sullivan, E. V. *et al.* Cognitive, emotion control, and motor performance of adolescents in the ncanda study: Contributions from alcohol consumption, age, sex, ethnicity, and family history of addiction. *Neuropsychology* **30**, 449 (2016).

47. Insel, T. *et al.* Research domain criteria (rdoc): toward a new classification framework for research on mental disorders (2010).

48. Infurna, M. R. *et al.* Associations between depression and specific childhood experiences of abuse and neglect: A meta-analysis. *J. affective disorders* **190**, 47–55 (2016).

49. Medeiros, G. C. *et al.* Positive and negative valence systems in major depression have distinct clinical features, response to antidepressants, and relationships with immunomarkers. *Depress. anxiety* **37**, 771–783 (2020).

50. Engle, R. W., Sedek, G., Von Hecker, U. & McIntosh, D. N. *Cognitive limitations in aging and psychopathology* (Cambridge University Press, 2005).

51. Hu, C., Wang, C., Liu, W. & Wang, D. Depression and reasoning ability in adolescents: Examining the moderating role of growth mindset. *Front. Psychol.* **13**, 636368 (2022).

52. Burt, D. B., Zembar, M. J. & Niederehe, G. Depression and memory impairment: a meta-analysis of the association, its pattern, and specificity. *Psychol. bulletin* **117**, 285 (1995).

53. Brown, R., Scott, L., Bench, C. & Dolan, R. J. Cognitive function in depression: its relationship to the presence and severity of intellectual decline. *Psychol. medicine* **24**, 829–847 (1994).

54. Clark, D. C. *et al.* Intellectual functioning and abstraction ability in major affective disorders. *Compr. psychiatry* **26**, 313–325 (1985).

55. Marrelec, G. *et al.* Partial correlation for functional brain interactivity investigation in functional mri. *Neuroimage* **32**, 228–237 (2006).

56. Jha, A. K., Fisher, E. S., Li, Z., Orav, E. J. & Epstein, A. M. Racial trends in the use of major procedures among the elderly. *New Engl. J. Medicine* **353**, 683–691 (2005).

57. Lawrence, R. C. *et al.* Estimates of the prevalence of arthritis and selected musculoskeletal disorders in the united states. *Arthritis & Rheum. Off. J. Am. Coll. Rheumatol.* **41**, 778–799 (1998).

58. Mehta, B. *et al.* Race, discharge disposition, and readmissions after elective hip replacement: analysis of a large regional dataset. *Heal. Equity* **3**, 628–636 (2019).

59. Duncan, G. J. & Magnuson, K. Socioeconomic status and cognitive functioning: moving from correlation to causation. *Wiley Interdiscip. Rev. Cogn. Sci.* **3**, 377–386 (2012).

60. Okike, K. *et al.* Association of race and ethnicity with total hip arthroplasty outcomes in a universally insured population. *JBJS* **101**, 1160–1167 (2019).

61. Singh, J. A., Lu, X., Rosenthal, G. E., Ibrahim, S. & Cram, P. Racial disparities in knee and hip total joint arthroplasty: an 18-year analysis of national medicare data. *Annals rheumatic diseases* **73**, 2107–2115 (2014).

62. Belmont Jr, P. J. *et al.* Morbidity and mortality in the thirty-day period following total hip arthroplasty: risk factors and incidence. *The J. arthroplasty* **29**, 2025–2030 (2014).

63. Dowsey, M. M., Spelman, T., Choong, P. F. *et al.* What is the impact of advancing age on the outcomes of total hip arthroplasty? *The J. Arthroplast.* **33**, 1101–1107 (2018).

64. Inacio, M. C. *et al.* Sex and risk of hip implant failure: assessing total hip arthroplasty outcomes in the united states. *JAMA internal medicine* **173**, 435–441 (2013).

65. Shah, C. K. *et al.* Nonelective primary total hip arthroplasty: the effect of discharge destination on postdischarge outcomes. *The J. arthroplasty* **32**, 2363–2369 (2017).

66. Zimmerman, F. J. & Anderson, N. W. Trends in health equity in the united states by race/ethnicity, sex, and income, 1993-2017. *JAMA network open* **2**, e196386–e196386 (2019).

67. Braveman, P., Egerter, S. & Williams, D. R. The social determinants of health: coming of age. *Annu. review public health* **32**, 381–398 (2011).

Author contributions statement

T.B. led the analysis, conducted neuroscience and synthetic experiments, and wrote initial draft of manuscript. L.S. and J.W. led the code development and integration into the InterpretML package. Y.W. and B.M. provided clinical healthcare expertise, and conducted the experiments on social determinants of health. J.D. prepared and analyzed the neuroscience data. O.K. and F.B. provided clinical neuroscience expertise. Q.Z. developed the fMRI segmentation methodology and led its application to the neuroscience experiments. R.C. conceived of the presented idea and developed the theory. K.M. collected and analyzed the neuroscience data. Both R.C. and K.M. provided guidance, were closely involved in writing the manuscript, and led the direction of the project. All authors contributed to discussion and reviewing the manuscript.

Additional information

Competing interests The authors declare no competing interests.

Figures

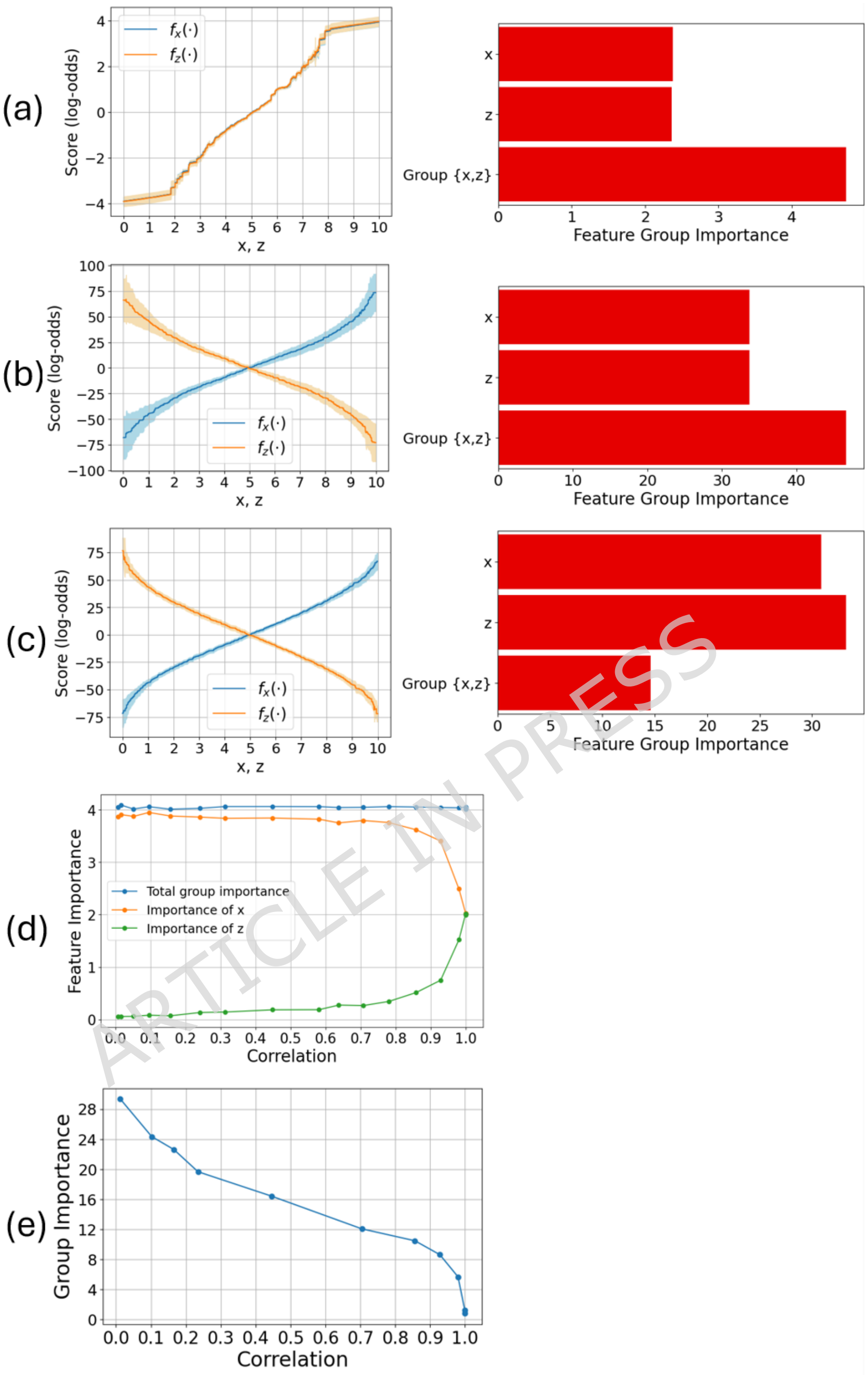


Figure 1. Group importance of the Explainable Boosting Machine (EBM) on various synthetic datasets with two variables x and z, and target y. By design, these variables can either be additive or conflicting in predicting target y; that is, the shape functions of x and z could have similar or opposing shapes. Part (a) shows additive signal through z=x. Part (b) shows independent, conflicting signals, while part (c) shows correlated, conflicting signals. While the group importance ranking in (b) is larger than each single feature importance, the group importance in (c) shows that the two signals almost always cancel each other since the features x and z are highly correlated and close in value. Part (d) demonstrates group importance as a function of correlation for additive signals, while part (e) visualizes group importance as a function of correlation for conflicting signals. Note that the case of zero correlation in part (d) corresponds to the experiment in part (b). High correlation in part (e) corresponds to the experiment in part (c), and the extreme case of perfect correlation in part (e) is explored in Supplementary Information Fig. S3.

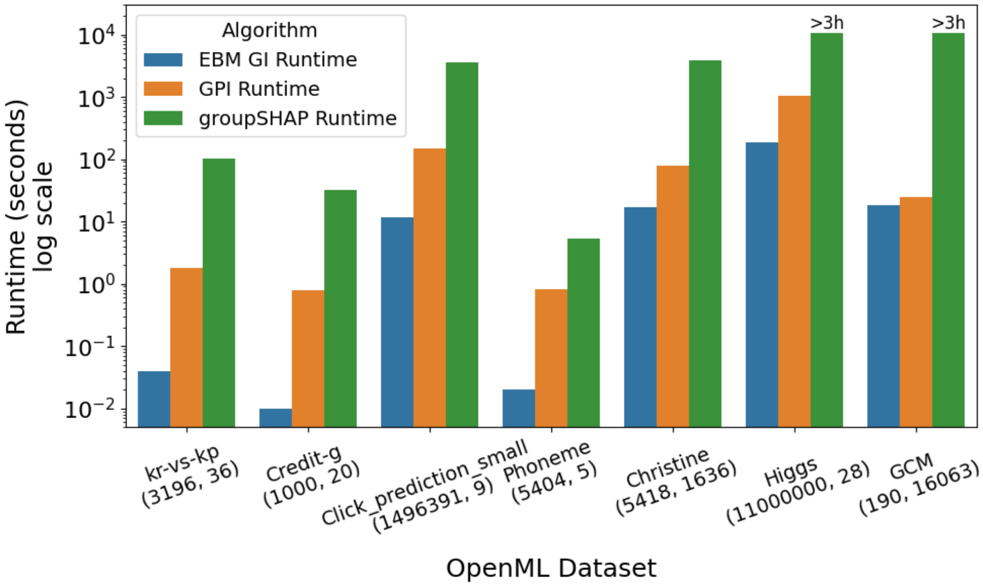


Figure 2. Computational cost in terms of runtime of various metrics of group importance. Our method is compared to Grouped Permutation Importance (GPI)[12] and groupSHAP[16] for a variety of datasets picked to represent different characteristics, such as number of data samples and variables. These seven datasets were downloaded from OpenML, an open-source platform for sharing datasets that is commonly used for benchmarking machine learning models[33]. We were unable to reproduce the runtimes for the permutations-based ranking[8], and the runtime of groupSHAP for Higgs and GCM was cutoff at $>$ 3h. Our method is an order of magnitude faster to compute on all OpenML datasets but GCM, where it is very marginally better. The experiments were performed on an Intel(R) Xeon(R) CPU @ 2.20GHz with 12GB RAM.

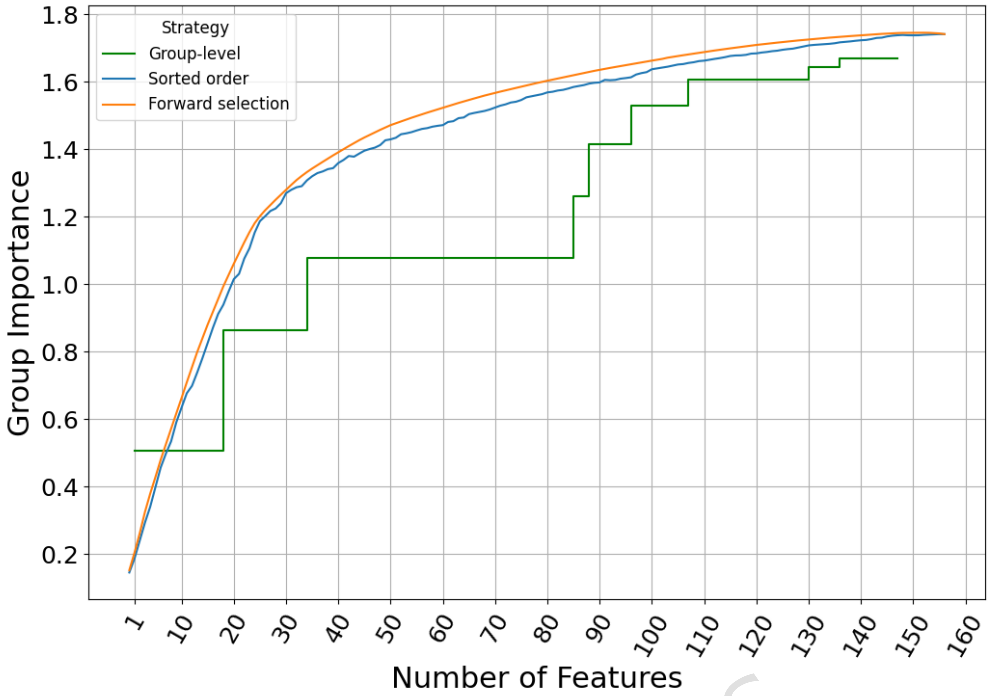
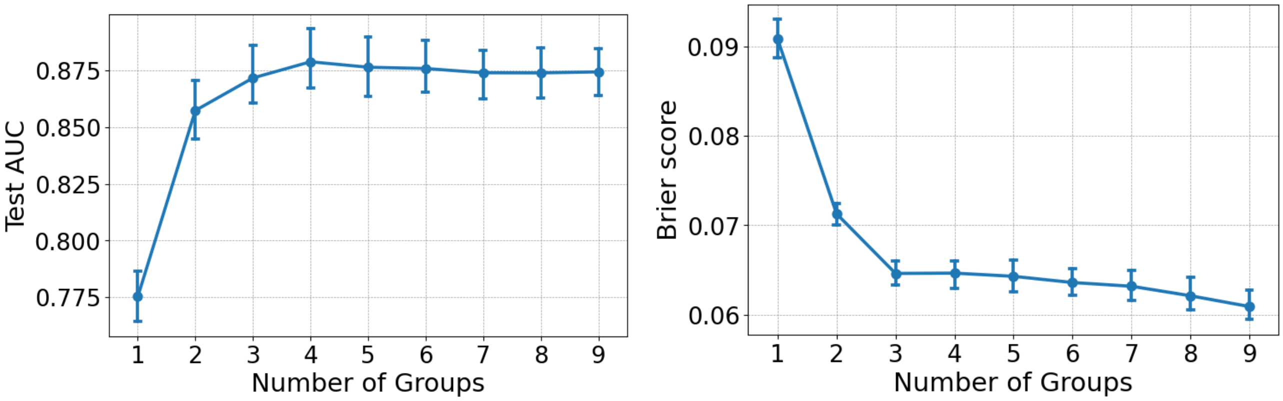


Figure 3. Cumulative group importance when successively adding feature groups. We compare the proposed group-level ranking to both the sorted order of individual feature importances and a greedy forward selection strategy. As expected, group-level selection exhibits a stepwise pattern and is less efficient on a per-feature basis than greedy forward selection, which operates at the level of individual features. The purpose of this comparison is to illustrate the behavior of group-level selection when features can only be selected in predefined groups, showing that selecting meaningful feature groups can provide a practical approximation to forward selection under such constraints.



(a) Test AUC (higher is better).

(b) Brier loss (lower is better).

Figure 4. Mean area under the ROC curve (AUC) and Brier loss, which measures model calibration, when the EBM is trained on the top 1 through 9 groups of features ranked by group importance. Using just the top 3 groups yields nearly optimal performance, indicating feature selection may be performed effectively on a group level.

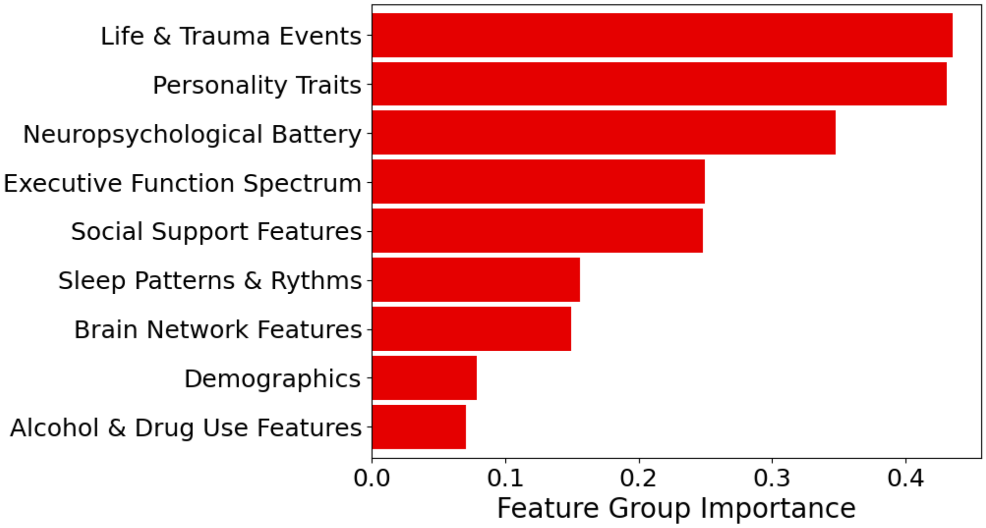


Figure 5. Group importance scores computed using our method for identifying depressive symptoms using data from the National Consortium on Alcohol and Neurodevelopment in Adolescence (NCANDA)[37,38]. The groups Life & Trauma Events and Personality Traits are the most important according to model attribution, highlighting the significance of behavioral and environmental factors.

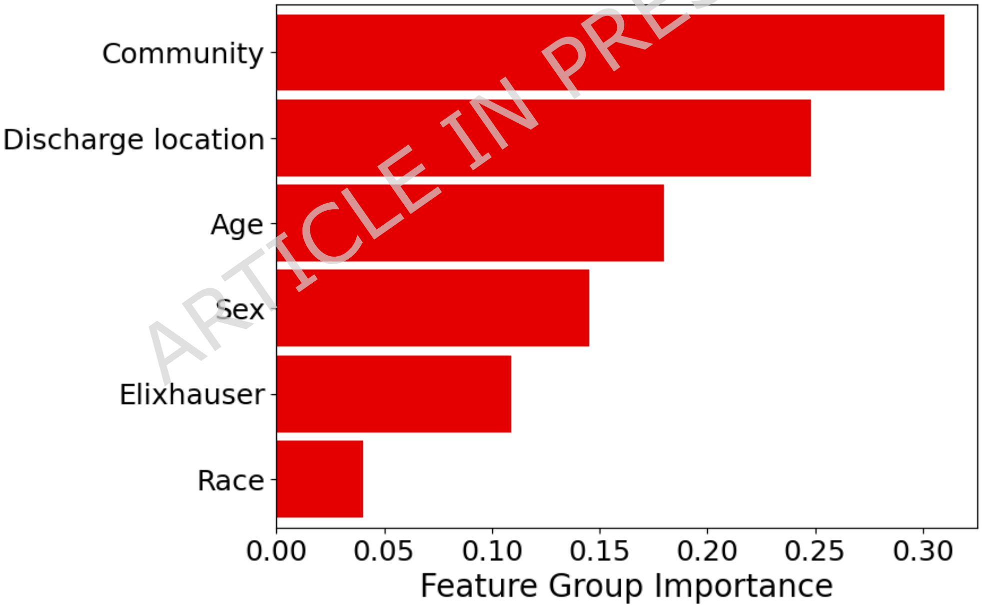


Figure 6. Group and individual feature importances for predicting 90-day mortality after total hip arthroplasty on the Pennsylvania Health Care Cost Containment Council (PHC4) database. Community-level social determinants of health collectively are attributed larger importance by the model than individual predictors such as age, sex, and comorbidities, traditionally considered the main risk factors[56,60–64]. Note that feature and group importances reflect the importance of these terms in the learned model, not necessarily the causal importance of these factors in the true generating function.

Tables

Table 1. Ranking of the group importances of the groups of features in the neuroscience case study using NCANDA data[37,38]. The 23 resting state fMRI (rs-fMRI) features and 131 demographic and behavioral measurements were grouped into nine disjoint groups[8,40]. The permutations-based ranking has been previously reported[8] on NCANDA data excluding rs-fMRI brain network scores. The group permutation importance (GPI)[12] and groupSHAP[16] metrics were also used for comparison. While there is a lot of similarity in the various rankings, our ranking is consistent with the literature[50–54], and is most similar to the ranking obtained by groupSHAP, with the top three most important features being identical. All methods agree on the two most important groups, i.e., Life & Trauma Events and Personality Traits, while Demographics and Alcohol & Drug Use are the least important across all metrics. To evaluate the statistical significance of the ordering of features, a Friedman test was applied on the full group rankings as well as on the top three features per method, showing statistical significance. On the other hand, the Wilcoxon signed-rank test further indicates that the relative ordering of the top two groups is not statistically distinguishable for any of the methods considered.

Rank	Ours	groupSHAP[16]	GPI[12]	Permutations[8]
1	Life & Trauma	Life & Trauma	Personality	Personality
2	Personality	Personality	Life & Trauma	Life & Trauma
3	Neuropsychology	Neuropsychology	Executive Function	Executive Function
4	Executive Function	Brain Networks	Social Support	Sleep Patterns
5	Social Support	Executive Function	Neuropsychology	Alcohol & Drug Use
6	Sleep Patterns	Sleep Patterns	Sleep Patterns	Social Support
7	Brain Networks	Social Support	Brain Networks	Neuropsychology
8	Demographics	Demographics	Demographics	Demographics
9	Alcohol & Drug Use	Alcohol & Drug Use	Alcohol & Drug Use	
p-value top 2	0.641	0.275	0.439	
p-value top 3	0.031	0.023	0.045	
p-value full ranking	$9.5 \cdot 10^{-13}$	$1.7 \cdot 10^{-13}$	$8.2 \cdot 10^{-11}$	