ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240605

Explaining a Probabilistic Prediction on the Simplex with Shapley Compositions

Paul-Gauthier Noé[a], Miquel Perelló-Nieto[b], Jean-François Bonastre[a] and Peter Flach[b]

[a]Laboratoire Informatique d'Avignon, Avignon Université, France
[b]University of Bristol, United Kingdom
ORCID (Paul-Gauthier Noé): https://orcid.org/0000-0002-2304-9830, ORCID (Miquel Perelló-Nieto): https://orcid.org/0000-0001-8925-424X, ORCID (Jean-François Bonastre): https://orcid.org/0000-0001-7741-3346, ORCID (Peter Flach): https://orcid.org/0000-0001-6857-5810

Abstract. Originating in game theory, Shapley values are widely used for explaining a machine learning model's prediction by quantifying the contribution of each feature's value to the prediction. This requires a scalar prediction as in binary classification, whereas a multiclass probabilistic prediction is a discrete probability distribution, living on a multidimensional simplex. In such a multiclass setting the Shapley values are typically computed separately on each class in a one-vs-rest manner, ignoring the compositional nature of the output distribution. In this paper, we introduce *Shapley compositions* as a well-founded way to properly explain a multiclass probabilistic prediction, using the Aitchison geometry from compositional data analysis. We prove that the Shapley composition is the unique quantity satisfying linearity, symmetry and efficiency on the Aitchison simplex, extending the corresponding axiomatic properties of the standard Shapley value. We demonstrate this proper multiclass treatment in a range of scenarios.

1 Introduction

Many machine learning approaches are regarded as black-boxes, making them unreliable for real-life applications where the model's predictions need to be understood or explained. In recent years, the interest in more interpretable models and explainability methods has therefore increased in the machine learning literature [5, 18]. One group of approaches, known as *local explanation*, aims to measure the contribution of each input feature's value to the computation of the model's output. Shapley values are widely used for this purpose [28, 8], especially since the release of the SHAP toolkit [20][1].

Shapley values were introduced in cooperative game theory where a group of players work together to maximise a payoff. A set of Shapley values distributes the payoff over all the players according to their individual contribution to the total. The Shapley value is the unique quantity that satisfies a set of desired axiomatic properties [26]. For explaining a machine learning model's prediction, features are treated as players and the scalar output of the model as the total payoff.

The Shapley value is designed for a one-dimensional function's codomain. In game theory, the characteristic function takes a coalition of players and gives a payoff. In machine learning, for a given instance, the characteristic function takes a group of features and gives a scalar output measuring how the prediction changes when the values of the features are considered. For a two-class probabilistic classification, the prediction is essentially a scalar since the probabilities for the two classes sum to one. Therefore, the Shapley value framework can simply be applied to the logit transform of one of the probabilities[2].

For more than two possible classes the output of the model is a discrete probability distribution or the output of a softmax function as commonly used by neural networks. Hence the output lives on a $(D-1)$-dimensional simplex, where D is the number of classes. In this case, the Shapley value framework cannot be directly applied. Of course, one can compute Shapley values on each output probability separately, but this ignores the structure of the simplex where the relative values between the probabilities is what really matters, rather than the absolute value of a single probability.

This paper presents *Shapley composition* as an extension of the Shapley value to the space of discrete probability distributions, using the *Aitchison geometry of the simplex* from the field of compositional

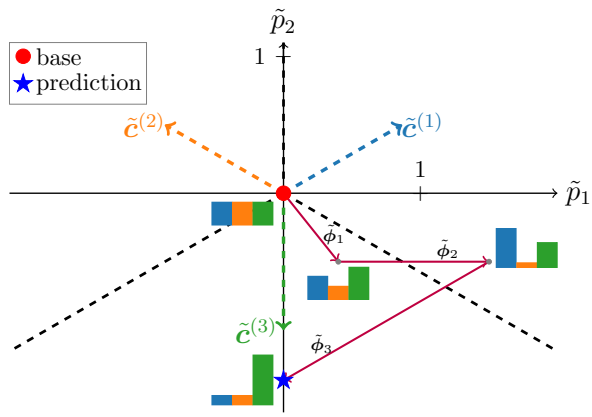

Figure 1. A synthetic example of a Shapley composition-based explanation. This shows a 2-dimensional space isomorphic to the 3-class-simplex where each point is a probability distribution as visualised with the histograms. The dashed black rays separate the maximum probability regions for each class and the dashed coloured vectors show the direction in favour of one class and against the other two. The space is additive such that the features' contributions $\{\tilde{\phi}_i\}_{\{1,2,3\}}$ translate the base distribution to the prediction.

[1] https://shap.readthedocs.io/en/latest

[2] The logit maps the domain $]0,1[$ to the additive real line $\mathbb{R}$.

data analysis. Compositional data [2, 23] are vectors – known as compositions – living on a simplex (not necessarily a probability simplex). Compositional data analysis has been applied to geological and chemical data, for example, but also to discrete probability distributions [10, 11, 21]. In the present paper, the probability distributions given by a classifier will be treated as compositional data in order to extend the Shapley value to multiclass classification.

Figure 1 shows a synthetic example to provide some intuitions. It shows a 2-dimensional space isomorphic to the 3-class-simplex where each point is a probability distribution also visualised as a histogram. The maximum probability regions, for each class, are clearly visible and separated by dashed black rays. Importantly, this space of probability distribution is additive thanks to the Aitchison geometry. The vectors show how the contribution of each feature changes – in an additive manner – the probability distribution (the ordering of the features can be chosen freely but the final point is fixed).

In the example, the base distribution (the average prediction over all data[3]) is modified by the contribution $\tilde{\phi}_1$ of the first feature. This goes mostly against class 2 such that the resulting distribution has the lowest probability for class 2. The angle between $\tilde{\phi}_1$ and the class-3 direction being the lowest among the classes, the resulting distribution has the highest probability for this class. The second feature moves the distribution into the class-1 region, perpendicular to the class-3 direction. The probability for class 1 is now maximum by reducing the probability for class 2, keeping the relative weight for class 3 unchanged. The third feature moves the distribution away from class 1. The resulting distribution being on the class-3 direction, the probability is maximum for class 3 and uniform for the other two.

We fully formalise the approach in this paper, making the following contributions:

- We define *Shapley composition* as a principled multidimensional extension of the Shapley value to the probability simplex,
- We prove that the Shapley composition is the unique quantity satisfying the set of desired properties known as *linearity*, *symmetry* and *efficiency* on the simplex equipped with the Aitchison geometry,
- We demonstrate the advantages of Shapley compositions for explaining a multiclass probabilistic prediction in machine learning.

The paper is structured as follows. Section 2 briefly reviews related work. Section 3 recalls the standard definition of the Shapley value and its use in binary classification. Section 4 presents the necessary tools from compositional data analysis: in particular, the Aitchison geometry of the simplex and the isometric log-ratio transformation. Section 5 defines the Shapley composition as an extension of the Shapley value framework to the multidimensional simplex using the Aitchison geometry. Section 6 shows with intuitive examples and visualisations how Shapley compositions can be used for explaining multiclass probabilistic predictions. Section 7 provides a short discussion and concludes the paper[4].

2 Related work

There is a plethora of methods in the literature to explain and better understand predictive models. They focus on different aspects of the task, from possible dataset biases, the feature importance with respect to the target, the parameters of a model after training, or the model predictions [27]. Some methods explain the influence of features on the model's performance: e.g., Permutation Feature Importance for random forest [7], which was later extended to the model agnostic Model Reliance [13]. Other methods focus on how individual features influence the model's predictions: e.g., Local Interpretable Model-agnostic Explanations (LIME) [25], Individual Conditional Explanation [16], Partial Dependence-based Feature Importance [17], Marginal Effect [6], Accumulated Local Effect [6], and Shapley value-based approaches [28, 8, 20].

We base our work on the Shapley value framework as one of the most well-founded feature influence methods. Inherently two-class, it has been applied to multiclass problems by explaining the influence of the features in a one-vs-one or one-vs-rest manner [32, 19], hence losing information that can be obtained by properly considering the full distribution. Utkin et al. [29, 30] explicitly consider the classifier output as a probability distribution, and measure the change in prediction in terms of statistical distance or divergence rather than in terms of difference between scalar predictions. However, even if this approach can measure the strength of a feature's value effect, it loses its directional information.

A recent work presented by Franceschi et al. [14] (later extended [15]) introduces stochastic characteristic functions to deal with models that output a random variable. With a categorical random variable, their approach can be used for explaining a multiclass classifier by allowing probabilistic statements about the likelihood of a feature to flip the decision from one class to another. In contrast, the approach we propose does not require an additional stochastic process but does not permit such a probabilistic statement. Instead, our approach is geometrical, by measuring how a feature moves the prediction on the probability simplex. In this way, it constitutes a natural extension of the standard Shapley value to the simplex for multiclass applications.

3 The Shapley value in machine learning

This section briefly recalls Shapley values as used for explaining features' contribution on a scalar prediction in machine learning. Let $f : \mathcal{X} \to \mathbb{R}$ be a learned model one wants to *locally* explain where $f(\boldsymbol{x})$ is the prediction on the instance $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d$. Let Pr be the probability distribution of the data over $\mathcal{X}$ (usually unknown but approximated by empirical averages). Let $S \subseteq \mathcal{I} = \{1, 2, \dots d\}$ be a subset of indices where d is the number of features. $\boldsymbol{x}_S$ refers to an instance $\boldsymbol{x}$ restricted to the features with indices in S.

When an instance $\boldsymbol{x}$ is observed, the expected value of the prediction is simply $\mathbb{E}[f(\boldsymbol{X}) \mid \boldsymbol{x}] = f(\boldsymbol{x})$. However, when only $\boldsymbol{x}_S$ is given with $\mathcal{S} \neq \mathcal{I}$, there is uncertainty about the non-observed features and the expected prediction given $\boldsymbol{x}_S$ is computed as $\mathbb{E}_{\text{Pr}}[f(\boldsymbol{X}) \mid \boldsymbol{x}_S] = \int_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}) \text{Pr}(\boldsymbol{x} \mid \boldsymbol{x}_S) d\boldsymbol{x}$. The change in prediction when the values of the features indexed by S are observed is measured by the characteristic function:

$$\begin{aligned} v_{f,\boldsymbol{x},\text{Pr}} : 2^{\mathcal{I}} &\to \mathbb{R}, \\ S &\mapsto \mathbb{E}_{\text{Pr}}[f(\boldsymbol{X}) \mid \boldsymbol{x}_S] - \mathbb{E}_{\text{Pr}}[f(\boldsymbol{X})], \end{aligned} \tag{1}$$

where $2^{\mathcal{I}}$ is the set of all subsets of $\mathcal{I}$. The contribution of the feature indexed by $i \notin S$ to the prediction, given the known values for the features indexed by S, is given by:

$$c_{f,\boldsymbol{x},\text{Pr}}(i, S) = v_{f,\boldsymbol{x},\text{Pr}}(S \cup \{i\}) - v_{f,\boldsymbol{x},\text{Pr}}(S). \tag{2}$$

The total contribution of the ith feature is computed by averaging this quantity over all possible coalitions S as follows:

$$\phi_i\left(f, \boldsymbol{x}, \text{Pr}\right) = \frac{1}{d!} \sum_{\pi} c_{f,\boldsymbol{x},\text{Pr}}(i, \pi^{<i}), \tag{3}$$

[3] Note that this does not need to be the uniform distribution, i.e., the origin.

[4] The code and Jupyter notebooks are available on the github page: https://github.com/shapley-composition.

where π is a permutation of the set $\mathcal{I}$ of indexes and $\pi^{<i}$ is the set of indexes before i in the ordering given by π. For better clarity, "$f, \boldsymbol{x}, \Pr$" or simply "$\boldsymbol{x}, \Pr$" will be dropped from the equations.

This quantity is known as the Shapley value for the ith feature. It originates from cooperative game theory and is the unique quantity respecting a set of desired axiomatic properties [26, 28]:

Linearity with respect to the model:
$\alpha, \beta \in \mathbb{R}, \forall i \in \mathcal{I},\ \phi_i(\alpha f + \beta g) = \alpha\phi_i(f) + \beta\phi_i(g)$;
Symmetry:
$\forall S \subseteq \mathcal{I} \backslash \{i, j\}, v(S \cup \{i\}) = v(S \cup \{j\}) \Rightarrow \phi_i = \phi_j$;
Efficiency: The "centered" prediction is additively separable with respect to the Shapley values:

$$f(\boldsymbol{x}) - \mathbb{E}_{\Pr}[f(\boldsymbol{X})] = \sum_{i=1}^{d} \phi_i(f, \boldsymbol{x}, \Pr). \tag{4}$$

Efficiency ensures that the change in prediction when the features are observed is distributed among them. In other words, the cumulative sum of the Shapley values moves the averaged prediction (also called *base* prediction) to the actual one.

The Shapley value is designed for a characteristic function with a scalar codomain. For explaining machine learning models which output multidimensional discrete probability distributions, like in multiclass classification, one could explain each output dimension separately, resulting in a one-vs-rest comparison. However, this approach ignores the relative information between each probability and ignores the compositional nature of the discrete probability distributions. Indeed, the probabilistic output of a classifier lives on a multidimensional simplex. The latter is the sample space of *compositional data* briefly reviewed in the next section.

4 Compositional data

Compositional data carries relative information. Each element of a composition *describes a part of some whole* [23], such as vectors of proportions, concentrations, and discrete probability distributions. An D-part composition is a vector of D non-zero positive real numbers that sum to a constant k. Each element of the vector is a part of the *whole* k. The sample space of compositional data is the $(D-1)$-dimensional simplex:

$$\mathcal{S}^D = \left\{ \boldsymbol{x} = [x_1, x_2, \dots x_D]^T \in \mathbb{R}_+^{*D} \mid \sum_{i=1}^{D} x_i = k \right\}. \tag{5}$$

In a composition, only the relative information between parts matters and John Aitchison introduced the use of log-ratios of parts to handle this [2]. He defined several operations on the simplex which leads to what is called the *Aitchison geometry of the simplex*.

4.1 The Aitchison geometry of the simplex

Only the relative information between the parts of a composition matters. Compositions are therefore scale-invariant. This is materialised by the closure operator defined for $k > 0$ as:

$$\mathcal{C}(\boldsymbol{x}) = \left[\frac{kx_1}{\|\boldsymbol{x}\|_1}, \frac{kx_2}{\|\boldsymbol{x}\|_1}, \dots \frac{kx_D}{\|\boldsymbol{x}\|_1} \right]^T \in \mathcal{S}^D, \tag{6}$$

where $\boldsymbol{x} \in \mathbb{R}_+^{*D}$ and $\|\boldsymbol{x}\|_1 = \sum_{i=1}^{D} |x_i|$. Aitchison defined on the simplex the following three operations [3]:

Perturbation: $\boldsymbol{x} \oplus \boldsymbol{y} = \mathcal{C}([x_1 y_1, \dots x_D y_D])$, seen as an addition between two compositions $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}^D$;
Powering: $\alpha \odot \boldsymbol{x} = \mathcal{C}([x_1^\alpha, \dots x_D^\alpha])$, seen as a multiplication by a scalar $\alpha \in \mathbb{R}$;
Inner product:

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{A}} = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \log \frac{x_i}{x_j} \log \frac{y_i}{y_j}. \tag{7}$$

In this paper, since we are interested in classification problems where the set of classes represents a set of exhaustive and mutually exclusive hypotheses, the output of a probabilistic classifier is a discrete probability distribution over the set of classes. We therefore restrict ourselves to the *probability simplex* where $k = 1$.

4.2 The isometric log-ratio transformation

A $(D-1)$-dimensional orthonormal basis of the simplex, referred to as an *Aitchison* orthonormal basis, can be built. The projection of a composition into this basis defines an isometric isomorphism between $\mathcal{S}^D$ and $\mathbb{R}^{D-1}$. This is known as an isometric log-ratio (ILR) transformation [12] and allows to express a composition into a Cartesian coordinate system preserving the metric of the Aitchison geometry. Within this real space, the perturbation, the powering and the Aitchison inner product are respectively the standard addition between two vectors, the multiplication of a vector by a scalar, and the standard inner product.

Given a composition $\boldsymbol{p} = [p_1, \dots p_D]^T \in \mathcal{S}^D$ we write its ILR transformation as $\tilde{\boldsymbol{p}} = \operatorname{ilr}(\boldsymbol{p}) = [\tilde{p}_1, \dots \tilde{p}_{D-1}]^T \in \mathbb{R}^{D-1}$. The ith element $\tilde{p}_i$ of $\tilde{\boldsymbol{p}}$ is obtained as: $\tilde{p}_i = \langle \boldsymbol{p}, \boldsymbol{e}^{(i)} \rangle_{\mathcal{A}}$ where the set $\{\boldsymbol{e}^{(i)} \in \mathcal{S}^D\}_{1 \leq i \leq D-1}$ forms an *Aitchison* orthonormal basis of the simplex. The basis can be obtain through the Gram-Schmidt procedure or by building a sequential binary partition [12, 9]. Examples are discussed in Section 6.2.

In the introductory example of Figure 1, the 2-dimensional ILR space isomorphic to the 3-class probability simplex was constructed as follows:

$$\tilde{p}_1 = \frac{1}{\sqrt{2}} \log \frac{p_1}{p_2}, \quad \tilde{p}_2 = \sqrt{\frac{2}{3}} \log \frac{\sqrt{p_1 p_2}}{p_3}.$$

Hence, the x-axis compares the probabilities for classes 1 and 2, the y-axis compares the probability for class 3 with the geometric mean of p_1 and p_2, and the origin corresponds to the uniform distribution, i.e., the neutral element for the perturbation. Note that the perturbation can be seen as a Bayesian update: the perturbation of a prior by a likelihood function gives the posterior. In the space of isometric log-ratio transformed distributions, the Bayes update is a vector translation.

5 Shapley composition on the simplex

In this section we will use the Aitchison geometry to extend the Shapley value from Section 3 to the simplex for explaining a multiclass probabilistic prediction. Let $\boldsymbol{f} : \mathcal{X} \to \mathcal{S}^D$ be a learned model which outputs a prediction on the $(D-1)$-dimensional probability simplex $\mathcal{S}^D$. In order to properly consider the structure of the simplex and the relative information between the probabilities, the model's output is treated as compositional data using the operators from the Aitchison geometry of the simplex. We therefore rewrite the characteristic function and the contribution of Equations 1 and 2 as follows:

$$\begin{aligned} \boldsymbol{v}_{\boldsymbol{f},\boldsymbol{x},\Pr} &: 2^{\mathcal{I}} \to \mathcal{S}^D, \\ & S \mapsto \mathbb{E}^{\mathcal{A}}_{\Pr}[\boldsymbol{f}(\boldsymbol{X}) \mid \boldsymbol{x}_S] \ominus \mathbb{E}^{\mathcal{A}}_{\Pr}[\boldsymbol{f}(\boldsymbol{X})]. \end{aligned} \tag{8}$$

$$\boldsymbol{c}_{\boldsymbol{f},\boldsymbol{x},\Pr}(i,S) = \boldsymbol{v}_{\boldsymbol{f},\boldsymbol{x},\Pr}(S\cup\{i\}) \ominus \boldsymbol{v}_{\boldsymbol{f},\boldsymbol{x},\Pr}(S), \tag{9}$$

where $\boldsymbol{a} \ominus \boldsymbol{b}$ is the perturbation $\boldsymbol{a} \oplus ((-1)\odot \boldsymbol{b})$ which corresponds to a subtraction between two compositions, and where the $\mathcal{A}$ in superscript highlights the fact that the expectation is taken with respect to the Aitchison measure. This can be computed as: $\mathbb{E}^{\mathcal{A}}[\boldsymbol{Y}] = \text{ilr}^{-1}(\mathbb{E}[\text{ilr}(\boldsymbol{Y})])$, where $\mathbb{E}^{\mathcal{A}}$ refers to the expectation with respect to the Aitchison measure while $\mathbb{E}$ refers to the expectation with respect to the Lebesgue measure [23].

The Shapley quantity expressing the contribution of the ith feature's value on a prediction can be expressed on the simplex as the composition $\boldsymbol{\phi}_i$ given by:

$$\boldsymbol{\phi}_i(\boldsymbol{f},\boldsymbol{x},\Pr) = \frac{1}{d!}\odot \bigoplus_{\pi} \boldsymbol{c}_{\boldsymbol{f},\boldsymbol{x},\Pr}(i,\pi^{<i}). \tag{10}$$

We call this quantity *Shapley composition*. Note that the average is here with respect to the Aitchison geometry, i.e. with perturbations and a powering rather than sums and a scaling.

The following is the main theoretical result of the paper.

Theorem. *The Shapley composition is the unique quantity that satisfies the following properties on the Aitchison simplex:*

Linearity *with respect to the model:*
$\alpha,\beta\in\mathbb{R}, \forall i\in\mathcal{I}$,
$\boldsymbol{\phi}_i(\alpha\odot\boldsymbol{f}\oplus\beta\odot\boldsymbol{g}) = \alpha\odot\boldsymbol{\phi}_i(\boldsymbol{f})\oplus\beta\odot\boldsymbol{\phi}_i(\boldsymbol{g})$;
Symmetry:
$\forall S\subseteq\mathcal{I}\backslash\{i,j\}$, $\boldsymbol{v}(S\cup\{i\}) = \boldsymbol{v}(S\cup\{j\}) \Rightarrow \boldsymbol{\phi}_i = \boldsymbol{\phi}_j$;
Efficiency:

$$\bigoplus_{i=1}^{d}\boldsymbol{\phi}_i(\boldsymbol{f},\boldsymbol{x},Pr) = \boldsymbol{f}(\boldsymbol{x}) \ominus \mathbb{E}^{\mathcal{A}}_{Pr}[\boldsymbol{f}(\boldsymbol{X})]. \tag{11}$$

A proof of this result is given in the supplementary material [22]. Shapley compositions are thus the natural multidimensional extension of the Shapley value framework on the Aitchison simplex. In the next section we give a number of compelling examples of how this can be used to explain multiclass probabilistic predictions.

6 Explaining a multiclass prediction with Shapley compositions

Given a probabilistic prediction $\boldsymbol{f}(\boldsymbol{x})\in\mathcal{S}^D$, the Shapley composition $\boldsymbol{\phi}_i(\boldsymbol{f},\boldsymbol{x},\Pr)$ describes the contribution of the ith feature value to the prediction. The efficiency property shows how the probability distribution is perturbed from the *base* distribution $\mathbb{E}^{\mathcal{A}}_{\Pr}[\boldsymbol{f}(\boldsymbol{X})]$, i.e. the expected prediction regardless of the current input, to the actual prediction $\boldsymbol{f}(\boldsymbol{x})$. In the standard Shapley formulation recalled in Section 3, the prediction is one-dimensional such that the Shapley quantity is a scalar. In applications where there are more than two possible classes, the prediction is multidimensional such that the Shapley quantity (the Shapley composition) is too. Both live in the same space: the probability simplex. In this section, we discuss how the set of Shapley compositions can be analysed to better understand the contribution and influence of each feature's value on the prediction.

6.1 Visualisation in an isometric-log-ratio space

The Shapley compositions can be visualised in a $(D-1)$-dimensional Euclidean space isometric to the simplex with the ILR transformation presented in Section 4.2. As we will see, this space is intuitive since it is a standard real vector space and it is additive. In what follows, we discuss some examples of Shapley composition-based explanations in an ILR space.

6.1.1 Three classes

Our first illustration uses the well-known Iris classification dataset consisting of a set of flowers described by 4 features: sepal length and width, and petal length and width. The aim of the classification task is to predict to which of the three species (*setosa*, *versicolor* and *virginica*) a flower belongs.

In the present example, a Support Vector Machine (SVM) with a radial basis function (rbf) kernel is used as a classifier. Pairwise coupling [31] is used to obtain a probabilistic prediction. Figure 2 shows the explanation of the classifier prediction for one *versicolor* instance where the Shapley compositions move the distribution from the base to the prediction. Having the highest norm, the petal width and length are the features contributing the most to the prediction and move the base distribution into the *versicolor* maximum probability region (maximum probability region boundaries are the dashed gray rays). Class-compositions are represented by coloured dashed vectors.

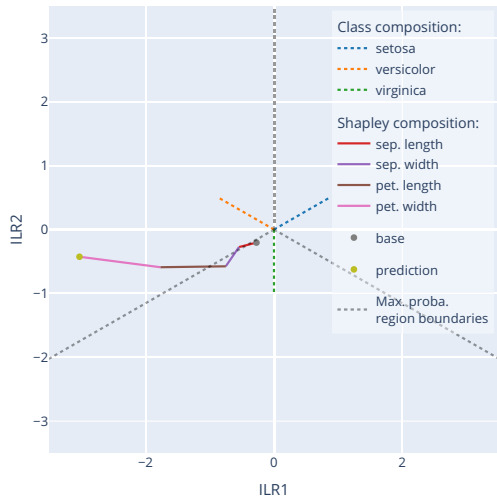

Figure 2. The sum of the Shapley compositions in an ILR space from the base distribution to the prediction for the classification of an Iris instance.

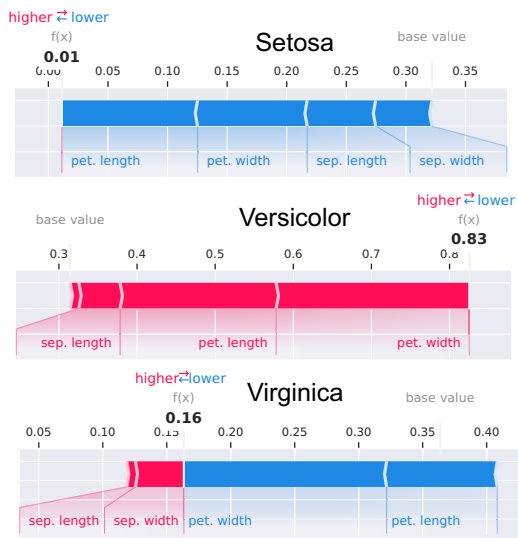

Figure 3. Visualisation of the Shapley values for each class in a one-vs-the-rest manner for the same instance as in Figure 2, obtained using the SHAP toolkit [20]. The red/blue bars represent positive/negative contributions of each feature on the prediction.

A class-composition is defined as a unit norm composition going straight to the direction of one class and uniformly against all the others (see the supplementary material [22] for a formal definition). Note that the class-compositions are not mutually orthogonal. This is because a positive contribution toward one class has necessarily to lead to a negative contribution toward at least another class to preserve the structure of the simplex.

The Shapley composition for the petal length is almost orthogonal to the *virginica* class-composition: for this instance, this feature does not contribute to the weight of the predicted probability for this class. Having a Shapley composition going straight to the opposite direction of one class-composition would suggest that the corresponding feature's value contributes to rejecting this class. This is somewhat the case for the sepal length. However, because its Shapley composition has a low norm, this feature contributes little to the prediction.

Alternatively, one could analyse this instance by applying the standard Shapley value in a one-vs-rest manner, explaining the feature contributions separately for each class. Figure 3 shows how each explanation is usually visualised with the SHAP toolkit [20]. The prediction is explained for each class one-by-one independently from one another, which makes it hard to appreciate the influence of one feature on the full distribution. Moreover, there is no guarantee that the intermediate full distribution remains on the simplex. In contrast, with our approach, the influence of one feature's value on the full prediction can be analyse with a single quantity, the Shapley composition, in a single coherent and easily interpretable plot.

6.1.2 Four classes

In a four-class example, the simplex is 3-dimensional. We illustrate this with a simple handwritten digit recognition task[5]. It consists of classifying an 8×8 image as representing one of the digits 0, 1, 2 or 3. Since there are 64 pixels, considering each pixel as a feature would correspond to 64 Shapley compositions. Moreover, the pixels will be highly correlated. Since our goal here is to provide simple illustrative examples, we reduce the number of features to 6 using a principal component analysis for better clarity and conciseness. An SVM with a rbf kernel and pairwise coupling is again used as a probabilistic classifier. A similar analysis as before can be applied here but within a 3-dimensional plot as illustrated in Figure 4.

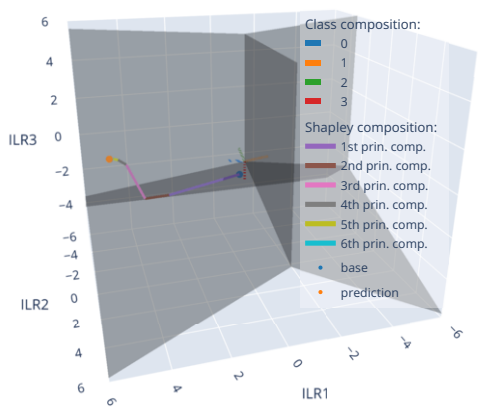

Figure 4. Shapley compositions in a 3-dimensional ILR space for a four classes digit recognition task. The Shapley compositions are summed in the ILR space from the base distribution to the prediction. The gray transparent walls mark out the four maximum probability decision regions.

[5]We use the scikit-learn's digits dataset [24].

To better understand how this space is divided into four regions–each representing the maximum probability region for one class–one can think about the shape of a methane molecule. The hydrogens correspond to the vertices and the carbon to the center of a tetrahedron i.e. a 3-dimensional simplex. The relative positions of the class-compositions in the ILR space are the same as the bonds between the carbon and hydrogen: the angles are $\approx 109.5°$. In this example, the tested instance is a 0[6].

6.2 More classes: groups of classes and balances

When more than three classes are involved, the dimensions of the ILR space cannot be visualised all at once. However, 2 or 3-dimensional subspaces can still be visualised. In order to select the ILR components to investigate, one needs to understand what they refer to. In this section, we briefly discuss the interpretation of the ILR components.

A component of an ILR space can be interpreted as a *balance*, i.e. a log-ratio of two geometrical means of probabilities [12, 9, 23]: one giving the central values of the probabilities in one group of classes and one for another group of classes. Therefore, a balance is here comparing the weight of two groups of classes. The set of balances is built such that they are geometrically orthogonal meaning they provide non-redundant information[7]. This can be illustrated by a sequential binary partition or bifurcation tree. Two examples are given in Figures 5 and 6. Figure 5 shows the bifurcation tree corresponding to the basis obtained with the Gram-Schmidt procedure as in [12] which is the one used in the examples of Figures 2 and 4 with respectively $D = 3$ and $D = 4$. Each node of the tree is a balance, i.e., an ILR component. The first balance $\tilde{p}_1$ compares the probability for class 1 with the probability for class 2. Each next balance then recursively compares the probability for the next class with the probabilities for the previous classes independently of all the others.

In some applications, one may be interested in particular comparisons of groups of classes not necessarily given by a basis in the form of Figure 5. For instance, as in an example presented in [9], if one wants to compare political parties or groups, it may be pertinent to have a balance comparing left and right-wing groups. But sometimes there are no obvious relevant comparisons to study. In the handwritten digit recognition problem, one may want to compare odd with even numbers or primes with non-primes (although, being essentially a shape recognition problem, and the shape of the numbers being independent of their arithmetic properties, such comparisons may not be pertinent).

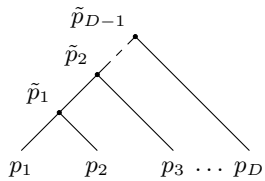

Figure 5. Bifurcation tree corresponding to the basis obtained with the Gram-Schmidt procedure as in [12] and used in the examples of Figures 2 and 4.

We use the basis of Figure 6 for a 10-class digit recognition task. In this example, the bifurcation tree is obtained from the dendrogram of an agglomerative clustering of classes: for each class, the set of predictions is modelled by a logistic-normal [4], with equal covariance,

[6]More examples and better visualisations can be obtained from the Jupyter notebooks: https://github.com/orgs/shapley-composition.

[7]Not to be confused with statistical uncorrelation [23].

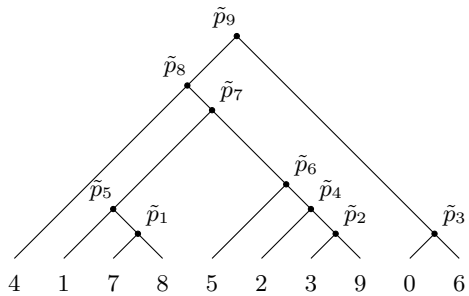

Figure 6. Bifurcation tree used in the 10-class digit recognition task discussed in Section 6.2 and in Figure 7.

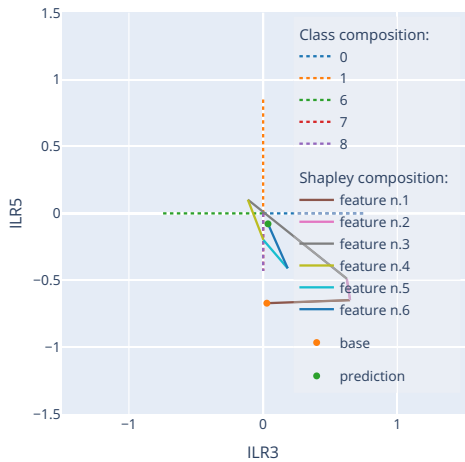

Figure 7. The sum of the Shapley compositions from the base to the prediction in the ILR subspace made of $\tilde{p}_3$ and $\tilde{p}_5$ for a test instance from class 2. $\tilde{p}_3$ compares the probability assigned for class 0 with the probability assigned for class 6 and $\tilde{p}_5$ compares the probability assigned for class 1 with the group of probabilities assigned for class 7 and 8. The color dashed vectors represent the class-compositions with non-zero projection.

and classes are recursively merged with respect to the Mahalanobis distance. Consider, in Figure 7, the third and fifth ILR dimensions ($\tilde{p}_3$ and $\tilde{p}_5$). Effectively, we are saying that we are interested in comparing the probability assigned for class 0 with the probability assigned for class 6, and in comparing the probability assigned for class 1 with the group of probabilities assigned for classes 7 and 8. $\tilde{p}_3$ depends only on the probability for the digits 0 and 6, and $\tilde{p}_5$ depends only on the probabilities for the digits 1, 7 and 8. The class-compositions for the other digits have a zero projection within this subspace and are therefore discarded in Figure 7. The class-compositions for 0 and 6 are orthogonal to the class-compositions for classes 1, 7 and 8. Indeed, the set of classes making the balance $\tilde{p}_3$ and the set of classes making $\tilde{p}_5$ have no intersection.

In contrast, in the example of Figure 2, $\tilde{p}_1$ is comparing the probabilities for the class *setosa* with the probability for the class *versicolor* and $\tilde{p}_2$ is comparing the probabilities for the class *virginica* with the group of probabilities for *setosa* and *versicolor*. In Figure 2, the class-compositions are exhaustively present and are therefore geometrically dependent and none of them are orthogonal. In Figure 7, the classes are not all represented such that the class-compositions projections can be orthogonal. In other words, since we look at only a subspace of an ILR space, we are not looking at the full probability distribution.

In the example of Figure 7, since $\tilde{p}_5$ is comparing 1 with the group of digits 7 and 8, the projection on this line of the class-compositions for 1 goes in an opposite direction than the one for the class-compositions for 7 and 8. The latter two are equal and half as long as the former. In this way, $\tilde{p}_5$ compares the probability for 1 with the group of probabilities for 7 and 8 with the same weight. In other words, in this subspace, the class-compositions for 7 and 8 are reweighted such that this group of two classes has the same weight as the group made of the single class 1.

Within this space, Shapley compositions can be explored as in the examples of Figures 2 and 4, keeping in mind that this is a subspace of a full ILR space.

6.3 *Angles, norms and projections*

An explanation can be summarised by sets of angles, norms and projections:

- The norm of a Shapley composition gives the strength of the contribution of the feature's value to the prediction. This measures the overall contribution of the feature, regardless of its direction.
- The angle between two Shapley compositions informs about their orthogonality. Orthogonality suggests that the features are non-redundant for the given instance. A negative angle would suggest that the features have an opposite influence on the prediction.
- The projections of a Shapley composition on the set of class-compositions inform in favour of, or against, which classes a feature's value is contributing.

To give a few examples, for the Iris example of Figure 2, the norms for each Shapley composition are ≈ 1.27, 1.02, 0.36 and 0.28 respectively for the petal width, length and sepal width and length, confirming the features' importance one would expect from Figure 2. The projection of the petal length's composition on the *virginica* class-composition is ≈ 0.01 confirming the low influence of this feature on the probability for this class. Finally, note that the cosine similarity between the Shapley compositions for petal length and width is close to one (≈ 0.99) which confirms these features are moving the distribution toward the same direction while the compositions for sepal and petal width have a cosine similarity of 0.45 confirming they point to complementary directions.

6.4 *Histograms and parallel coordinates*

For a classification problem with at most 4 classes, an ILR space can be fully visualised within a single figure. However, for more classes we cannot visualise the full ILR space and therefore have to explore subspaces. In this section we discuss alternative visualisations.

The Shapley composition can be visualised using a bar plot like discrete probability distributions. Figure 8 shows the Shapley compositions of the Iris classification example as histograms. Note that in Figure 1, the histograms were showing the probability distributions as the successive perturbation of the base by the features' contribution. The histograms in this section refer to the visualisation of Shapley compositions for each feature separately. A more uniform histogram reflects less contribution of the feature's value to the change of the probability distribution (e.g. the sepal length in Figure 8). In contrast, the Shapley compositions for the petal length and width have a high value for the *versicolor* class, in comparison to the others. This confirms the contribution of these features toward this class.

As another illustration, Figure 9 shows the Shapley compositions of the 10-class digit recognition example. Here, and contrary to the visualisation of the compositions in an ILR space as in Section 6.2, one can analyse all parts of each composition within a single plot. In this

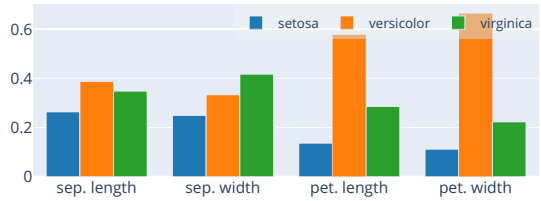

Figure 8. Shapley compositions visualised as histograms for the Iris classification example.

example, the high value for digit 2 for the first principal component confirms the contribution of this feature toward this class.

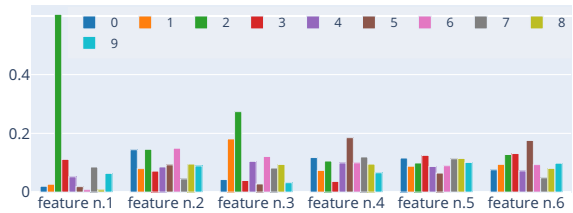

Figure 9. Shapley compositions visualised as histograms for the ten classes digit recognition example.

Another way to visualise the full compositions is with parallel coordinates. After sorting the features by their contribution (i.e. the norm of their Shapley composition), the successive perturbation of the distribution can be visualised as probability lines from the base distribution to the prediction. Figure 10 shows such a plot for the digit recognition example. With this visualisation, we can compactly see how the probability distribution is transformed by each feature contribution from the base distribution to the predicted one. In this example, the probability for digit 2 increases the most with the contribution of feature 1. This feature does not contribute in the change of the probability for digit 3 as suggested by the horizontal red segment. The next feature continues to increase the probability for digit 2 while decreasing the others.

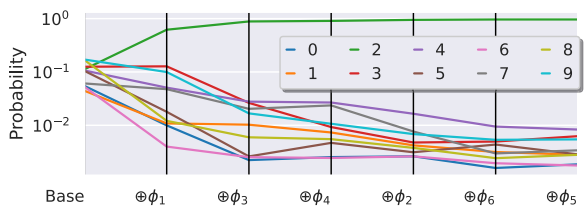

Figure 10. Parallel coordinates visualisation of the successive perturbation of the base distribution by Shapley compositions (ordered by importance, i.e., norm). The final distribution on the right side is the prediction.

6.5 The estimation algorithm

The estimation algorithm used in this work for computing the Shapley compositions is an adaptation of Algorithm 2 in [28]. Since the resulting Shapley compositions are approximations, the efficiency property does not necessarily hold. In order for the set of estimated Shapley compositions to respect the efficiency property, each Shapley composition is adjusted following a similar method as in the sampling approximation in the SHAP toolkit[8]. We refer the reader to the supplementary material [22] for details.

Note that the Shapley composition framework can be applied on many different types of data, such as images. However, the main limitation is algorithmic: the complexity of the algorithm increases with the number of features, as with the standard Shapley value framework. Moreover, the estimation algorithm assumes that the features are independent. Estimating Shapley values without such assumption has been discussed in the literature [1]. We leave the exploration of estimation algorithms of the Shapley compositions without the features-independence assumption and for data with a large number of features for future work.

[8] https://github.com/shap/shap/blob/master/shap/explainers/_sampling.py

7 Discussion and conclusion

The use of standard Shapley values for explaining multiclass machine learning models has been rarely discussed in the literature. However, the computation of the Shapley values on each output dimension one-by-one can be encountered. To be more precise, for an D-class problem ($D > 2$), it may first sound natural to compute a Shapley value on the logit of the probability for each class resulting in a D-dimensional vector of the Shapley values. Even if the efficiency property holds with the standard addition, i.e. the sum of the element-wise logit of the base distribution with such vectors for each feature is equal to the element-wise logit of the prediction, the path from the base to the prediction may go out of the simplex, i.e., the space of probability distributions, which is counter-intuitive and indeed incoherent. Moreover, such a strategy would require running D independent explanations contrary to the Shapley composition approach which requires a single explanation process.

As far as we are aware, this paper is the first to propose an extension of the Shapley value framework to the multidimensional simplex for explaining a multiclass probabilistic prediction in machine learning. We saw how the formalisation of the standard Shapley value naturally extends to the simplex using the Aitchison geometry. The resulting Shapley quantity is a composition (distribution), i.e. a vector living on the probability simplex. It is referred as *Shapley composition* and explicates the contribution of a feature's value to a prediction. To be more precise, it tells how a feature's value moves the distribution from the base one to the predicted one on the simplex. We saw that the Shapley composition is the unique quantity that satisfies the linearity, symmetry and efficiency on the Aitchison simplex.

The Aitchison geometry gives to the simplex an Euclidean vector space structure. For explaining a prediction, Shapley compositions can be visualised and analysed through angles, norms and projections. They inform on both the strength and the direction of each feature's value effect. Living on the probability simplex, i.e. the same space as discrete probability distributions, the Shapley compositions can also be visualised as histograms. Parallel plots of probabilities can also be visualised to keep track of the change in the distribution induced by each feature's value.

The literature about the use of Shapley values in machine learning is extensive. Many estimation algorithms have been developed, many applications of the Shapley value have emerged, and large-scale experiments have been conducted. In contrast, our paper presents limited experimental results as simple proofs of concept and illustrations. However, the main contribution of this work is theoretical and methodological. We believe this work lays proper foundations to foster the research in explainable machine learning, especially for multidimensional and multiclass predictions.

Acknowledgements

The work of PF and MPN was supported by TAILOR[9], a European research network funded by the EU Horizon 2020 research and innovation programme under GA No 952215. This work wouldn't have happened without a research visit of PGN at the University of Bristol made possible by the TAILOR Connectivity Fund. The work of PGN and JFB was also supported by the LIAvignon chair.

We thank Telmo de Menezes e Silva Filho from the University of Bristol for suggesting parallel coordinates to visualise Shapley compositions. We also thank the anonymous reviewers for helpful comments.

References

[1] K. Aas, M. Jullum, and A. Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298:103502, 2021.

[2] J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):139–177, 1982.

[3] J. Aitchison. Simplicial inference. In D. S. P. R. Marlos A. G. Viana, editor, *Algebraic Methods in Statistics and Probability*, Contemporary Mathematics 287. American Mathematical Society, 2001.

[4] J. Aitchison and S. M. Shen. Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2):261–272, 1980.

[5] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson. Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5):e1424, 2021.

[6] D. W. Apley and J. Zhu. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4):1059–1086, 06 2020.

[7] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. Publisher: Springer.

[8] A. Datta, S. Sen, and Y. Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 598–617, 2016.

[9] J. J. Egozcue and V. Pawlowsky-Glahn. Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, 37(7): 795–828, 2005.

[10] J. J. Egozcue and V. Pawlowsky-Glahn. Evidence information in bayesian updating. *Proc. International Workshop on Compositional Data Analysis*, 05 2011.

[11] J. J. Egozcue and P.-G. Vera. Evidence functions: a compositional approach to information. *SORT-Statistics and Operations Research Transactions*, 1(2):101–124, Dec. 2018.

[12] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical geology*, 35(3):279–300, 2003.

[13] A. Fisher, C. Rudin, and F. Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177): 1–81, 2019.

[14] L. Franceschi, C. Zor, M. B. Zafar, G. Detommaso, C. Archambeau, T. Madl, M. Donini, and M. Seeger. Explaining multiclass classifiers with categorical values: A case study in radiography. In H. Chen and L. Luo, editors, *Trustworthy Machine Learning for Healthcare*, pages 11–24, Cham, 2023. Springer Nature Switzerland.

[15] L. Franceschi, M. Donini, C. Archambeau, and M. Seeger. Explaining probabilistic models with distributional values. *arXiv preprint arXiv:2402.09947*, 2024.

[16] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 2015. Publisher: Taylor & Francis.

[17] B. M. Greenwell, B. C. Boehmke, and A. J. McCarthy. A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755*, 2018.

[18] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang. Xai—explainable artificial intelligence. *Science Robotics*, 4 (37):eaay7120, 2019. doi: 10.1126/scirobotics.aay7120.

[19] A. Lamens and J. Bajorath. Explaining multiclass compound activity predictions using counterfactuals and shapley values. *Molecules*, 28(14), 2023.

[20] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[21] P.-G. Noé. *Representing evidence for attribute privacy: bayesian updating, compositional evidence and calibration*. PhD thesis, Université d'Avignon, 2023.

[22] P.-G. Noé, M. Perelló-Nieto, J.-F. Bonastre, and P. Flach. Explaining a probabilistic prediction on the simplex with shapley compositions. *arXiv preprint arXiv:2408.01382*, 2024. This version contains the supplementary material: https://arxiv.org/abs/2408.01382.

[23] V. Pawlowsky-Glahn, J. J. Egozcue, and R. Tolosana-Delgado. *Modeling and Analysis of Compositional Data*. John Wiley & Sons, 2015.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[25] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier. KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[26] L. S. Shapley et al. A value for n-person games. In *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press Princeton, 1953.

[27] K. Sokol and P. Flach. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 56–67, New York, NY, USA, 2020. Association for Computing Machinery.

[28] E. Štrumbelj and I. Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41:647–665, 2014.

[29] L. V. Utkin, A. V. Konstantinov, and K. A. Vishniakov. An imprecise SHAP as a tool for explaining the class probability distributions under limited training data. *arXiv preprint arXiv:2106.09111*, 2021.

[30] L. V. Utkin, A. Petrov, and A. Konstantinov. Modifications of shap for local explanation of function-valued predictions using the divergence measures. In D. G. Arseniev and N. Aouf, editors, *Cyber-Physical Systems and Control II*, pages 52–64, Cham, 2023. Springer International Publishing.

[31] T.-F. Wu, C.-J. Lin, and R. Weng. Probability estimates for multi-class classification by pairwise coupling. *Advances in Neural Information Processing Systems*, 16, 2003.

[32] T. K. Yoo, I. H. Ryu, H. Choi, J. K. Kim, I. S. Lee, J. S. Kim, G. Lee, and T. H. Rim. Explainable Machine Learning Approach as a Tool to Understand Factors Used to Select the Refractive Surgery Technique on the Expert Level. *Translational Vision Science & Technology*, 9(2):8–8, 02 2020.

[9] https://tailor-network.eu