Chaos, Solitons and Fractals 204 (2026) 117724

Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

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

Information-theoretic quantification of high-order feature effects in classification problems

Ivan Lazic [a], Chiara Barà [b], Marta Iovino [b], Sebastiano Stramaglia [c,d], Karolina Kasas-Lazetic [a], Nikša Jakovljević [a], Luca Faes [a,b,*]

[a] Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovica 6, Novi Sad, 21102, Serbia
[b] Department of Engineering, University of Palermo, Viale delle Scienze, Palermo, 90128, Italy
[c] Università degli Studi di Bari Aldo Moro, Dipartimento Interateneo di Fisica M. Merlin, Via Amndola 173, Bari, 70125, Italy
[d] Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Via E. Orabona 4, Bari, 70125, Italy

ARTICLE INFO

Keywords:
Feature importance
Conditional Mutual Information
Nearest neighbors entropy estimation
Redundancy and synergy

ABSTRACT

Understanding the contribution of individual features in predictive models remains a central goal in interpretable machine learning, and while many model-agnostic methods exist to estimate feature importance, they often fall short in capturing high-order interactions and disentangling overlapping contributions. In this work, we present an information-theoretic extension of the High-order interactions for Feature importance (Hi-Fi) method, leveraging Conditional Mutual Information (CMI) estimated via a k-Nearest Neighbor (kNN) approach working on mixed discrete and continuous random variables. Our framework decomposes feature contributions into unique, synergistic, and redundant components, offering a richer, model-free understanding of their predictive roles. We validate the method using synthetic datasets with known Gaussian structures, where ground truth interaction patterns are numerically derived, and further test it on non-Gaussian and real-world gene expression data from TCGA-BRCA. Results indicate that the proposed estimator accurately recovers theoretical and expected findings, providing a potential use case for developing feature selection algorithms or model development based on the interaction analysis.

1. Introduction

One of the central aspects of model explainability in machine learning is the identification and interpretation of feature importance [1]. Understanding which features drive the predictions, on both local and global levels, not only enhances the user's trust in the automated decision process but also provides valuable insights into the underlying data structure and feature interactions within the prediction process. As a consequence, numerous methods have been proposed for assessing feature contributions in a model-agnostic setting, ranging from simple error-based permutation importance tests [2] and basic information-theoretic approaches [3], to more sophisticated techniques such as SHAP, which relies on the weighted average contribution of features across possible subsets [4].

Despite their success, these methods often fail to isolate true marginal contributions of individual features, particularly in the presence of high-order interactions, cases where the importance of a feature depends not only on its contribution but also on its combined influence with other features. Such interactions can make it difficult to determine whether a feature provides unique information alone or in conjunction with specific feature subsets, leading to synergistic or redundant effects. To address this, König et al. [5]

* Corresponding author at: Department of Engineering, University of Palermo, Viale delle Scienze, Palermo, 90128, Italy.
E-mail address: luca.faes@unipa.it (L. Faes).

https://doi.org/10.1016/j.chaos.2025.117724
Received 5 July 2025; Received in revised form 2 November 2025; Accepted 1 December 2025

propose a model-agnostic approach to disentangle standalone, interaction, and dependency effects by decomposing the explained variance of a pretrained model's prediction. While capable of disentangling interaction effects, the approach assumes a mapping from inputs to continuous outputs, limiting its applicability to regression settings and restricting its use in models with discrete or categorical targets. From an information-theoretic standpoint, such interactions can alternatively be characterized using Partial Information Decomposition (PID) [6], which provides partial information atoms that exhaustively decompose the total information contained within the system, which in turn can form the unique, synergistic, and redundant information components. Although comprehensive, this decomposition is computationally demanding, as the number of atom calculations scales according to the Dedekind series. Levaraging these concepts, Ontivero-Ortega et al. [7] extend the Leave-One-Covariate-Out (LOCO) approach [8] to resolve the prediction error reduction in a regression problem into unique, synergistic, and redundant contributions of the features. The crucial novelty of this approach is that it allows to quantify the High-order effects in Feature importance (Hi-Fi method), capturing both the marginal effect of excluding a feature and its interactions with subsets of other features in contributing to the model's output.

In the context of quantifying higher-order interactions, the proposed approach introduces several important advances over existing methods. Unlike the work in [5], which assumes a model, the proposed method maintains model-free properties without compromising scalability, as in the PID framework. Furthermore, in contrast to the Hi-Fi version of LOCO, which is primarily applied to continuous data analysis, our method extends its applicability to classification problems (specifically, discrete target, continuous input). In this work, we build upon this foundation to bring the Hi-Fi method into the context of information theory, which enables the decomposition of feature contributions into interpretable informational components, providing a rich, model-free characterization of feature importance beyond marginal effects. As noted in [7], based on the work by Stramaglia et al. on the decomposition of Granger causal effects [9], the change in predictive performance induced by omitting a feature can be interpreted as a proxy for the Conditional Mutual Information (CMI) between the target and the source variable, conditioned on the remaining features. Furthermore, to make the approach computationally reliable in practical classification problems, we design a data-efficient estimator of the CMI, based on the k-Nearest Neighbor (kNN) method, computed between a discrete output class variable and several continuous input feature variables. We validate the approach utilizing simulated data with known ground truth values, as well as simulated and real-world data without ground truth, but with expected behavior derived from controlled interaction setups or supported by existing literature.

2. Method

Let us consider a system with $n+2$ variables, out of which one represents the class variable Y, while the rest represent continuous features. For the analysis, we isolate the variable X, denoted as the source feature, for which we assess the importance in relation to the other features collected in $\boldsymbol{Z} = \{Z_1, \ldots, Z_n\}$. Our goal is to determine the importance of X to explain Y in the context of the variables $\boldsymbol{Z}$. To this end, we introduce an information-theoretic equivalent of the LOCO approach [8] exploiting the CMI between the class and the source feature conditioned to the remaining features, i.e.,

$$I(Y; X|\boldsymbol{Z}) = I(Y; X, \boldsymbol{Z}) - I(Y; \boldsymbol{Z}). \tag{1}$$

Intuitively, the CMI (1) captures the information provided to the target Y by the source X above and beyond the information brought by the other sources in $\boldsymbol{Z}$. This measure was investigated in [10] in terms of PID, showing that it collects the unique and synergistic effects obtained from the source feature, providing a well-principled criterion for feature selection. Here, using the searching approach devised in [7], we take another perspective on higher-order effects by looking for sets of variables in $\boldsymbol{Z}$ that minimize or maximize the CMI between the class Y and the source X, i.e., $I(Y; X|\boldsymbol{Z}_{min})$ and $I(Y; X|\boldsymbol{Z}_{max})$. In this way we disentangle the unique contribution of X in determining Y from that deriving from its redundant and synergistic interaction with the variables in $\boldsymbol{Z}$ without the need to perform PID. Specifically, the maximum information shared between the class Y and the source X can be decomposed as follows [9]:

$$I(Y; X|\boldsymbol{Z}_{max}) = S(Y; X|\boldsymbol{Z}) + R(Y; X|\boldsymbol{Z}) + U(Y; X|\boldsymbol{Z}), \tag{2}$$

with

$$\begin{aligned} S(Y; X|\boldsymbol{Z}) &= I(Y; X|\boldsymbol{Z}_{max}) - I(Y; X), \\ R(Y; X|\boldsymbol{Z}) &= I(Y; X) - I(Y; X|\boldsymbol{Z}_{min}), \\ U(Y; X|\boldsymbol{Z}) &= I(Y; X|\boldsymbol{Z}_{min}), \end{aligned} \tag{3}$$

representing the synergistic, redundant, and unique contribution, respectively.

2.1. Searching algorithm

Due to the exponential computational cost of an exhaustive search, which requires evaluating all possible conditioning subsets (scaling as $\mathcal{O}(2^n)$), a greedy approach is adopted to obtain $\boldsymbol{Z}_{min}$ and $\boldsymbol{Z}_{max}$ in practice. Specifically, at each iteration of the sequential procedure, variables are added to the two subsets one at a time, minimizing or maximizing the CMI quantities with respect to the set of previously selected variables, reducing the computational complexity to approximately $\mathcal{O}(n^2)$ evaluations. In practice, however, for high-dimensional systems, the greedy search stops after $m \ll n$ iterations, further reducing the complexity to $\mathcal{O}(mn)$. The overall flow of the greedy variable selection is illustrated in Fig. 1, which summarizes the minimization procedures used to determine $\boldsymbol{Z}_{min}$,

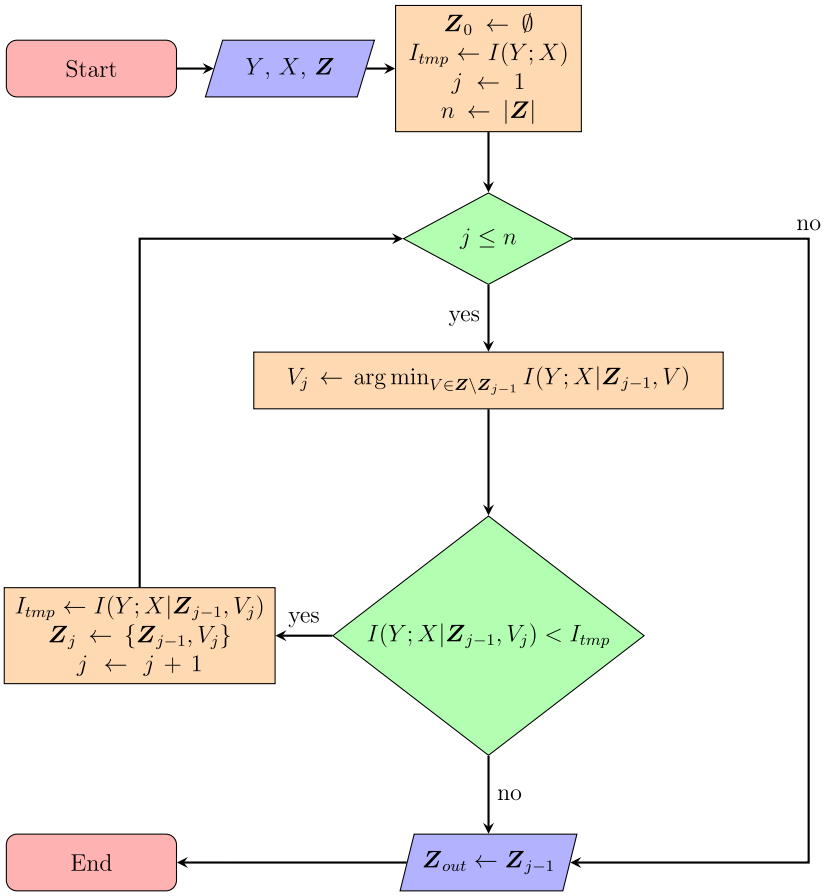

Fig. 1. Flowchart of the theoretical greedy selection procedure for determining $\boldsymbol{Z}_{min}$. At each iteration, a candidate variable V_j is selected, provided it has a minimal CMI. n represents the number of features in $\boldsymbol{Z}$. Selection of the $\boldsymbol{Z}_{max}$ set is similar, where V_j is selected as the feature that maximizes the CMI out of the set of candidate features, $V_j \leftarrow \arg\max_{V \in \boldsymbol{Z} \setminus \boldsymbol{Z}_{j-1}} I(Y;X|\boldsymbol{Z}_{j-1}, V)$, along with the comparison $I(Y;X|\boldsymbol{Z}_{j-1}, V_j) > I_{tmp}$ before the update step.

while the selection of $\boldsymbol{Z}_{max}$ follows a similar approach. The two searches follow the same iterative structure, differing only in the selection criterion applied to the candidate feature V_j (minimizing or maximizing the CMI). The procedure is initialized representing the subsets of selected variables as the empty set, $\boldsymbol{Z}_0 = \emptyset$. Then, to determine the set $\boldsymbol{Z}_{min}$, at any given step j ($1 \leq j \leq n$) the variable $V_j \in \boldsymbol{Z} \setminus \boldsymbol{Z}_{j-1}$ is determined such that

$$V_j = \underset{V \in \boldsymbol{Z} \setminus \boldsymbol{Z}_{j-1}}{\arg\min}\ I(Y;X|\boldsymbol{Z}_{j-1}, V). \tag{4}$$

If the CMI $I(Y;X|\boldsymbol{Z}_{j-1}, V_j)$ decreases significantly below $I(Y;X|\boldsymbol{Z}_{j-1})$, the variable V_j is retained, the set of selected variables is updated as $\boldsymbol{Z}_j = \{\boldsymbol{Z}_{j-1}, V_j\}$ and the counter j is updated; otherwise, if the decrease in CMI is not significant, V_j is discarded and the search terminates with $\boldsymbol{Z}_{min} = \boldsymbol{Z}_{j-1}$. Similarly, to determine the set $\boldsymbol{Z}_{max}$ which maximizes the CMI starting again from the empty set $\boldsymbol{Z}_0 = \emptyset$, at each step $j \geq 1$ the variable $V_j \in \boldsymbol{Z} \setminus \boldsymbol{Z}_{j-1}$ is determined such that

$$V_j = \underset{V \in \boldsymbol{Z} \setminus \boldsymbol{Z}_{j-1}}{\arg\max}\ I(Y;X|\boldsymbol{Z}_{j-1}, V). \tag{5}$$

V_j is retained updating the set of selected features as $\boldsymbol{Z}_j = \{\boldsymbol{Z}_{j-1}, V_j\}$ if the increase of the CMI $I(Y;X|\boldsymbol{Z}_{j-1}, V_j)$ above $I(Y;X|\boldsymbol{Z}_{j-1})$ is significant, or it is otherwise discarded terminating the search with $\boldsymbol{Z}_{max} = \boldsymbol{Z}_{j-1}$. After the two searches, the unique, redundant, and synergistic contributions can be determined applying (3) with the selected $\boldsymbol{Z}_{min}$ and $\boldsymbol{Z}_{max}$.

2.2. Estimation approach

To perform the searching algorithm and find $\boldsymbol{Z}_{min}$ and $\boldsymbol{Z}_{max}$, at the generic step j we need to compute the two CMI terms to be compared for quantifying the variation in the predictive improvement brought by the candidate variable V, i.e.,

$$I(Y;X|\boldsymbol{Z}_{j-1},V) = I(Y;X,\boldsymbol{Z}_{j-1},V) - I(Y;\boldsymbol{Z}_{j-1},V),$$
$$I(Y;X|\boldsymbol{Z}_{j-1}) = I(Y;X,\boldsymbol{Z}_{j-1}) - I(Y;\boldsymbol{Z}_{j-1}). \tag{6}$$

In the following, we first describe the estimator developed in this work to compute the two CMI values in (6) from a dataset of N observations of the class and feature variables, and then describe the procedure followed to assess the statistical significance of the CMI variation obtained adding the selected V_j to $\boldsymbol{Z}_{j-1}$ inside the conditioning set.

The two CMI values are estimated using a mixed variable approach based on the kNN estimator of the mutual information (MI) [11]. Specifically, the mixed approach considers the class variable Y as discrete, with alphabet $\mathcal{A}_Y$, and the feature variables X and $\boldsymbol{Z}$ as continuous, with domains D_X and $D_{\boldsymbol{Z}}$, respectively. Then, a mixed kNN estimation approach is followed [12] to compute the four MI terms in (6), adapting the kNN MI estimator [11] to consider the estimation bias possibly arising from the comparison across different neighborhood spaces [13]. In order to minimize such bias, a neighbor search is performed first in the highest-dimensional space $\{X,\boldsymbol{Z}_{j-1},V\}$, after which the determined distances are projected to the lower-dimensional spaces $\{\boldsymbol{Z}_{j-1},V\}$, $\{X,\boldsymbol{Z}_{j-1}\}$, and $\{\boldsymbol{Z}_{j-1}\}$. In detail, each MI term is calculated as the weighted sum of the specific MI relevant to the outcome y of the class Y, where the weights are given by the probability of the event $Y=y$, i.e., $P(\{Y=y\})=P(y)$. The specific MI terms are computed as:

$$I(Y=y;X,\boldsymbol{Z}_{j-1},V) = \psi(N) - \psi(N_y) + \psi(k) - \langle\psi(m_{x,z_{j-1},v}+1)\rangle, \tag{7}$$
$$I(Y=y;\boldsymbol{Z}_{j-1},V) = \psi(N) - \psi(N_y) + \langle\psi(m^y_{z_{j-1},v}+1)\rangle - \langle\psi(m_{z_{j-1},v}+1)\rangle, \tag{8}$$
$$I(Y=y;X,\boldsymbol{Z}_{j-1}) = \psi(N) - \psi(N_y) + \langle\psi(m^y_{x,z_{j-1}}+1)\rangle - \langle\psi(m_{x,z_{j-1}}+1)\rangle, \tag{9}$$
$$I(Y=y;\boldsymbol{Z}_{j-1}) = \psi(N) - \psi(N_y) + \langle\psi(m^y_{z_{j-1}}+1)\rangle - \langle\psi(m_{z_{j-1}}+1)\rangle, \tag{10}$$

where $\langle\cdot\rangle$ denotes the average over observations, $\psi(\cdot)$ is the digamma function, N is the total number of samples, N_y is the number of samples associated to the outcome $Y=y$, k is the number of neighbors searched for each sample in the space $\{X,\boldsymbol{Z}_{j-1},V|Y=y\}$, referring to samples in $\{X,\boldsymbol{Z}_{j-1},V\}$ associated to the outcome $Y=y$, $m^y_{z_{j-1},v}$, $m^y_{x,z_{j-1}}$, and $m^y_{z_{j-1}}$ are the number of samples in $\{\boldsymbol{Z}_{j-1},V|Y=y\}$, $\{X,\boldsymbol{Z}_{j-1}|Y=y\}$ and $\{\boldsymbol{Z}_{j-1}|Y=y\}$ for the determined distance, respectively, and $m_{x,z_{j-1},v}$, $m_{z_{j-1},v}$, $m_{x,z_{j-1}}$, and $m_{z_{j-1}}$ are those counted considering all samples in $\{X,\boldsymbol{Z}_{j-1},V\}$, $\{\boldsymbol{Z}_{j-1},V\}$, $\{X,\boldsymbol{Z}_{j-1}\}$ and $\{\boldsymbol{Z}_{j-1}\}$.

Given that the set $\boldsymbol{Z}_{min}$ minimizes the information shared between X and Y, that determines the unique contribution of the source variable to the class, and that the redundant contribution of the variables in $\boldsymbol{Z}_{min}$ is obtained as the difference between $I(Y;X)$ and $I(Y;X|\boldsymbol{Z}_{min})$, it is necessary to assess first the statistical significance of $I(Y;X)$ to determine if there is any potential unique or redundant contribution. Therefore, in the initial step, the MI term $I(Y;X)$ is computed as

$$I(Y;X) = \sum_{y\in\mathcal{A}_Y} P(y)I(Y=y;X), \tag{11}$$

with the specific MI term being

$$I(Y=y;X) = \psi(N) - \psi(N_y) + \psi(k) - \langle\psi(m_X+1)\rangle; \tag{12}$$

here, k is the number of neighbors searched in the space $\{X|Y=y\}$, and m_X is the number of samples counted in the same range for the space $\{X\}$. A surrogate test is performed to test the null hypothesis of independence between Y and X. To simulate the null distribution, the samples of X are randomly shuffled, and the MI in (11) is recomputed. This procedure is repeated N_{surr} times, producing a distribution of surrogate values. The significance of the original measure is then evaluated by comparing it to the 95th percentile of the surrogate distribution. If the observed value exceeds this threshold, the null hypothesis is rejected and the MI is assumed to be greater than 0. Otherwise, this would indicate that the empty set is the final set $\boldsymbol{Z}_{min}=\emptyset$, thus shortening the search procedure.

Similarly, for determined values of $I(Y;X|\boldsymbol{Z}_{j-1},V_j)$ and $I(Y;X|\boldsymbol{Z}_{j-1})$, a statistical analysis of their difference is performed to test the null hypothesis that the candidate V_j does not alter the conditional information shared by X and Y given $\boldsymbol{Z}_{j-1}$; if the null hypothesis is verified, V_j will not be included in the sets of selected variables $\boldsymbol{Z}_{min}$ or $\boldsymbol{Z}_{max}$ and the search is terminated. Specifically, the significance of $I(Y;X|\boldsymbol{Z}_{j-1},V_j) - I(Y;X|\boldsymbol{Z}_{j-1})$ is tested for the $\boldsymbol{Z}_{max}$ search, and that of $I(Y;X|\boldsymbol{Z}_{j-1}) - I(Y;X|\boldsymbol{Z}_{j-1},V_j)$ for the $\boldsymbol{Z}_{min}$ search. With the same null hypothesis, a test is done by randomly permuting the investigated variable V_j N_{surr} times, producing a distribution of surrogate values of the difference measure. If the original measurement exceeds the 95th percentile threshold, the difference is deemed significant, and the contributing variable V_j is added to the appropriate set.

The specific selection process, considering the estimation approach, for identifying the subsets $\boldsymbol{Z}_{min}$ and $\boldsymbol{Z}_{max}$ is summarized in Algorithms 1 and 2, respectively (see Appendix). Once they are found, all the terms needed to obtain the contributions in (3), i.e., $I(Y;X)$, $I(Y;X|\boldsymbol{Z}_{min})$ and $I(Y;X|\boldsymbol{Z}_{max})$, are computed again considering as neighbor search space $\{X,\boldsymbol{Z}_{min}\cup\boldsymbol{Z}_{max}\}$. Specifically, Eqs. (9) and (10) are employed, replacing $\boldsymbol{Z}_{j-1}$ with $\boldsymbol{Z}_{min}$ and $\boldsymbol{Z}_{max}$. The overall computational cost of the search algorithms and final mutual information calculations is dominated by the kNN neighborhood search. Using k-d tree implementations, the average time complexity is $\mathcal{O}(N\log N)$ for a low-dimensional setting, with a corresponding space complexity of $\mathcal{O}(N)$ [14]. For a high-dimensional setting the time complexity degrades to $\mathcal{O}(N^2)$.

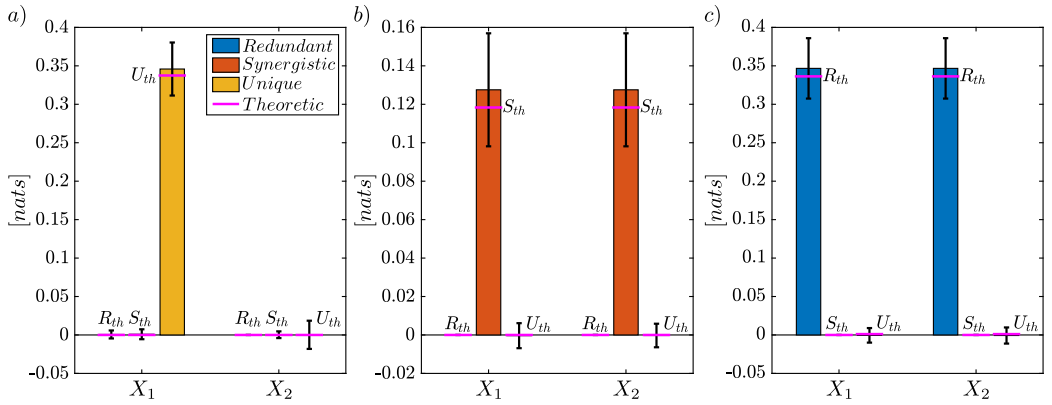

Fig. 2. Decomposition into unique, redundant, and synergistic components of the maximum information shared between the binary class variable Y and the feature X_i $(i = 1, 2)$ for the simulated systems with parameters given by Eq. (17) (a), Eq. (18) (b), Eq. (19) (c). Barplots represent the mean values (with ±2 SD) of the various information contributions, for features X_1 and X_2 showcasing unique (a), synergistic (b), and redundant (c) predictive information; magenta lines represent the theoretical values, computed as in Section 3.1.

3. Validation on synthetic data with mixed Gaussian distributions

To establish a quantitative reference for evaluation, we introduce a Gaussian-based framework in which the CMI can be computed theoretically. Assuming that the conditional joint distribution $\{X, \boldsymbol{Z} \mid Y = y\}$ is multivariate Gaussian, it allows for the derivation of ground-truth CMI values via Monte Carlo (MC) estimation. These theoretical measures provide a benchmark against which the proposed estimator can be validated. The following subsections detail the derivation of these theoretical values and the validation experiments designed to assess the estimator's performance under controlled dependency structures.

3.1. Theoretical computation

Under the Gaussian-based framework, we derive the theoretical values of the CMI for a known system. With the assumption that for a given discrete realization $Y = y$ the underlying joint conditioned space $\{X, \boldsymbol{Z}|Y = y\}$ follows a multivariate Gaussian distribution, the joint distribution $p(X, \boldsymbol{Z})$ becomes a Gaussian Mixture Model (GMM), formed by the class weights and conditional probability densities, i.e.,

$$p(X, \boldsymbol{Z}) = \sum_{\mathcal{A}_Y} P(y)p(X, \boldsymbol{Z}|Y = y). \tag{13}$$

To evaluate the CMI values, starting from the definition

$$I(Y; X|\boldsymbol{Z}) = \sum_{\mathcal{A}_Y} \int_{D_X} \int_{D_{\boldsymbol{Z}}} p(y, x, \boldsymbol{z}) \log \left(\frac{p(\boldsymbol{z})p(y, x, \boldsymbol{z})}{p(x, \boldsymbol{z})p(y, \boldsymbol{z})} \right) dx d\boldsymbol{z}, \tag{14}$$

we have

$$I(Y; X|\boldsymbol{Z}) = \sum_{\mathcal{A}_Y} P(y) \int_{D_X} \int_{D_Z} p(x, \boldsymbol{z}|y) \log \left(\frac{p(\boldsymbol{z})p(x, \boldsymbol{z}|y)}{p(x, \boldsymbol{z})p(\boldsymbol{z}|y)} \right) dx d\boldsymbol{z}, \tag{15}$$

with each of the terms, apart from $P(y)$, representing the density for either a multivariate Gaussian distribution or a GMM. The integral represents the specific CMI for a given realization $Y = y$, i.e., $I(Y = y; X|\boldsymbol{Z})$.

Given the initial distribution assumption constraints, this formulation allows us to simulate complex, multimodal dependencies in the joint space while retaining full knowledge of the underlying generative processes. Since a closed-form expression for the specific CMI is not available in this setting, we resort to high-precision numerical estimation using the MC method [15], with the integral approximated as

$$I(Y = y; X|\boldsymbol{Z}) = \hat{\mathbb{E}}_{p(x,z|y)} \left[\log \left(\frac{p(\boldsymbol{z})p(x, \boldsymbol{z}|y)}{p(x, \boldsymbol{z})p(\boldsymbol{z}|y)} \right) \right]. \tag{16}$$

To obtain this estimate, we draw 10^6 samples from the conditional distribution space $\{X, \boldsymbol{Z}|Y = y\}$, and for each sample we evaluate the entire log-ratio expression in (16). The MC estimate of $I(Y = y; X|\boldsymbol{Z})$ is then computed as the average of these expressions across all sampled instances.

3.2. Computation on simulated data

This section presents a series of experiments on synthetically generated data designed in a way such that the underlying generative models are fully specified, allowing to capture the distinct components of information in structured settings. The goals of these experiments are to investigate the true values of the information-theoretic contributions of individual feature variables as

a function of the simulation parameters, and to quantify the bias and variance of the proposed estimator relative to the theoretical values of the information measures. To this end, we first construct a series of simple setups that focus on showcasing each possible type of individual contributions of the feature variables to the target (i.e., unique, redundant, and synergistic), and then design an additional, more complex simulation covering all possible types of interaction.

In detail, we consider a class variable Y and a set of m feature variables $\boldsymbol{X} = \{X_1, \dots, X_m\}$; the features are assumed to have a multivariate Gaussian distribution when associated with each specific outcome of the class, i.e., $\boldsymbol{X}|Y = y \sim \mathcal{N}(\mu_{\boldsymbol{X}}^{(y)}, \Sigma_{\boldsymbol{X}}^{(y)})$, where $\mu_{\boldsymbol{X}}^{(y)}$ and $\Sigma_{\boldsymbol{X}}^{(y)}$ are the mean vector and the covariance matrix of $\boldsymbol{X}$ evaluated when $Y = y$. Inspired by the examples in [7,10], we outline a set of simulations inducing behaviors that elicit specific types of information contributions. This is achieved by imposing changes in the mean of a feature variable across class states (i.e., imposing different values of $\mu_{X_i}^{(y)}$ at varying y) to reproduce relevant contributions, with different levels of high-order interaction determined by the correlation among features (i.e., by the off-diagonal elements of $\Sigma_{\boldsymbol{X}}^{(y)}$), and by imposing no changes of the mean and variance of a feature across states to reproduce an absence of contribution. In the first three simulations we consider only $m = 2$ features and show how unique, synergistic, or redundant behaviors can be isolated acting on the simulation parameters, while in the last simulation with $m = 6$ we demonstrate how combined changes in the distribution parameters of several features can give rise to coexisting contributions. The kNN estimation method was applied over simulations lasting $N_{sample} = 1000$ samples per realization $Y = y$, considering the typical setting $k = 10$ for the number of neighbors [16,17]; the statistical significance of the selected variables was tested generating $N_{surr} = 100$ surrogate sequences. Each simulation was repeated $N_{sim} = 100$ times to assess the confidence intervals of each measure, while theoretical values were obtained using $N_{MC} = 10^6$ MC samples for each realization $Y = y$.

3.2.1. Unique components

First, we consider a 2-class ($\mathcal{A}_Y = \{y_1, y_2\}$), 2-variable ($\boldsymbol{X} = \{X_1, X_2\}$) setup with the following model parameters:

$$\begin{aligned} &P(y_1) = P(y_2) = 0.5, \\ &\mu_{\boldsymbol{X}}^{(y_1)} = [1, 1], \\ &\mu_{\boldsymbol{X}}^{(y_2)} = [-1, 1], \\ &\Sigma_{\boldsymbol{X}}^{(y_1)} = \Sigma_{\boldsymbol{X}}^{(y_2)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \end{aligned} \tag{17}$$

The class-specific distributions are $X_1|Y = y_1 \sim \mathcal{N}(1, 1)$, $X_1|Y = y_2 \sim \mathcal{N}(-1, 1)$ and $X_2|Y = y_1 \equiv X_2|Y = y_2 \sim \mathcal{N}(1, 1)$, showing that the mean of X_1 shifts between the two classes while the distribution of X_2 remains the same. The diagonal covariance implies that the two features are uncorrelated.

The results of the analysis are shown in Fig. 2(a), documenting that the predictive information is fully determined by the purely unique contribution of X_1, which can be ascribed to the shift in mean between the two class labels; the feature X_2, being uncorrelated with X_1 and equally distributed within the two classes, displays no relevancy to the class variable Y; the uncorrelation between the features explains also the absence of redundant or synergistic effects. The distribution of the information measures estimated over N_{sim} realizations shows the accuracy of the estimates, which exhibit a small negative bias and limited variability.

3.2.2. Synergistic components

This simulation reproduces again a 2-class, 2-variable setup, now characterized by the following parameters:

$$\begin{aligned} &P(y_1) = P(y_2) = 0.5, \\ &\mu_X^{(y_1)} = \mu_X^{(y_2)} = [0, 0], \\ &\Sigma_X^{(y_1)} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \\ &\Sigma_X^{(y_2)} = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}, \end{aligned} \tag{18}$$

corresponding to identical standard normal distributions of the two feature variables within the two classes ($X_i|Y = y_j \sim \mathcal{N}(0, 1)$, $i, j = 1, 2$). Moreover, as seen in the covariance matrix, the two features are positively correlated when the class takes the value y_1, and anticorrelated when the class takes the value y_2.

The results for this setup, reported in Fig. 2(b), document the absence of unique contributions, reflecting the fact that the class-specific mean and variance remained unchanged across the realizations of Y. However, their joint interaction, here captured by the class-specific covariance, carries discriminative information about the class, which is encoded in the significant synergistic contribution observed in the absence of redundancy. In this case, synergy arises from the fact that the relationship between the features differs across the realizations of Y, so that both X_1 and X_2 need to be known to predict the class.

3.2.3. Redundant components

Another 2-class, 2-variable setup is then considered, featuring the following parameters:

$$\begin{aligned} &P(y_1) = P(y_2) = 0.5, \\ &\mu_X^{(y_1)} = [1, 1], \\ &\mu_X^{(y_2)} = [-1, -1], \\ &\Sigma_X^{(y_1)} = \Sigma_X^{(y_2)} = \begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix}. \end{aligned} \tag{19}$$

In this case, the class-specific distributions are the same for the two features, i.e., $X_i|Y = y_1 \sim \mathcal{N}(1,1)$, $X_i|Y = y_2 \sim \mathcal{N}(-1,1)$, $i = 1,2$, showing a shift in the mean between the two classes. Moreover, the features are strongly correlated as documented by the covariance value of 0.99.

The analysis of this configuration reveals a fully overlapping contribution of the two features to the prediction of the class labels, as seen in Fig. 2(c) where the predictive information is entirely redundant, with zero values for the unique and synergistic components. These results are explained by the fact that the two features share the same class-dependent change in the mean value.

3.2.4. Multiple interactions

The last simulation, which displays a setup with more complex interactions, including combined contributions of multiple components, is designed to portray the robustness of the proposed estimation strategy, considering a higher-dimensional interaction system. Specifically, a 2-class ($\mathcal{A}_Y = \{y_1, y_2\}$), 6-variable ($\boldsymbol{X} = \{X_1, \dots, X_6\}$) system is formed with the following model parameters:

$$
\begin{aligned}
&P(y_1) = P(y_2) = 0.5,\\
&\mu_X^{(y_1)} = [1, 0.5, 0.5, 0, 1, 0],\\
&\mu_X^{(y_2)} = [-1, -0.5, -0.5, 0, -1, 0],\\
&\Sigma_X^{(y_1)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 0.25 & 0 & 0 & 0\\ 0 & 0.25 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 0.5 & 0\\ 0 & 0 & 0 & 0.5 & 1 & 0\\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},\\
&\Sigma_X^{(y_2)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 0.25 & 0 & 0 & 0\\ 0 & 0.25 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & -0.5 & 0\\ 0 & 0 & 0 & -0.5 & 1 & 0\\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.
\end{aligned}
\tag{20}
$$

According to (20), the joint distribution of the features display the following characteristics: the feature X_1 exhibits different class-specific mean while remaining uncorrelated with the other features; X_2 and X_3 have the same different class-specific means and additionally display a positive correlation for both realizations of Y; while X_4 does not exhibit a change in the class-specific mean, X_5 does, and they jointly have a class-dependent covariance value; the feature X_6 has no class-specific change in the probability distribution and is uncorrelated with the other features.

Fig. 3 shows the results of the kNN estimator applied to the defined system, for which several expected contributions are observed as follows. The change in the mean value across class realizations enables the variables X_1, X_2, X_3, and X_5 to be predictive for the class, thereby providing unique contributions, with a higher magnitude for X_1 and X_5 which display the largest mean shift. Moreover, since these features exhibit similar class-dependent changes in the mean, their contribution is also partially redundant; we note that redundancy arises not only for the correlated variables X_2 and X_3, but also for X_1 and X_5 which are not correlated with each other or with X_2 and X_3. Additionally, a synergistic contribution is detected for X_4 and X_5, reflecting the change in sign of their correlation across the output values of the class variable Y. Finally, X_6 does not provide any contribution to Y, as expected by the fact that its probability density is not class-dependent and it is uncorrelated with the other features. Remarkably, the accuracy of the kNN estimation remains high also in this higher-dimensional system, as documented by the very small bias and limited variance of the estimates of unique, redundant, and synergistic components computed across N_{sim} realizations (barplots and errorbars in Fig. 3).

Fig. 4 depicts the selection of the variables included for this simulation into the conditioning sets $\boldsymbol{Z}_{min}$ and $\boldsymbol{Z}_{max}$ during the search for redundant and synergistic contributions to the class label, both theoretical based on the exact CMI values and practical based on the estimated values and the surrogate test for significance. Moreover, Fig. 5 shows the order according to which the variables are selected by the sequential procedure for the formation of the sets $\boldsymbol{Z}_{min}$ and $\boldsymbol{Z}_{max}$.

The group of features X_1, X_2, X_3 or X_5 are selected into the conditioning set $\boldsymbol{Z}_{min}$ when one of them is taken as source variable (Fig. 4(a), columns 1,2,3,5), in an order generally prioritizing X_1 and X_5 followed by X_2 and X_3 (Fig. 5(a)). This selection is expected, given the previously discussed marginal predictive capability of these features, which leads to redundant contributions. Moreover, X_4 is also included theoretically into the set $\boldsymbol{Z}_{min}$ when X_1, X_2, or X_3 are taken as sources; in the practical estimation this selection occurred only in a fraction of realizations (89%, 16%, and 17% for X_1, X_2, and X_3, Fig. 4(a)) and at the third or fourth iteration of the searching procedure (Fig. 5(a)). The inclusion of X_4 is more nuanced, as it exhibits a synergistic interaction with X_5, and its selection along with X_5 further decreases the CMI.

With regard to the procedure of maximization of the CMI, the only significant selections involve X_4 and X_5, which are always reciprocally included into $\boldsymbol{Z}_{max}$ at the first iteration (Fig. 4(b), Fig. 5(b)). This selection is meaningful as it relates to the synergistic contribution to the class labels expected for the features X_4 and X_5. Only minor spurious inclusions, not expected theoretically and related to false positive selections, are observed in the other cases.

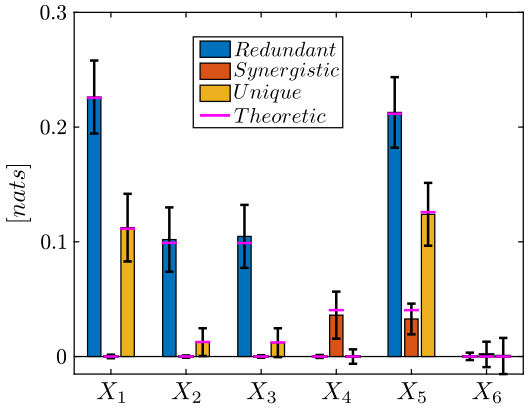

Fig. 3. Decomposition into unique, redundant and synergistic components of the maximum information shared between the binary class variable Y and the feature X_i ($i = 1, \dots, 6$) for the simulated systems with parameters given by Eq. (20). Barplots represent the mean values (with ±2 SD) of the various information contributions estimated across several realizations, while magenta lines represent the theoretical values, computed as in Section 3.1. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

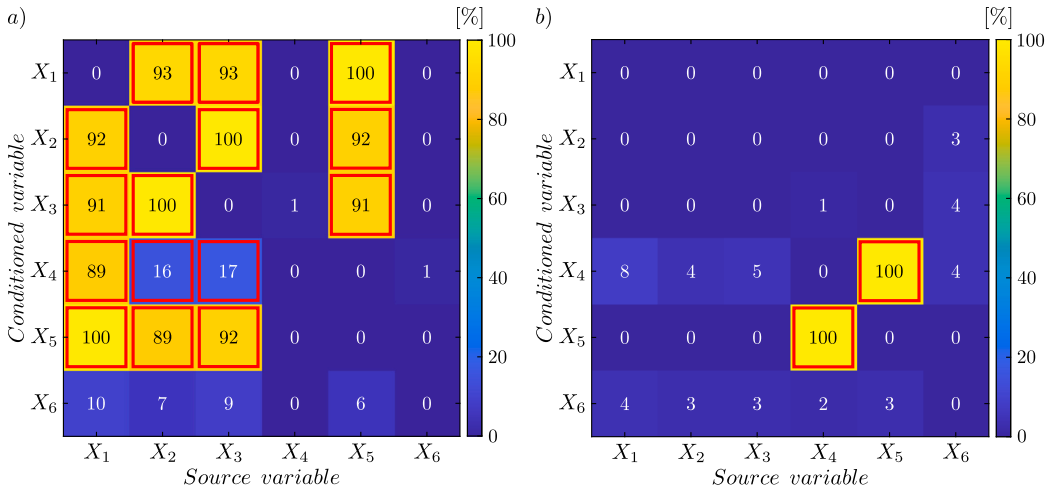

Fig. 4. Selection of the features to be included in the conditioning set during the search for redundancy and synergy in the simulated system with parameters given by Eq. (20). Heatmaps show, for any source variable (columns), the selection of the conditioning variables (rows; cells with red border: theoretical selection; cell color and number: percentage of selections over N_{sim} realizations) included within the set $\boldsymbol{Z}_{min}$ (a) or within the set $\boldsymbol{Z}_{max}$ (b). In panel (a), the source variables X_1, X_2, X_3, and X_5 consistently contribute to minimizing the CMI value, with X_4 further decreasing it, except when the starting variable is X_5. Panel (b) shows that the only synergy-induced CMI increase happens between X_4 and X_5. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4. Application to non-Gaussian data

To investigate the behavior of the kNN CMI estimator when dealing with distributions for which the theoretical values of the information components cannot be computed, we design two validation scenarios: one using synthetic data and the other using real-world data. In the synthetic case, we interpret the results according to the expected interaction effects based on the experimental setup, while in the real-world case, we validate the findings against known analyses reported in the literature.

4.1. Synthetic data

In this setup, we simulate from four independent uniform distributions, $X_i \sim \mathcal{U}(-1, 1)$, $i \in \{1, \dots, 4\}$. The class discrete variable Y is formed as $Y = \Theta(X_1 \cdot X_2) + \Theta(X_3) + \Theta(X_4)$, where $\Theta(\cdot)$ denotes the Heaviside function, resulting in a class alphabet $A_Y = \{0, 1, 2, 3\}$. Across $N_{sim} = 100$ simulations, the proposed estimator receives as inputs the class variable Y and the set of feature variables $X_1, \dots, X_4$, along with two additional features X_5 and X_6 defined as $X_5 = X_4 + M$ (a noisy version of X_4 with $M \sim \mathcal{N}(0, 1)$), and as an independent uniform variable $X_6 \sim \mathcal{U}(-1, 1)$. Based on this setup, we expect a synergistic effect between X_1 and X_2, as

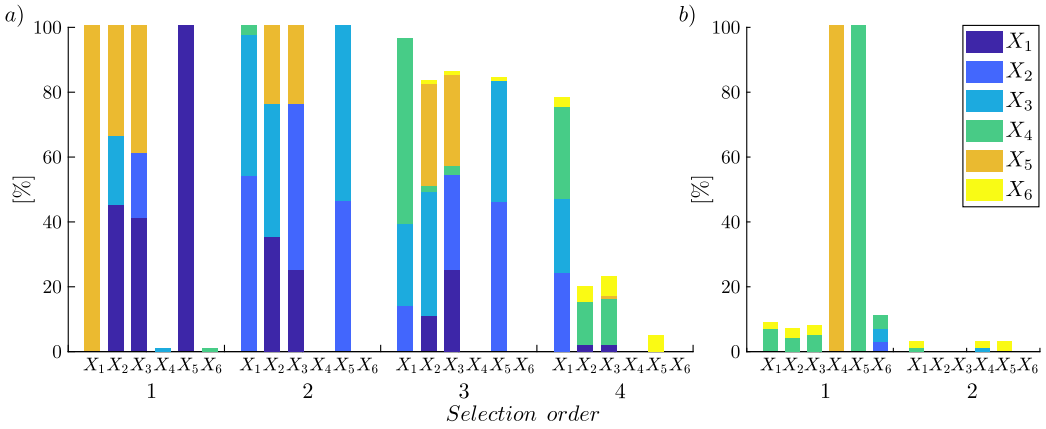

Fig. 5. Order and percentage of occurrence of the features selected as conditioning variables for the MI between each source feature and the target class during the search for redundancy ((a), inclusion into $\boldsymbol{Z}_{min}$) and synergy ((b), inclusion into $\boldsymbol{Z}_{max}$). The height of each bar quantifies the percentage of times over N_{sim} realizations for which a variable was selected for the source feature X_i ($i = 1, \ldots, 6$) at each iteration $j \in \{1, 2, \ldots\}$; the color information represents which variable was selected. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

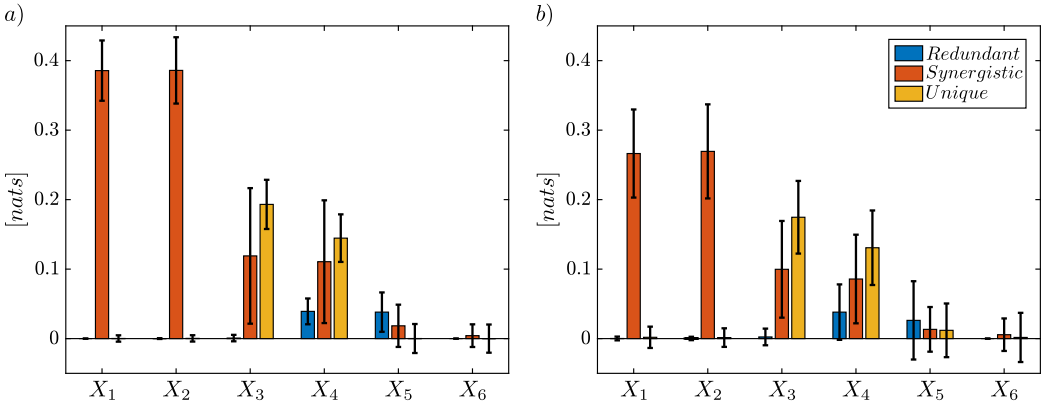

Fig. 6. Decomposition into unique, redundant and synergistic components of the maximum information shared between the 4-class variable Y and each of the six source feature variables simulated as in Section 4.1. Barplots represent the mean value (±2 SD) of the various information contributions estimated across several realizations lasting $N_{sample} = 1000$ (a) and $N_{sample} = 300$ (b). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

they interact in a product-wise manner (like in the XOR logic-gate), while X_3 and X_4 are expected to provide unique contributions. Moreover, due to the additive nature of the formulation of the class variable Y, we anticipate some amount of synergistic interaction among these four variables as well. The variable X_5, being a noisy version of X_4, is expected to share with it a certain amount of redundant predictive information. Finally, X_6, being independent from all other variables, should show no contributions to the target class. To evaluate the sensitivity of the measures to sample size, the analysis was repeated using $N_{sample} = 1000$ and $N_{sample} = 300$ as sample size.

Results in Fig. 6 confirm the expected synergistic interaction of X_1 and X_2, as well as the unique contributions of X_3 and X_4. Additionally, we correctly identify redundancy in X_4 since X_5 represents a noisy version of it. Finally, X_6, which is independent of the other variables, shows no contributions concerning the class. We additionally confirm the expectation of the synergistic effect caused by the additive effect of X_1, X_2 with X_3 and X_4. Notably, the observed synergistic effect has lower values when using a smaller sample size.

4.2. Real-world data

For this analysis, we utilize a subset of The Cancer Genome Atlas - Breast Invasive Carcinoma (TCGA-BRCA) RNA sequencing dataset [18] containing gene expression measurements from breast cancer tissues. The data, obtained from the Marginal Contribution Feature Importance (MCI) [19] (GitHub page), comprises 572 samples characterized by the expression levels of 50 genes, spanning four molecular breast cancer subtypes: 304 Luminal A (LumA), 114 Luminal B (LumB), 105 Basal, and 49 HER2-enriched. The approach introduced in this work is used to investigate the contribution of gene expressions in determining the molecular subtype, considered as class outputs in the analysis. To estimate the confidence intervals of the informational contribution of each feature,

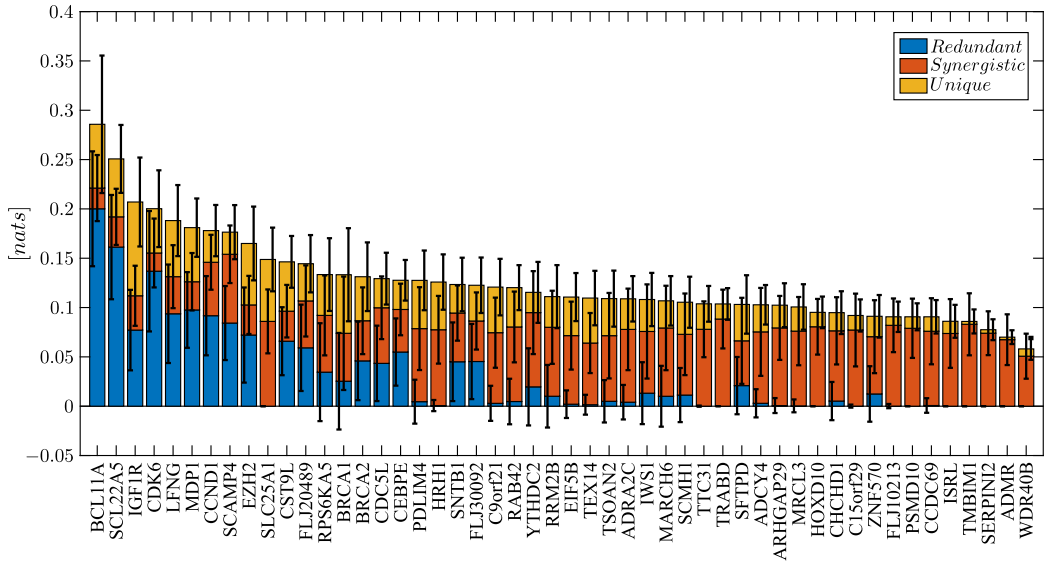

Fig. 7. TCGA-BRCA feature analysis results. Barplots represent the contributions obtained with the kNN-based estimation approach for each gene, shown according to a decreasing CMI. The bar and confidence intervals represent the mean ±2 SD across 100 bootstrap samples obtained for the unique, synergistic, and redundant contributions for each gene. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

we generate a bootstrapped dataset with 572 samples and $N_{sim} = 100$ sampling repetitions, while ensuring the same class counts. Using the kNN estimator, we fix the number of neighbors at $k = 10$, and employ $N_{surr} = 100$ surrogate samples for the selection criteria.

The results shown in Fig. 7 display the ranked top contributors according to the average CMI estimated between the target class and each observed feature, conditioned on the subset of features that maximizes this measure, $I(Y; X|\boldsymbol{Z}_{max})$. According to [20,21], the 10 genes BCL11A, SLC22A5, CDK6, IGF1R, LFNG, CCND1, EZH2, BRCA1, BRCA2, and TEX14, are known breast cancer-associated markers; in our analysis, most of these genes exhibit high CMI values based on their mean contributions. The findings, based on the selected set of features, are consistent with previous, similar works in [19,21,22], that identify the top contributing set of genes. However, as also noted by Catav et al. [19], we observe substantial contributions from MDP1, SCAMP4, SLC25A1, and CST9L genes, some of which are not linked to breast cancer based on the literature. SLC25A1, however, has been reported by Catalina-Rodriguez et al. [23] to exhibit elevated expression in breast cancer tissues, suggesting a potential relevance to subtype differentiation.

Considering the 10 selected genes, the proposed decomposition allows for the observation of component-wise interactions, noting that the majority of the significant contribution comes from the redundancy components. A similar study by Westphal et al. [20] performs a PID-like analysis to determine synergistic and redundant contributions in addition to the MI that each feature has with the target, $I(Y; X_i)$. However, while their results indicate a lower contribution of synergy and redundancy, we note that the nature of these components is different. Specifically, they define Feature-Wise Synergy and Redundancy (FWS and FWR) through the concept of interaction information. To obtain FWS they maximize $I(Y; X_i; \boldsymbol{Z}^{ms})$, where $\boldsymbol{Z}^{ms}$ is the subset of features for maximum synergy, while the FWR is determined as FWS – $I(Y; X_i; \boldsymbol{Z}_{\setminus X_i})$, where $\boldsymbol{Z}_{\setminus X_i}$ is the subset of features without X_i.

For our approach, the redundancy measure for a given gene indicates shared information when the inclusion of another gene results in a reduced CMI. Analyzing the selection order (not shown in the work), we observe shared redundancy between BCL11A and SLC22A5, which is expected given their subtype-specific roles, as BCL11A is associated with the basal-like phenotype [24], whereas SLC22A5 is linked to luminal phenotypes and is known to be an estrogen receptor regulated gene [25,26]. Furthermore, LFNG and BCL11A exhibit redundant information, likely due to their shared association with the basal-like subtype [27].

To assess how the identified features contribute to the final classification performance, we compared the ranking obtained using the proposed approach with those derived from popular model-free feature importance ranking algorithms, MRMR [28] and RELIEFF [29] (using $k = 10$ nearest neighbors). The ranking produced by the proposed method is based on the total CMI measure, representing the overall informational relevance of each feature with respect to the class variable.

An SVM classifier with parameter $C = 1$ (selected heuristically) and leave-one-out cross-validation was employed to evaluate the classification accuracy as a function of the number of top-ranked features included in the model, ranging from 1 to 50. For completeness, the same evaluation was also performed using the reversed feature order (starting from the least informative features for each method), to confirm that features ranked lowest indeed contribute the least to classification.

The results, shown in Fig. 8, demonstrate that the features prioritized by the CMI and RELIEFF consistently achieve higher accuracy with fewer features compared to MI-based MRMR. On the other hand, the lowest-ranked CMI features consistently yielded the lowest classification accuracy when included first. It should be noted, however, that a direct value-wise comparison across

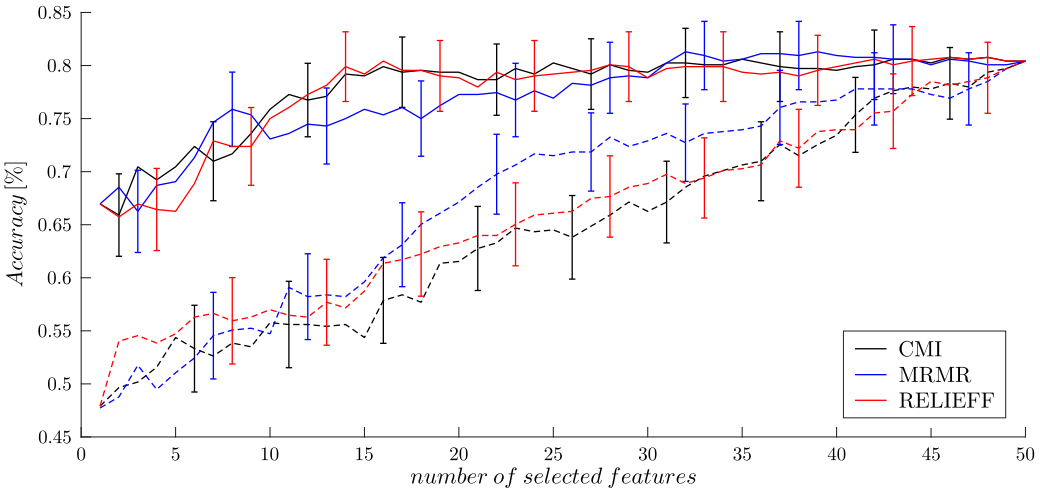

Fig. 8. Classification accuracy obtained using features ranked by the total CMI obtained in Fig. 7, MRMR, and RELIEFF (with $k = 10$ nearest neighbors), for an SVM classifier with parameter $C = 1$ under leave-one-out cross-validation. Solid lines show the accuracy as a function of the number of top-ranked features included. Dashed lines represent results using the reversed feature order, starting from the least informative features, for each method. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

features is affected by a small degree of estimation bias, as the total CMI values are computed through independent estimation procedures for each feature. Nevertheless, the observed trend clearly indicates that the CMI-based ranking captures the discriminative information for the given classification problem. Beyond this comparative evaluation, the proposed framework opens further possibilities for feature selection strategies that explicitly exploit the decomposed informational roles identified by the method. In particular, separating redundant, synergistic, and unique contributions provides a means to design more targeted selection criteria.

5. Conclusions

In this work, we presented a method to assess feature importance accounting for high-order interactions among features, designed by fully leveraging an information-theoretic approach based on CMI, estimated using the kNN algorithm, with a focus on classification problems involving continuous features. This approach enables the disentanglement of individual feature contributions into unique, synergistic, and redundant components in a model-independent manner.

Through evaluation on synthetic datasets with known Gaussian distributions, we validated the accuracy of the proposed method against theoretical expectations. Additional experiments on non-Gaussian data, including a real-world breast cancer gene expression dataset, demonstrated that the estimator reliably captured the expected interaction patterns based on the experimental setup or established findings in the literature.

Compared to PID, the proposed CMI-based Hi-Fi extension offers an alternative perspective for analyzing high-order feature interactions. While PID typically focuses on decomposing the total mutual information into predefined atomic components requiring a redundancy measure, whose definition is debated in literature, our approach provides a more flexible, observation-driven decomposition based on traditional information-theoretic measures. Additionally, our method naturally accommodates direct CMI estimation without the need to exhaustively compute all subset combinations, offering scalability and practical relevance for larger systems.

The proposed framework offers potential applications in machine learning, particularly as a feature selection criterion by prioritizing features with predominantly unique or synergistic contributions. Similarly, it can assist in the development of models that better capture the interaction patterns by explicitly accounting for the information structure of the inputs. Nevertheless, further investigation is necessary to assess the estimator's performance in high-dimensional, highly interactive systems, where the curse of dimensionality may introduce challenges for CMI estimation, as well as to complete a more formal analysis comparing our approach with others proposed in the literature.

CRediT authorship contribution statement

Ivan Lazic: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation. **Chiara Barà:** Writing – original draft, Visualization, Software, Methodology, Investigation, Data curation. **Marta Iovino:** Writing – review & editing, Visualization, Software, Investigation, Formal analysis. **Sebastiano Stramaglia:** Writing – review & editing, Supervision, Investigation, Conceptualization. **Karolina Kasas-Lazetic:** Formal analysis, Visualization, Writing – review & editing. **Nikša Jakovljević:** Writing – review & editing, Validation, Supervision, Investigation. **Luca Faes:** Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

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.

Acknowledgments

This research was co-funded by the Italian Complementary National Plan PNC-I.1 "Research initiatives for innovative technologies and pathways in the health and welfare sector" D.D. 931 of 06/06/2022, "DARE - DigitAl lifelong pRevEntion" initiative, code PNC0000002, CUP: B53C22006460001, and by the project "HONEST - High-Order Dynamical Networks in Computational Neuroscience and Physiology: an Information-Theoretic Framework", Italian Ministry of University and Research (funded by MUR, PRIN 2022, code 2022YMHNPY, CUP: B53D23003020006). I.L., N.J., and K.K.-L. are supported by the Ministry of Science, Technological Development and Innovation (Contract No. 451-03-137/2025-03/200156) and the Faculty of Technical Sciences, University of Novi Sad, through the project "Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad 2025" (No. 01-50/295).

Appendix. Subset search algorithms

The pseudocode implementations of the $\boldsymbol{Z}_{min}$ and $\boldsymbol{Z}_{max}$ searching procedures described in Section 2.2 are reported here. These correspond to the algorithmic steps referred to in the main text, where the main section provides their methodological description and rationale.

Data: Y, X, $\boldsymbol{Z}$, N_{surr}
Result: $\boldsymbol{Z}_{min}$
$\boldsymbol{Z}_0 \leftarrow \emptyset$;
$n \leftarrow |\boldsymbol{Z}|$; /*number of features in the set*/
$j \leftarrow 1$;
$I_{min} \leftarrow I(Y; X \mid \boldsymbol{Z}_0) \equiv I(Y; X)$;
for $i \in \{1, \ldots, N_{surr}\}$ **do**
  $X^* \leftarrow \mathrm{shuffle}(X)$; /*Permutation surrogate*/
  $I_i^* \leftarrow I(Y; X^*)$;
end
if $I_{min} < \mathrm{percentile}(\boldsymbol{I}^*, 95)$ **then**
  $\boldsymbol{Z}_{min} \leftarrow \boldsymbol{Z}_0$
else
  while $j \leq n$ **do**
    $V_j \leftarrow \arg\min_{V \in \boldsymbol{Z} \setminus \boldsymbol{Z}_{j-1}} I(Y; X | \boldsymbol{Z}_{j-1}, V)$;
    determine $I(Y; X \mid \boldsymbol{Z}_{j-1}, V_j)$;
    determine $I(Y; X \mid \boldsymbol{Z}_{j-1})$; /*In space $\{X, \boldsymbol{Z}_{j-1}, V_j\}$*/
    $\Delta I \leftarrow I(Y; X \mid \boldsymbol{Z}_{j-1}) - I(Y; X \mid \boldsymbol{Z}_{j-1}, V_j)$;
    if $\Delta I < 0$ **then**
      break;
    end
    for $i \in \{1, \ldots, N_{surr}\}$ **do**
      $V_j^* \leftarrow \mathrm{shuffle}(V_j)$; /*Permutation surrogate*/
      determine $I(Y; X \mid \boldsymbol{Z}_{j-1}, V_j^*)$;
      determine $I(Y; X \mid \boldsymbol{Z}_{j-1})$; /*In space $\{X, \boldsymbol{Z}_{j-1}, V_j^*\}$*/
      $\Delta I_i^* \leftarrow I(Y; X \mid \boldsymbol{Z}_{j-1}) - I(Y; X \mid \boldsymbol{Z}_{j-1}, V_j^*)$;
    end
    if $\Delta I < \mathrm{percentile}(\Delta \boldsymbol{I}^*, 95)$ **then**
      break;
    else
      $\boldsymbol{Z}_j \leftarrow \{\boldsymbol{Z}_{j-1}, V_j\}$;
      $j \leftarrow j + 1$;
    end
  end
  $\boldsymbol{Z}_{min} \leftarrow \boldsymbol{Z}_{j-1}$;
end

Algorithm 1: Greedy set $\boldsymbol{Z}_{min}$ search for given feature X

Data: Y, X, $\boldsymbol{Z}$, N_{surr}
Result: $\boldsymbol{Z}_{max}$
$\boldsymbol{Z}_0 \leftarrow \emptyset$;
$n \leftarrow |\boldsymbol{Z}|$; /*number of features in the set*/
$j \leftarrow 1$;
while $j \leq n$ **do**
  $V_j \leftarrow \arg\max_{V \in \boldsymbol{Z} \setminus \boldsymbol{Z}_{j-1}} I(Y; X | \boldsymbol{Z}_{j-1}, V)$;
  determine $I(Y; X \mid \boldsymbol{Z}_{j-1}, V_j)$;
  determine $I(Y; X \mid \boldsymbol{Z}_{j-1})$; /*In space $\{X, \boldsymbol{Z}_{j-1}, V_j\}$*/
  $\Delta I \leftarrow I(Y; X \mid \boldsymbol{Z}_{j-1}, V_j) - I(Y; X \mid \boldsymbol{Z}_{j-1})$;
  if $\Delta I < 0$ **then**
    break;
  end
  for $i \in \{1, \ldots, N_{surr}\}$ **do**
    $V_j^* \leftarrow \text{shuffle}(V_j)$; /*Permutation surrogate*/
    determine $I(Y; X \mid \boldsymbol{Z}_{j-1}, V_j^*)$;
    determine $I(Y; X \mid \boldsymbol{Z}_{j-1})$; /*In space $\{X, \boldsymbol{Z}_{j-1}, V_j^*\}$*/
    $\Delta I_i^* \leftarrow I(Y; X \mid \boldsymbol{Z}_{j-1}, V_j^*) - I(Y; X \mid \boldsymbol{Z}_{j-1})$;
  end
  if $\Delta I < \text{percentile}(\Delta \boldsymbol{I}^*, 95)$ **then**
    break;
  else
    $\boldsymbol{Z}_j \leftarrow \{\boldsymbol{Z}_{j-1}, V_j\}$;
    $j \leftarrow j + 1$;
  end
  $\boldsymbol{Z}_{max} \leftarrow \boldsymbol{Z}_{j-1}$;
end

Algorithm 2: Greedy set $\boldsymbol{Z}_{max}$ search for given feature X

Data availability

Data will be made available on request.

References

[1] Arsenault P-D, Wang S, Patenaude J-M. A survey of explainable artificial intelligence (XAI) in financial time series forecasting. ACM Comput Surv 2025;57(10):1–37. http://dx.doi.org/10.1145/3729531.
[2] Breiman L. Random forests. Mach Learn 2001;45(1):5–32. http://dx.doi.org/10.1023/A:1010933404324.
[3] Vergara JR, Estévez PA. A review of feature selection methods based on mutual information. Neural Comput Appl 2013;24(1):175–86. http://dx.doi.org/10.1007/s00521-013-1368-0.
[4] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;30.
[5] König G, Günther E, von Luxburg U. Disentangling interactions and dependencies in feature attribution. 2024, arXiv preprint arXiv:2410.23772. URL https://arxiv.org/abs/2410.23772.
[6] Williams PL, Beer RD. Nonnegative decomposition of multivariate information. 2010, arXiv preprint arXiv:1004.2515. URL https://arxiv.org/abs/1004.2515.
[7] Ontivero-Ortega M, Faes L, Cortes JM, Marinazzo D, Stramaglia S. Assessing high-order effects in feature importance via predictability decomposition. Phys Rev E 2025;111(3):L033301.
[8] Lei J, G'Sell M, Rinaldo A, Tibshirani RJ, Wasserman L. Distribution-free predictive inference for regression. J Amer Statist Assoc 2018;113(523):1094–111. http://dx.doi.org/10.1080/01621459.2017.1307116, arXiv:https://doi.org/10.1080/01621459.2017.1307116.
[9] Stramaglia S, Faes L, Cortes JM, Marinazzo D. Disentangling high-order effects in the transfer entropy. Phys Rev Res 2024;6(3):L032007.
[10] Wollstadt P, Schmitt S, Wibral M. A rigorous information-theoretic definition of redundancy and relevancy in feature selection based on (partial) information decomposition. J Mach Learn Res 2023;24(131):1–44, URL http://jmlr.org/papers/v24/21-0482.html.
[11] Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. Phys Rev E 2004;69(6):066138.
[12] Barà C, Antonacci Y, Iovino M, Lazic I, Faes L. Partial information decomposition for discrete target and continuous source random variables. Phys Rev E 2025;112:L012301. http://dx.doi.org/10.1103/58bg-5n9s, URL https://link.aps.org/doi/10.1103/58bg-5n9s.
[13] Xiong W, Faes L, Ivanov PC. Entropy measures, entropy estimators, and their performance in quantifying complex dynamics: Effects of artifacts, nonstationarity, and long-range correlations. Phys Rev E 2017;95:062114. http://dx.doi.org/10.1103/PhysRevE.95.062114, URL https://link.aps.org/doi/10.1103/PhysRevE.95.062114.
[14] Brown RA. Building a balanced k-d tree in $O(kn \log n)$ time. J Comput Graph Tech (JCGT) 2015;4(1):50–68, URL http://jcgt.org/published/0004/01/03/.
[15] Metropolis N, Ulam S. The Monte Carlo method. J Amer Statist Assoc 1949;44(247):335–41, URL http://www.jstor.org/stable/2280232.
[16] Barà C, Pernice R, Catania CA, Hilal M, Porta A, Humeau-Heurtier A, Faes L. Comparison of entropy rate measures for the evaluation of time series complexity: Simulations and application to heart rate and respiratory variability. Biocybern Biomed Eng 2024;44(2):380–92. http://dx.doi.org/10.1016/j.bbe.2024.04.004, URL https://www.sciencedirect.com/science/article/pii/S0208521624000287.

[17] Xiong W, Faes L, Ivanov PC. Entropy measures, entropy estimators, and their performance in quantifying complex dynamics: Effects of artifacts, nonstationarity, and long-range correlations. Phys Rev E 2017;95(6). http://dx.doi.org/10.1103/physreve.95.062114.
[18] Tomczak K, Czerwińska P, Wiznerowicz M. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Wspolczesna Onkol 2015;1A:68–77. http://dx.doi.org/10.5114/wo.2014.47136.
[19] Catav A, Fu B, Zoabi Y, Meilik ALW, Shomron N, Ernst J, Sankararaman S, Gilad-Bachrach R. Marginal contribution feature importance - an axiomatic approach for explaining data. In: Meila M, Zhang T, editors. Proceedings of the 38th international conference on machine learning. Proceedings of machine learning research, vol. 139, PMLR; 2021, p. 1324–35, URL https://proceedings.mlr.press/v139/catav21a.html.
[20] Westphal C, Hailes S, Musolesi M. Partial information decomposition for data interpretability and feature selection. 2024, arXiv preprint arXiv:2405.19212. URL https://arxiv.org/abs/2405.19212.
[21] Janssen J, Guan V, Robeva E. Ultra-marginal feature importance: Learning from data with causal guarantees. In: International conference on artificial intelligence and statistics. PMLR; 2023, p. 10782–814.
[22] Covert I, Lundberg SM, Lee S-I. Understanding global feature contributions with additive importance measures. Adv Neural Inf Process Syst 2020;33:17212–23.
[23] Catalina-Rodriguez O, Kolukula VK, Tomita Y, Preet A, Palmieri F, Wellstein A, Byers S, Giaccia AJ, Glasgow E, Albanese C, Avantaggiati ML. The mitochondrial citrate transporter, CIC, is essential for mitochondrial homeostasis. Oncotarget 2012;3(10):1220–35. http://dx.doi.org/10.18632/oncotarget.714.
[24] Katnik E, Gomulkiewicz A, Piotrowska A, Grzegrzolka J, Rusak A, Kmiecik A, Ratajczak-Wielgomas K, Dziegiel P. BCL11A expression in breast cancer. Curr Issues Mol Biol 2023;45(4):2681–98. http://dx.doi.org/10.3390/cimb45040175.
[25] Orrantia-Borunda E, Anchondo-Nuñez P, Acuña-Aguilar LE, Gómez-Valles FO, Ramírez-Valdespino CA. Subtypes of breast cancer. In: Breast cancer. Exon Publications; 2022, p. 31–42.
[26] Wang C, Uray IP, Mazumdar A, Mayer JA, Brown PH. SLC22A5/OCTN2 expression in breast cancer is induced by estrogen via a novel intronic estrogen-response element (ERE). Breast Cancer Res Treat 2012;134(1):101–15. http://dx.doi.org/10.1007/s10549-011-1925-0.
[27] Xu K, Usary J, Kousis PC, Prat A, Wang D-Y, Adams JR, Wang W, Loch AJ, Deng T, Zhao W, Cardiff RD, Yoon K, Gaiano N, Ling V, Beyene J, Zacksenhaus E, Gridley T, Leong WL, Guidos CJ, Perou CM, Egan SE. Lunatic fringe deficiency cooperates with the met/caveolin gene amplicon to induce basal-like breast cancer. Cancer Cell 2012;21(5):626–41. http://dx.doi.org/10.1016/j.ccr.2012.03.041.
[28] Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. J Bioinform Comput Biol 2005;3(02):185–205.
[29] Kononenko I, Šimec E, Robnik-Šikonja M. Overcoming the myopia of inductive learning algorithms with RELIEFF. Appl Intell 1997;7(1):39–55.