A Local Non-Additive Framework for Explaining Black-Box Predictive Models

Majid Mohammadi^{a;*}, Ilaria Tiddi^a and Annette ten Teije^a

^aVrije Universiteit Amsterdam

Abstract. Understanding the rationale behind the predictions made by machine learning models holds paramount importance across numerous applications. Various explainable models have been developed to shed light on these predictions by assessing the individual contributions of features to the outcome of black-box models. However, existing methods often overlook the crucial aspect of interactions among features, restricting the explanation to isolated feature attributions. In this paper, we introduce a novel Choquet integral-based explainable method, termed ChoquEx, which not only considers the interactions among features but also enables the computation of contributions for any subset of features. To achieve this, we propose an innovative algorithm based on support vector regression that efficiently estimates the contributions of all feature subsets. Intriguingly, we leverage gametheoretic concepts, including Shapley values and interaction index, to calculate both the feature importance and interaction strength. This approach adds further interpretability and insight into the model's decision-making process. To evaluate the effectiveness of ChoquEx, we conduct extensive experiments on diverse real-world scenarios. Our results convincingly demonstrate the superiority of the proposed model over existing explainable techniques. With its ability to unravel feature interactions and furnish comprehensive explanations, ChoquEx significantly enhances our understanding of predictive models, opening new avenues for applying machine learning in critical domains that require transparent decision-making.

1 Introduction

The use of machine learning has witnessed remarkable growth across various domains of science and technology, showcasing promising predictive capabilities. Evaluating these predictions often involves multiple performance metrics, designed to quantify the distinctions in performance among different models. However, a drawback of this approach lies in the lack of comprehensible explanations for model predictions, leading to limited understanding by human users leveraging such models. This dearth of justification is particularly concerning in critical domains like healthcare, where decisions can directly impact the well-being and lives of individuals. Without transparent explanations for predictions, domain experts may hesitate to trust predictive models, potentially impeding their adoption.

To address this challenge, significant efforts have been dedicated to developing explainable methods for machine learning models (for a review, refer to [3, 19]). These methods can be broadly categorized into model-specific, providing explanations for a specific machine learning model (e.g., random forest), and model-agnostic, offering explanations for any trained model. Explanations can be either global, revealing insights into the overall functioning of the predictive model, or local, focusing on explaining predictions for individual samples.

Local explainable methods often resort to assigning contributions to individual features, typically overlooking the interactions among features [28, 21]. Some approaches attempt to capture interactions, but due to the exponential growth of model parameters, they are restricted to 2-way interactions [21, 24, 30, 23, 27, 32]. In contrast, black-box models like deep neural networks exhibit remarkable performance due to their ability to model complex patterns, including feature interactions.

This paper introduces the concept of non-additive explanations, leveraging the Choquet integral [6], a nonlinear aggregation function widely used in decision theory [22]. Building upon this notion, we develop a local explainable model capable of computing contributions for any subset of features. A significant challenge in computing contributions for all subsets arises from the exponential increase in parameters that need estimation concerning the number of features. To address this, we propose an algorithm based on support vector regression, where the problem dimension scales proportionately to the number of samples (and **not** features), enabling efficient discovery of feature interactions.

The key contributions of this paper are as follows:

- Introduction of non-additive explanations using the Choquet integral, presenting an explainable yet nonlinear model, distinguishing it from conventional approaches that rely on linear surrogate models.
- Development of an efficient algorithm based on support vector regression to estimate the parameters of the Choquet integral, required for computing contributions of each feature subset. The parameter estimation involves solving a number of linear equations, with the equation count being proportional to the number of samples used for explanation.
- Calculation of feature importance and interactions using gametheoretic notions, such as Shapley importance and interaction index
 [2]. It is important to note that unlike other explainable methods, we do not estimate feature importance solely based on the Shapley value; instead, we estimate the parameters essential for computing such values.

The structure of this paper is as follows: Section 2 discusses related works, while Section 3 introduces the fundamental concept of the Choquet integral. Section 4 presents the concept of non-additive explanations based on the Choquet integral, along with the efficient algorithm for parameter estimation using support vector regression. In Section 5, we demonstrate the calculation of individual feature

^{*} Corresponding Author. Email: majid.mohammadi690@gmail.com

importance and interaction values using Shapley value and interaction index. Section 6 provides details on the experiments conducted to validate the proposed method, and finally, Section 8 concludes the paper.

2 Related Work

There are many methods available for explaining black-box models. LIME [28] provides explanations by learning a linear model at the neighborhood of the sample under explanation and uses the coefficients of the model as an explanation. SHAP [21] estimates the feature importance based on Shapley value, whose model-agnostic version, known as *Kernel SHAP*, estimates the feature importance by using a weighted linear model [7]. There are other variations of SHAP, for instance, for tree-based models [20] or for global explanations [8].

There are several explainable methods that relax the feature independence assumption, imposed by many methods including SHAP and LIME. In [15], authors take into account the causal structure of the data (if available) and then estimate the feature importance accordingly. In particular, they use the causal structure to ignore some of the generated samples for explanations. Another model for considering the causal structure is developed by [12], where the weight in the Shapley value formula is replaced based on the given causal structure. Also, the work in [1] considers the correlation between features and provides a new formulation to estimate the feature importance by the Shapley value.

All of these methods estimate only the importance of individual features and provide limited information regarding the feature interaction. Identifying feature interactions has recently drawn much attention [30, 32, 23, 24, 33, 9]. In [30], authors employ an axiomatic approach and put forward a new interaction index that can be used to compute the interaction between features. A general axiomatic framework for interaction detection is developed in [32], according to which the interactions among features are computed based on mixed partial derivatives. The work in [9] estimates the global feature interactions by using a Bayesian neural network, but computations are limited to binary interactions only. Estimating binary interactions using the Shapley interaction index for the neural network is also studied in [33]. Contextual importance [24] is an explainable method for LSTM [16] and can estimate the interactions among consecutive features. Generally, the explainable methods with feature interactions are typically restricted to 2-way interaction only, as higher-order interactions increase exponentially the number of parameters.

The Choquet integral has been successfully applied as an aggregation operator in many settings [6, 22]. Recently, the use of Choquet integral in machine has also been investigated [18, 11, 31, 17]. A ridge regression model based on the Choquet integral is developed in [18], while a nonlinear monotone logistic regression using the Choquet integral is developed in [11]. The theoretical behavior of the Choquet integral in the machine learning settings is studied in [17, 31], where it is shown that the VC-dimension of models based on Choquet integral ($\approx 2^d$, d is the number of features) is much higher than that of linear models, yet the Choquet integral remains explainable.

3 Choquet Integral and Mobius Transformation

This section reviews the basic notions of the non-additive measures and Choquet integral that will be used to present the non-additive explanation.

Let $\{x_i, y_i\}_{i=1}^n$ be *n* data points and their corresponding label, $x_i \in \mathbb{R}^d$, and $F = \{f_1, f_2, ..., f_d\}$ be the finite set of features. The

measure μ is a set function that maps any subset to a real number, so $\mu : \Sigma \to \mathbf{R}$, where Σ is the power set and \mathbf{R} is the real space. If μ is additive, then $\mu(A \cup B) = \mu(A) + \mu(B)$ for all $A, B \subset F$. The linear function typically used in different machine learning models could be seen as a result of an additive assumption on the features, i.e.,

$$w^{T}x_{i} = \sum_{j=1}^{d} w_{j}x_{ij} = \sum_{j=1}^{d} \mu(\{f_{j}\})x_{ij},$$
(1)

where w is the coefficient vector. Equation (1) is typically used for explainable methods, where $\mu(\{f_i\})$ is interpreted as the importance of the corresponding features. Also, this equation ignores the interaction between the features that could possibly influence the prediction as well. In order to consider the interactions, we need a more generalized measure, referred to as *non-additive* measures, that has the following properties [29]:

$$\mu(\emptyset) = 0$$

$$\mu(A) \le \mu(B) \quad \text{for all } A \subseteq B \subseteq F.$$
(2)

The last property in equation (2) refers to as monotonicity of the measure, but other non-monotone measures are also put forward [26]. Also, measure μ is called k-additive if k is the smallest integer where $\mu(A) = 0$ for each $A \subseteq F$ and |A| > k.

Given the non-additive measure, we require a proper function to aggregate the feature values and their associated importance and interactions. The Choquet integral is a satisfactory manner to build such an aggregation function [6, 14]. The departure point for the Choquet integral is that it is an alternative way to compute the area under a curve, where it decomposes the area under the curve horizontally instead of vertically. Figure 1 shows the intuition behind the Choquet integral. The panel on the left shows the linear model over three features, which is computed as the sum of the product of their width (i.e., weight) to their corresponding height (i.e., feature value). This computation could be rewritten as in the middle panel, where the vertical rectangles are replaced by horizontal ones, which provide the same value as the one on the left panel. However, this representation would allow us to incorporate the interactions of features, as shown by an exemplary case in the right panel of Figure 1. Intuitively, in the right panel, we see that the measure $\mu(x_1, x_2, x_3)$ is more than the sum of weights in the other panels, highlighting a positive interaction between the three features. In contrast, the middle horizontal rectangle in the right panel suggests that $\mu(x_2, x_3)$ is less than the sum of their weight, implying a negative interaction, or redundancy, between them.

That being said, the Choquet integral of x with respect to a nonadditive measure μ , shown by $C_{\mu}(x)$, is defined as [6]:

$$C_{\mu}(x) = \sum_{j=1}^{d} \left(x_{(j)} - x_{(j-1)} \right) \mu(F_{(j)}), \tag{3}$$

where (.) is a permutation of features such that $0 \le x_{(1)} \le x_{(2)} \le \dots \le x_{(d)}, x_{(0)} = 0$ by definition, and $F_{(j)} = \{f_{(j)}, f_{(j+1)}, \dots, f_{(d)}\}$. For the additive case, one can write:

$$C_{\mu}(x) = \sum_{j=1}^{d} \left(x_{(j)} - x_{(j-1)} \right) \mu(F_{(j)})$$

= $\sum_{j=1}^{d} \left(x_{(j)} - x_{(j-1)} \right) \left(\mu(\{f_{(j)}\}) + \dots + \mu(\{f_{(d)}\}) \right)$
= $\sum_{j=1}^{d} x_{j} \mu(\{f_{j}\}),$ (4)



Figure 1: An example of the Choquet integral.

which is the weighted sum, identical to equation (1).

The Choquet integral as in equation (3) requires ordering of x, which might not be efficient for practical purposes. A representation of non-additive measures, that can be used for computing the Choquet integral, is the *Mobius transformation*, defined as [14]:

$$\mu(B) = \sum_{A \subseteq B} m_{\mu}(A), \tag{5}$$

where the Mobius transform m_{μ} is defined as follows:

$$m_{\mu}(A) = \sum_{C \subseteq A} (-1)^{|A| - |C|} \mu(C).$$
(6)

Thus, there is a one-to-one mapping between measure μ and its Mobius transformation m_{μ} . The Choquet integral with respect to the Mobius transform could be written as [11]:

$$C_{\mu}(x) = \sum_{f \subseteq F} m_{\mu}(f) \times \min_{i \in f} x_i.$$
(7)

Equation (7) obviates the need of ordering x for computing the Choquet integral, but we need to find the minimum values in a set.

4 A Local Non-additive Nonlinear Explainable Model

In this section, we first present the notion of the non-additive explanation based on the Choquet integral and then develop an efficient algorithm based on support vector regression to estimate the parameters in the Choquet integral (i.e., the contributions of any subset of features).

4.1 Non-additive explanation

Let f be the trained model and g be a local surrogate model for the explanation. We seek to explain the prediction f(x) for a single input x. For doing so, explanation methods often use a simplified version of x, shown here by x', where the original data could be estimated by the mapping $x = h_x(x')$. The simplified version of the sample is different for various types of inputs. For text classification, a simplified input is a binary vector showing the presence or absence of a word, or for image classification, it is a binary vector showing the presence or absence of super-pixels (i.e., a batch of similar pixels in a neighborhood). In any case, the local methods try to build a model like g whose prediction for x' matches the original prediction, i.e., $g(x') = f(h_x(x'))$. We show the original sample by $x \in \mathbb{R}^d$, the binary vector for the interpretation by $x' \in \{0, 1\}^{d'}$, and the perturbed samples from x' by z'. The following definition provides a basis for the non-additive explanation model.

Definition 1 (Non-additive Explanation) A non-additive explanation model is a nonlinear function of binary variables, defined as:

$$g(z') = \mu_0 + C_\mu(z'), \tag{8}$$

where $z' \in \{0, 1\}^{d'}$, $\mu_0 \in R$, and $\mu \in R^{2^{d'}-1}$, and d' is the number of input features to the explainable method.

According to Definition 1, we need to identify the parameters of the non-additive measure in the Choquet integral. These parameters identify, not only the attribution of each individual feature but also the contribution of any subset of features.

The non-additive explanation as in Definition 1 could also be written based on the Mobius transformation:

$$g(z') = \mu_0 + \sum_{i \subseteq \{1, \dots, d'\}} m_\mu(F_{\{i\}}) \times \min_{j \in i} z'_j, \tag{9}$$

where $F_{\{i\}} = \{f_j | j \in i\}$ is the Mobius coefficient for set *i*. To simplify the notation, let $I = \{1, ..., d'\}$ and denote its power set by Σ_I . Also, let $u : \Sigma_I \to N$ be a one-to-one function that gets a subset from *I* and returns a unique natural number less than $2^{d'} - 1$. For further simplification, we define $m = [m_1, m_2, ..., m_{2d'-1}]$ where $m_q = m_{\mu}(F_{\{u^{-1}(q)\}}) (u^{-1})$ is the inverse of *u*) and write equation (9) as:

$$g(z') = \mu_0 + \sum_{q=1}^{2^{d'}-1} m_q \times \min_{j \in u^{-1}(q)} z'_j.$$
(10)

This equation implies an additive representation of a non-additive measure, but the additive representation is in a higher-dimensional space (i.e., $2^{d'} - 1$). We use this representation to estimate the parameters of the Choquet integral. Recall that the number of parameters in equation (9) is the same as that presented in Definition 1.

4.2 Explanation model

To provide an explanation for a single prediction, a popular strategy is to create a neighborhood around the sample under the explanation [28, 7]. This is typically done by perturbing the corresponding binary vector z' and generating new samples around it. The influence of each generated sample on building the explanation model g could be proportionate to its distance to the original input z'. The proportionality is taken into account by using proper loss functions weighted on the distances of the samples. We show the distance function for the explaining instance x' is $\gamma_{x'}$.

Another important aspect of the explanation, also discussed in [28, 21], is the complexity of the provided explanation, as the complex explanation (like linear models with many non-zero coefficients or a deep decision tree) could not be essentially interpretable. This is why the complexity of the explanation model should be controlled. As a result, the following model is used for building an explanation model for x' [28, 21]:

$$\min_{g} L(f, g, \gamma_{x'}) + \Omega(g), \tag{11}$$

where $\Omega(g)$ is a function controlling the model complexity of g. LIME [28] entails the distances to the sample under explanation and the complexity of the explanation model by using the weighted lasso for estimating a linear g, where the loss is the weighted least square and Ω is the ℓ_1 regularization. Kernel SHAP (i.e., the model-agnostic version of SHAP) also uses a similar framework, but the weights for the model are computed based on the Shapley value formula [21, 7], and they include the ℓ_1 regularization in their implementation for controlling the complexity. Within this framework, we now develop a non-additive, nonlinear explainable method.

4.3 Support vector regression for estimating contributions

The non-additive explanation as presented in Definition 1 has many advantages. However, the number of parameters is still exponential to the number of input features. As a result, obtaining such an explanation could be very time-consuming.

We now develop an efficient algorithm based on support vector regression (SVR) that can estimate the non-additive measure parameters in the non-additive explanation. The reason that we think of the SVR is that there is an additive representation, as presented in equation (9), that can represent the non-additive measure in a higher-dimensional space. Intuitively, it resembles mapping to the feature space in the SVR, where it is assumed that the input data are linearly separable in a higher-dimensional space (e.g., infinite dimension for the radial basis function).

We then propose the following optimization model for the nonadditive explanation for x' based on the SVR:

$$\min_{\mu,\xi,b} \Omega(\mu) + C \sum_{i=1}^{n'} \gamma_i \xi_i^2$$

s.t. $C_{\mu}(z_i) + b - y_i = \xi_i, \quad i = 1, ..., n',$ (12)

where $\Omega(\mu)$ is the regularization controlling the complexity of the explanation model, b is the bias term, and C is the regularization parameters indicating the trade-off between the error and the regularization, n' is the number of samples generated for building explanation model, z_i represents a generated samples, and $\gamma_i = \gamma_{x'}(z_i)$. For purpose of solving this nonlinear problem, we use the Mobius representation

of the Choquet integral as in equation (9), as it does not require the ordering of values and can uniquely correspond to a non-additive measure. In order to do so, we assume that $F' = \{f'_1, ..., f'_{d'}\}$ is the input features to the explainable method, and then rewrite the Choquet integral for a sample $z \in \mathbb{R}^{d'}$ as:

$$C_{\mu}(z) = \mu_0 + \sum_{q=1}^{2^{d'}-1} m_q \times \min_{j \in u^{-1}(q)} z'_j = \sum_{q=1}^{2^{d'}-1} m_q \hat{z}_q = m^T \hat{z},$$

where $\hat{z} \in R^{2^{d'-1}}$ is the following vector:

$$\hat{z} = (z_1, z_2, ..., z_{d'}, \min\{z_1, z_2\}, \dots, \min\{z_{d'-1}, z_{d'}\},
\min\{z_1, z_2, z_3\}, \dots, \min\{z_1, ..., z_{d'}\}),$$
(13)

and *m* is a vector representing the Mobius coefficient of the corresponding set in \hat{z} . Using this equation, minimization (12) is rewritten based on the Mobius coefficient as:

$$\min_{m,\xi,b} \frac{1}{2} \|m\|_{2}^{2} + C \sum_{i=1}^{n'} \gamma_{i} \xi_{i}^{2}$$
s.t. $m^{T} \hat{z}_{i} + b - y_{i} = \xi_{i}, \quad i = 1, ..., n'.$ (14)

The Karush-Kuhn-Tucker (KKT) optimality conditions for this minimization can be summarized as the following linear system (see Appendix A for more detail):

$$\begin{bmatrix} K+Q & e \\ e^T & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ b \end{bmatrix} = \begin{bmatrix} y \\ 0 \end{bmatrix},$$
 (15)

where $K \in \mathbb{R}^{n' \times n'}$ is a matrix with elements $K_{ij} = \hat{z}_i^T \hat{z}_j$, Q is a diagonal matrix whose diagonal elements $Q_{ii} = 1/2C\gamma_i$, and e is a vector of one. The solution of the system is then calculated by:

$$b = \frac{e^T (K+Q)^{-1} y}{e^T (K+Q)^{-1} e} \quad \alpha = (K+Q)^{-1} (y-eb)$$

By calculating α based on the above equation systems, *m* is computed from the following equation (see Appendix A for more detail):

$$m = \sum_{i=1}^{n'} \alpha_i \hat{z}_i. \tag{16}$$

For applying the SVR, we need to calculate the matrix K according to \hat{z}_i 's. In the following theorem, we show that this matrix can be calculated very efficiently. See Appendix B for the proof.

Theorem 1 The elements of matrix K in equation (15) can be calculated as:

$$K_{ij} = \hat{z}_i^T \hat{z}_j = \sum_{j=1}^k \binom{q}{j},$$
 (17)

where q is the number of non-zero elements common in both z_i and z_j , and k denotes the k-additivity of the non-additive measure.

4.4 *Time Complexity*

The main advantage of using the SVR is that we only need to perform a few arithmetic operations, proportionate to the number of samples for explanation n', or alternatively, solve n' + 1 linear equation systems, where n' is independent of the number of input features d' and length of m (that is $2^{d'} - 1$). One way of obtaining the solution to the SVR problem is by computing the pseudoinverse of $n' + 1 \times n' + 1$ matrix (see Appendix A). The inverse is computed in $O(n'^3)$, and the solution to the SVR can be computed with a matrix-vector multiplication by n'^2 multiplications and summations. Also, for computing the coefficient m, we also need the $n'2^{d'}$ multiplications and summations. Note that the latter two operations are not iterative and we need to do the matrix-vector multiplications only once.

5 Importance and Interaction Indices

For an additive measure, the importance of a feature is basically identified by the corresponding coefficient. This implies that the influence of a feature like f_i is independent of the presence or absence of other features. For the non-additive measure, as used in the previous section, the importance of a feature is also influenced by the presence or absence of other features.

For a given non-additive measure μ over a set of features, a natural way to identify the importance of each feature is to look at the importance augmentation when a new feature like f_i is added to any subset $A \in F$. Thus, we need to compare $\mu(A \cup f_i)$ and $\mu(A)$ for all $A \subseteq F \setminus f_i$. The Shapley value, also called the importance index, is basically the average of the difference $\mu(A \cup f_i)$ and $\mu(A)$ for all possible subsets A:

$$\phi(f_i) = \sum_{A \subseteq F \setminus f_i} \frac{|A|!(n-|A|-1)!}{n!} \big(\mu(A \cup f_i) - \mu(A) \big),$$

where $\phi(f_i)$ (or simply ϕ_i) is the importance of f_i . For the additive case, $\mu(A \cup f_i) - \mu(A) = \mu(f_i)$ so that the Shapley value also boils down to $\phi_i = \mu(f_i) = w_i$.

The feature importance in SHAP is merely an estimation of ϕ_i 's, but here we estimate measure μ that includes all the parameters required to compute the importance of features ϕ_i . An important byproduct of such an approach is that we can compute the interaction indices as well. The interaction index between two features f_i and f_j is also defined as [25]:

$$I_{i,j} = \sum_{A \subseteq F \setminus \{f_i, f_j\}} \frac{1}{(m-1)\binom{|A|}{m-2}} \left(\mu(A \cup \{f_i, f_j\}) - \mu(A \cup \{f_i\}) - \mu(A \cup \{f_j\}) + \mu(A) \right),$$
(18)

that basically averages the marginal contribution of features f_i and f_j to any possible subset of features (excluding f_i and f_j). Generally, the interaction index can be extended for any subset of features like $T \subseteq F$, and is defined from the Mobius coefficient m as [13]:

$$I_T = \sum_{B|T \subseteq B} \frac{1}{|B| - |T| + 1} m(B)$$
(19)

As a result, identifying the non-additive measure enables us to compute the importance and interaction indices.

The importance and interaction indices based on the Shapley value are only a class of indices that can be used here. It is also possible to use other indices; examples are the Banzhaf power index as an importance index [10], and the recently proposed Shapley-Taylor interaction index [30]. In any case, the proposed method, which identifies the parameters of a non-additive measure, could be used to compute such quantities. The interaction index only indicates the interaction of the features for a prediction and not their joint importance/influence on the predicted value. For explainability, however, the importance of a subset of features to the prediction is desired.

Definition 2 (Joint feature importance) *The importance of feature set* $A \subseteq F$ *is defined as:*

$$\phi(A) = \sum_{f_i \in A} \phi(f_i) + \sum_{T \subseteq A} I_T.$$
 (20)

For instance, the joint importance of two features i and j based on the above definition could be computed as:

$$\phi(\{f_i, f_j\}) = \phi(f_j) + \phi(f_i) + I_{i,j}.$$
(21)

So, for nonzero $I_{i,j}$, the joint importance of two features f_i and f_j could be higher or lower than the sum of the importance of each individual feature, resulting in a non-additive explanation.

6 Experiments

This section provides some experiments regarding the proposed method and compares them with state-of-the-art explainable methods. We first use synthetic data sets with known ground truth explanations that enable us to compare different explainable methods objectively. We also use three models to compare the explainable methods on the post-hoc accuracy and execution time. First, we train a bidirectional LSTM on the IMDB reviews for sentiment analysis. We set the window size to 256, and the dropout to 0.2, with a dense layer with a sigmoid activation function. Second, we train a random forest with 50 trees on the Boston housing dataset. Third, we use a real case study, where we develop a BERT-based deep neural network to classify the clinical trials of a medical company as worthy (to be included in the database) and non-worthy. See Appendices C and D for other experiments. The Python implementation of ChoquEx is publicly available¹.

6.1 Synthetic data

We first experiment on several synthetic data sets to compare objectively the efficiency of the proposed method. We create three synthetic data sets that are with known ground truth for explanation in order to quantify the goodness of different methods. The data sets are variations of commonly used data sets in explainability and feature selection [4, 5]. The first dataset is a 2-dimensional XOR function, acting as a binary classification. The data set X is 10-dimensional and generated according to the standard normal distribution, and the target Y is computed as $P(Y = 1|X) \propto exp(X_1X_2)$. Every sample in this data set has the first two elements as true features, which are individually independent of the response variable, but jointly affect the response. For the second data set, we set the target variable as $Y \propto exp(-20cos(2X_2) + 2|X_3| + X_1 + exp(-X_4))$, which is a nonlinear additive function with respect to X. As a result, the first four features for each sample in this data set are the important ones, on which the response variable is dependent. For the third data set, we set the target variable as $Y \propto exp(\sum_{i=1}^{3} X_i^2)$, thus only the first three features influence the response.

The three data sets were subjected to the random forest with a maximum depth of ten, and the explanations for 300 test samples are

¹ https://github.com/Majeed7/choquex



Figure 2: The box plot of explainable methods on three synthetic data sets.

generated with ChoquEx, LIME, Kernel SHAP, BivariateSHAP [23], and $L2X^{2}$ [4]. BivariateSHAP can detect the pairwise interactions between features, and L2X is able to detect the interactions of higher orders through the use of mutual information. Since the true underlying feature for each sample is known, we can check if the explainable methods have correctly identified them. The selected features by the explainable methods are ranked based on their identified importance and we compute the median rank of the true features for each test sample. Figure 2 shows the box plot of the median ranks for the explainable methods on the three synthesized data sets. For the first data set, ChoquEx and BivariateSHAP have the best performance and outperform other explainable methods significantly. This is due to the fact that the first data set includes interaction between two features, so BivariateSHAP can also detect the interactions properly. In the other two data sets, ChoquEx and L2X have the best performances since they can capture highly nonlinear relationships with higher-order interactions.

6.2 Local accuracy and execution time

The fidelity of the local explanations is an important yardstick to compare different explainable methods. For doing so, we compare the prediction of the local explainable model with that of the predictive model: The closer the two predictions are, the better the explainable model could capture the behavior of the predictive model. Yet another important criterion is the execution time for generating the explanations.

To conduct a comparison, we generate explanations for a number of samples from the IMDB reviews, the clinical trials, and the Boston housing dataset, and compare the explanations generated by ChoquEx (with k = 5 and k = d'), LIME, Kernel SHAP³, L2X, and BivariateSHAP. Table 1 shows the means square error (MSE) of the local explanations and the execution time for different explainable methods. According to Table 1, the proposed method has better performance in terms of MSE, since it is a nonlinear model as opposed to LIME and Kernel SHAP which uses linear regression for creating the explainable model. Specifically, the ChoquEx performance increases when we increase the order of interactions from 5 to d'. BivariateSHAP and L2X have also very competitive results, implying that the accommodation of feature interactions can lead to a better surrogate model for an explanation. Regarding the average execution time, ChoquEx is competitive with LIME, both superior to Kernel SHAP. This means that we can reliably identify feature importance and interactions with the proposed method in a reasonable time. L2X is also very fast as it learns the explanations of all samples at once (by

training a neural network for explanation). Bear also in mind that the average execution time for BERT and LSTM is much higher since we need to get the predictions of those models for the samples generated for an explanation, and the forward pass in the LSTM and BERT is time-wise expensive.

6.3 IMDB sentiment classification

From the LSTM model trained on the IMDB review dataset, we generated explanations for 100 reviews and found the interactions between different words. Table 2 shows some of the interesting interactions detected by ChoquEx. The redundancy is captured properly by ChoquEx: In the second example of Table 2, the words *hated* and *words* convey the same message about the negative sentiment of the review. A similar redundancy exists in terms *good* and *very funny*, as well as in *boring* and *not* ... *exciting* in the fourth and fifth example of Table 2, respectively. In addition, in the third example, the words *must see* has strongly positive interactions. This is aligned with the context of the review since both words together are a stronger indication of the positiveness of the review.

To demonstrate the non-additive explanations, we plot the interaction and importance explanation provided by ChoquEx on the first IMDB review example in Table 2. Figure 3a shows the feature importance and interaction index of the top three compounds in the given example. According to this plot, the term *not* pushes the prediction of the review to have a negative sentiment ($\phi(not) = -0.43$), while good forces the prediction in the other direction ($\phi(good) = 0.24$). There is also a negative interaction between the two terms: $I_{not,good} = -0.33$. Since such values only indicate how the two features interact with each other, not how they influence the prediction, we can calculate the joint feature importance as in Definition 2. Accordingly, the joint importance of *not* and *good* is calculated as:

$$\phi(not \lor good) = \phi(not) + \phi(good) + I_{not,good} = -0.52$$

Figure 3b plots the joint feature importance of the IMDB review example. This highlights the non-additive explanation provided by the proposed ChoquEx.

7 Broader Impact

The main purpose of this work is to provide more insight into the black-box machine learning models. A critical risk of using ChoquEx is how the interactions between different features are interpreted, which might be confused with the importance of interacting features. For instance, the negative interactions between two features are a sign of redundancy of the features involved, not that negatively impacting the prediction. Also, the user of the method should not rule out other explainable methods for their experiment. For instance, if there is a

 $^{^2}$ L2X needs to train a neural network for generating explanations; we use the same architecture of the original paper for our experiments.

³ We use Kernel SHAP because it is model-agnostic and can be applied to all models; even though it sub-samples the training set.

Table 1: The comparison of explainable methods based on mean square error (MSE) of the local explanation and the average execution time in seconds (rounded to the first integer).

Method	MSE			Average time (s)		
	LSTM	BERT	RF	LSTM	BERT	RF
LIME	$12e^{-2}$	$14e^{-2}$	$3e^{-3}$	123	214	2
Kernel SHAP	$11e^{-2}$	$13e^{-2}$	$4e^{-3}$	310	426	29
ChoquEx(5)	$8e^{-2}$	$10e^{-2}$	$4e^{-3}$	136	251	3
ChoquEx(d')	$6e^{-2}$	$2e^{-2}$	$4e^{-3}$	139	267	5
L2X	$10e^{-2}$	$6e^{-2}$	$2e^{-3}$	73	210	2
BivariateSHAP	$7e^{-2}$	$5e^{-2}$	$4e^{-3}$	120	272	12

Table 2: Some examples of word interaction from the IMDB reviews.

Reviews (bold words with interactions)	interaction index	
the movie is not good at all	-0.32	
Hated it with all my being Worst movie ever It was that bad	-0.28	
A great film in its genre the direction A must see	0.45	
This is a good film This is very funny	-0.15	
the movie was absolutely boring there was not a single exciting event	-0.3	



(b) The ChoquEx joint feature importance explanation.

Figure 3: The ChoquEx interaction and importance plots on an exemplary review (first review in Table 2) from the IMDB dataset.

need to generate neuron-level or pixel-level explanations for respectively neural networks and image classifications, other methods might better fit such needs. There are also other domain-specific societal impacts. For instance, we use ChoquEx to provide explanations for the experts who want to decide to include clinical trials in their database. If the explanations convince the experts not to include the clinical trial, then the patients would not be able to see the clinical trial, and consequently, miss a relevant clinical trial.

8 Conclusion

This paper presented the non-additive explanation for predictive models based on the Choquet integral. In the given framework, a parameter in the non-additive measure is assigned to indicate the contribution of each individual and any subset of features, which increases the number of parameters exponentially. Nonetheless, we develop a support vector-based method to estimate the parameters of the non-additive measure very efficiently, whose complexity is only dependent on the number of samples generated for an explanation. The proposed model is a *nonlinear* surrogate model for the local explanation, which is unique as the common approach is to use a linear surrogate model for explainability.

There are many avenues for future research. An essential research complementary to this study would be to develop a model based on Choquet integral to provide global explanations. A by-product of such an approach is to capture a holistic interaction structure among different features, that could be further used as input for local explainable methods. Other types of Choquet integral should also be studied for explanations. For instance, the level-dependent Choquet integral could be used to provide local and global explanations simultaneously, obviating the need to build a local surrogate model for every sample.

ChoquEx uses the SVR as a workaround for the exponential number of parameters resulting from adding the feature interactions, nonetheless, we need to compute the primal solution that includes $2^{d'}$ parameters. Given the size of the parameters, the number of operations in ChoquEx is not high (i.e., n' multiplication summations for each element). However, such operations could be quite time-consuming for large-scale problems. One way to deal with this issue is to restrict the order of interactions. Alternatively, the Shapley value and interaction index (or other indices) could be rewritten based on the solution to the dual problem of the SVR, so that we can directly compute the Shapley values from the solution to the dual SVR problem.

Also, another aspect to investigate is the sample-generating mechanism. First, the number of samples in ChoquEx is deemed as a hyperparameter, similar to LIME and SHAP. In addition, the samples generated for the explanations ignore the causal structure in the data. Besides, feature interactions induce a causality between the features and their associated interactions. Thus, studying both feature interactions as well as the causal structure is an interesting problem not only for explainability but also for the broader machine-learning community.

References

- [1] Kjersti Aas, Martin Jullum, and Anders Løland, 'Explaining individual predictions when features are dependent: More accurate approximations to shapley values', *Artificial Intelligence*, **298**, 103502, (2021).
- [2] Robert J Aumann and Lloyd S Shapley, *Values of non-atomic games*, Princeton University Press, 2015.
- [3] Nadia Burkart and Marco F Huber, 'A survey on the explainability of supervised machine learning', *Journal of Artificial Intelligence Research*, 70, 245–317, (2021).
- [4] Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan, 'Learning to explain: An information-theoretic perspective on model interpretation', in *International Conference on Machine Learning*, pp. 883–892. PMLR, (2018).
- [5] Jianbo Chen, Mitchell Stern, Martin J Wainwright, and Michael I Jordan, 'Kernel feature selection via conditional covariance minimization', *Advances in Neural Information Processing Systems*, **30**, (2017).
- [6] Gustave Choquet, 'Theory of capacities', in *Annales de l'institut Fourier*, volume 5, pp. 131–295, (1954).
- [7] Ian Covert and Su-In Lee, 'Improving kernelshap: Practical shapley value estimation using linear regression', in *International Conference* on Artificial Intelligence and Statistics, pp. 3457–3465. PMLR, (2021).
- [8] Ian Covert, Scott M Lundberg, and Su-In Lee, 'Understanding global feature contributions with additive importance measures', *Advances in Neural Information Processing Systems*, 33, 17212–17223, (2020).
- [9] Tianyu Cui, Pekka Marttinen, and Samuel Kaski, 'Learning global pairwise interactions with bayesian neural networks', arXiv preprint arXiv:1901.08361, (2019).
- [10] Pradeep Dubey and Lloyd S Shapley, 'Mathematical properties of the banzhaf power index', *Mathematics of Operations Research*, 4(2), 99– 131, (1979).
- [11] Ali Fallah Tehrani, Weiwei Cheng, Krzysztof Dembczyński, and Eyke Hüllermeier, 'Learning monotone nonlinear models using the choquet integral', *Machine Learning*, 89(1), 183–211, (2012).
- [12] Christopher Frye, Colin Rowat, and Ilya Feige, 'Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability', *Advances in Neural Information Processing Systems*, **33**, 1229–1239, (2020).
- [13] Michel Grabisch, 'K-order additive discrete fuzzy measures and their representation', *Fuzzy sets and systems*, 92(2), 167–189, (1997).
- [14] Michel Grabisch, Jean-Luc Marichal, Radko Mesiar, and Endre Pap, Aggregation functions, volume 127, Cambridge University Press, 2009.
- [15] Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen, 'Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models', *Advances in neural information processing* systems, **33**, 4778–4789, (2020).
- [16] Sepp Hochreiter and Jürgen Schmidhuber, 'Long short-term memory', *Neural computation*, 9(8), 1735–1780, (1997).
- [17] Eyke Hüllermeier and Ali Fallah Tehrani, 'On the vc-dimension of the choquet integral', in *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pp. 42–50. Springer, (2012).
- [18] Siva K Kakula, Anthony J Pinar, Timothy C Havens, and Derek T Anderson, 'Choquet integral ridge regression', in 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–8. IEEE, (2020).
- [19] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis, 'Explainable ai: A review of machine learning interpretability methods', *Entropy*, 23(1), 18, (2020).
- [20] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee, 'From local explanations to global understanding with explainable ai for trees', *Nature Machine Intelligence*, 2(1), 2522–5839, (2020).
- [21] Scott M Lundberg and Su-In Lee, 'A unified approach to interpreting model predictions', *Advances in neural information processing systems*, 30, (2017).
- [22] Ksenija Mandic, Vjekoslav Bobar, and Boris Delibašić, 'Modeling interactions among criteria in mcdm methods: a review', in *International Conference on Decision Support System Technology*, pp. 98–109. Springer, (2015).
- [23] Aria Masoomi, Davin Hill, Zhonghui Xu, Craig P Hersh, Edwin K Silverman, Peter J Castaldi, Stratis Ioannidis, and Jennifer Dy, 'Explanations of black-box models based on directional feature interactions', in *International Conference on Learning Representations*, (2021).

- [24] W James Murdoch, Peter J Liu, and Bin Yu, 'Beyond word importance: Contextual decomposition to extract interactions from lstms', *arXiv* preprint arXiv:1801.05453, (2018).
- [25] T. Murofushi and S. Soneda, 'Techniques for reading fuzzy measures (iii): interaction index', *Processings of the 9th fuzzy system symposium*, 693–696, (1993).
- [26] Toshiaki Murofushi, Michio Sugeno, and Motoya Machida, 'Nonmonotonic fuzzy measures and the choquet integral', *Fuzzy sets and Systems*, 64(1), 73–86, (1994).
- [27] Giovanni Rabitti and Emanuele Borgonovo, 'A shapley-owen index for interaction quantification', SIAM/ASA Journal on Uncertainty Quantification, 7(3), 1060–1075, (2019).
- [28] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, "why should i trust you?" explaining the predictions of any classifier', in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, (2016).
- [29] Michio Sugeno, 'Fuzzy measures and fuzzy integrals—a survey', in Readings in fuzzy sets for intelligent systems, 251–257, Elsevier, (1993).
- [30] Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal, 'The shapley taylor interaction index', in *International conference on machine learning*, pp. 9259–9268. PMLR, (2020).
- [31] Ali Fallah Tehrani, 'The choquet kernel on the use of regression problem', *Information Sciences*, 556, 256–272, (2021).
- [32] Michael Tsang, Sirisha Rambhatla, and Yan Liu, 'How does this interaction affect me? interpretable attribution for feature interactions', *Advances in neural information processing systems*, 33, 6147–6159, (2020).
- [33] Hao Zhang, Yichen Xie, Longjie Zheng, Die Zhang, and Quanshi Zhang, 'Interpreting multivariate shapley interactions in dnns', in *Proceedings* of the AAAI Conference on Artificial Intelligence, volume 35, pp. 10877– 10886, (2021).

A Finding the optimal to the SVR

To solve problem (14), we introduce the following Lagrangian problem:

$$L(m,\xi,b,\alpha) = \frac{1}{2} ||m||_{2}^{2} + C \sum_{i=1}^{n'} \gamma_{i} \xi_{i}^{2} - \sum_{i=1}^{n'} \alpha_{i} (m^{T} \hat{z}_{i} + b - y_{i} - \xi_{i}), \qquad (22)$$

where α is the Lagrangian multiplier. Following the Karush-Kuhn-Tucker (KKT) conditions, we have:

$$\frac{\partial L}{\partial m} = 0 \Rightarrow m = \sum_{i=1}^{n'} \alpha_i \hat{z}_i,$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum \alpha_i = 0,$$

$$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow \xi_i = \frac{\alpha_i}{2C\gamma_i}$$

$$\frac{\partial L}{\partial \alpha_i} = 0 \Rightarrow m^T \hat{z}_i + b - y_i - \xi_i = 0.$$
(23)

By combining the above equations, the solution to minimization (14) follows from solving the following linear equation systems:

$$\begin{bmatrix} K+Q & e \\ e^T & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ b \end{bmatrix} = \begin{bmatrix} y \\ 0 \end{bmatrix},$$
 (24)

where $K \in \mathbb{R}^{n' \times n'}$ is a matrix with elements $K_{ij} = \hat{z}_i^T \hat{z}_j$, Q is a diagonal matrix whose diagonal elements $Q_{ii} = 1/2C\gamma_i$, and e is a vector of one. The solution of the system is then calculated by:

$$b = \frac{e^T (K+Q)^{-1} y}{e^T (K+Q)^{-1} e} \quad \alpha = (K+Q)^{-1} (y-eb)$$

By calculating α based on the above equation systems, m is computed by the first equation in (23), and the non-additive measure μ by equation (5).

B Proof of Theorem 1

Proof 1 The inner product is invariant to identical permutations of the two vectors. Thus, we permute the vectors z_i and z_j in a way that the first q elements are one in both vectors:

$$x = [\underbrace{1, 1, \dots, 1}_{q}, \dots]$$
$$y = [\underbrace{1, 1, \dots, 1}_{q}, \dots]$$
(25)

where x and y are, respectively, shown the permuted z_i and z_j on interest of simplicity. It follows:

$$\langle \hat{z}_i, \hat{z}_j \rangle = x_1 y_1 + \ldots + x_q y_q + \min\{x_1, x_2\} \min\{y_1, y_2\} + \ldots + \min\{x_{q-k}, \ldots, x_q\} \min\{y_{q-k}, \ldots, y_q\},$$
(26)

where the other elements are zero since either x or y is zero in those components, making the associated min operation zero. The total sum in equation (26) is indeed the number of the remaining terms, which could be counted based on the number of elements in the min

function (selecting the number of parameters in min from the total of q non-zero elements):

$$\underbrace{\begin{pmatrix} q\\ 1 \end{pmatrix}}_{one \ element} + \underbrace{\begin{pmatrix} q\\ 2 \end{pmatrix}}_{two \ elements} + \ldots + \underbrace{\begin{pmatrix} q\\ k \end{pmatrix}}_{k \ elements} = \sum_{j=1}^{k} \binom{q}{j}, \qquad (27)$$

that this completes the proof.

C Experiment on the entity linking use case

Entity linking refers to assigning a unique identity to entities appearing in a free body of text. It is typical that the combination of some words in the text might lead to linking any of the words to a different entity. To replicate such interactions, consider the following text about the central train station of Amsterdam:

This is *Amsterdam Central Station*, the transit hub that integrates two bicycle parking, two bus stations, two tram stops, and the central railway station for the city in the province of North Holland, Netherlands.

It is clear that the words "Amsterdam", "Central", and "Station" would refer to different terms (e.g., 'Amsterdam' refers to the city), but the three terms together refer to the main train station in the city of Amsterdam. We create a synthesized classifier that returns one if a given piece of text contains all three words consecutively, and zero otherwise. The goal is to explain this classifier with LIME and ChoquEx. Figure 4a shows the explanation provided by LIME on the given example. As expected, LIME cannot provide a reliable explanation nor capture the interaction and importance of the words in the text. Figure 4b provides the explanations provided by ChoquEx on the given example. For the purpose of completeness, we demonstrate all the feature importance and interactions captured by the proposed method. The importance of the terms "Amsterdam", "Central", and "Station" are equivalent in the explanation, which is also expected in the given example as the three words together would lead to a prediction of one from the classifier. For the joint importance values, the interaction between the three terms is also correctly captured by ChoquEx. Less clear interaction is the binary joint importance between any two terms of the three, but it is indeed from the ternary interactions of the word that affects the 2-way interaction as calculated by the Shapley interaction index (see equation (19)).

D Image classification

We also applied ChoquEx to explain the prediction of an image classifier. For doing so, we use Inception V3 as an image classifier and applied ChoquEx to three images for generating explanations. Figure 5 shows the explanation generated by ChoquEx. The first column depicts the original image, the second shows the most important superpixels for the prediction, and the other columns demonstrate the superpixels with positive interactions identified by ChoquEx. It is evident that ChoquEx can identify the most important superpixels for the classification. In addition, it assigns positive values for the interactions of those superpixels.



Figure 4: The LIME and ChoquEx explanation for the synthesized example.

