ECAI 2025
I. Lynce et al. (Eds.)

doi:10.3233/FAIA250902

Group Explainability Through Local Approximation

Mattia Setzu[a,*], Riccardo Guidotti[b], Dino Pedreschi[b] and Fosca Giannotti[b]

[a]University of Pisa
[b]University of Pisa, ISTI-CNR Pisa
ORCID (Mattia Setzu): https://orcid.org/0000-0001-8351-9999, ORCID (Riccardo Guidotti): https://orcid.org/0000-0002-2827-7613, ORCID (Dino Pedreschi): https://orcid.org/0000-0003-4801-3225, ORCID (Fosca Giannotti): https://orcid.org/0000-0003-3099-3835

Abstract. Machine learning models are becoming increasingly complex and widely adopted. Interpretable machine learning allows us to not only make predictions but also understand the rationale behind automated decisions through explanations. Explanations are typically characterized by their scope: local explanations are generated by local surrogate models for specific instances, while global explanations aim to approximate the behavior of the entire black-box model. In this paper, we break this dichotomy of locality to explore an underexamined area that lies between these two extremes: meso-level explanations. The goal of meso-level explainability is to provide explanations using a set of meso-level interpretable models, which capture patterns at an intermediate level of abstraction. To this end, we propose GROUX, an explainable-by-design algorithm that generates meso-level explanations in the form of feature importance scores. Our approach includes a partitioning phase that identifies meso groups, followed by the training of interpretable models within each group. We evaluate GROUX on a collection of tabular datasets, reporting both the accuracy and complexity of the resulting meso models, and compare it against other meso-level explainability algorithms. Additionally, we analyze the algorithm's sensitivity to its hyperparameters to better understand its behavior and robustness.

1 Introduction

Explainable Artificial Intelligence (XAI) and Interpretable Machine Learning have become an integral component of automated decision making pipelines [5, 34]. Whether the application is in financial [16], medical [37], judiciary [51], or commercial [23], XAI models provide explanations that empower a large set of stakeholders. Model developers are able to understand and thus improve the models. Auditors are able to understand the compliance of the model with current regulations, e.g., GDPR [39]. Users are able to understand if the model is behaving fairly and reasonably with respect to them.

Central to XAI are *explanations* [5]. While initial efforts in this research area have focused on *post-hoc* explainability, which provides explanations for black-box models, here we focus on explainability *by-design*, i.e., models which provide explanations built-in. Explanations are either local, or global. The former aim to understand model behavior *locally*, i.e., in a small neighborhood of it domain, while the later aim to understand it *globally*, i.e., over its entire domain.

Local explanations provide a local approximation of a model, and generally aim to approximate it with a local interpretable surrogate. In other words, they provide a point estimate of a model. Constraining the approximation locality eases the problem, and tends to improve the accuracy of the explanation algorithm [40]. Point estimates also enable a large set of families of explanations, both factual, and counterfactual [40, 29, 54, 24, 49]. Factual explanations define the *what* and *why* of a model, providing evidence of its behavior. Counterfactual explanations instead define the *why not* and *what if* of a model, providing conditional evidence of its behavior. Each explanation family has its own properties, and elucidates different interpretations and reasoning in its user [21] both according to its (counter)factual nature, and its form. Saliency explanations [40, 29], also referred to as feature importance, mark relevant features, highlighting where a model focuses its attention. Rule explanations [18, 41] instead conditionally define the behavior of a model, and describe under what circumstances a model behaves in one or another way. As such, both families they provide factual explanations. On the other hand, counterfactual explanations [24] provide rules triggering a change in model behavior, and as such, are contrastive explanations. Due to their variety and accuracy, local explanations are usually favored.

Global explanations, instead, aim to provide a complete approximation of the model, thus providing an interpretable surrogate for the model itself [10, 12]. For this, they face significantly a harder challenge, often due to the difference in model capacity and complexity. Moreover, global explanations must rely on a smaller set of explanation families [45]. Counterfactuals, which often find success and clear interpretation as local explanation, are inherently local, thus lose significance when applied on a global scale [17]. Saliency explanations are inadequate as well, either due to their reliance on a reference instance, or due to the heterogeneity of the data itself [56].

Still, locality is a false dichotomy, as locality exists on a spectrum. As humans, we may reason on different layers of generalization, from a local, gradually to a global one [7]. As such, it is natural for explanations to follow along the same line. Rather than understand the models on a single prediction, or on average, it would be beneficial to understand it on an intermediate, i.e., *meso*, level. Figure 1 provides a visual interpretation of the spectrum. On the one end, local explanations focus on a local understanding of the model, and provide a single model to understand the prediction on a single inference – see Figure 1a. On the other end, global explanations provide a single model to holistically understand a model – see Figure 1c. In between the two, the meso level provides a sweet spot between the precision of local explanations, and the holistic view of global ones, allowing us to identify behavioral patterns of the model at different

* Corresponding Author. Email: mattia.setzu@unipi.it

Figure 1: Locality of different explanations. Local explanations address a single instance with a single model. Global explanations address all instances with a single model. In between the two, meso explanations instead address each *meso group* of instances with a single model. Color-filled instances are explained by models of the same color.

(a) Local (b) Meso (c) Global

scales. This allows for finer-grain understanding and control of the complexity of the explanation.

To accommodate this view, in this paper, we propose GROUX (GROUp eXplainability), an algorithmic framework for by-design meso-level explainability. GROUX relies on a three-step pipeline which first encodes the data in a suitable space, then identifies groups within the data, and finally learns meso-level explainable by-design models on each of group. The encoding step facilitates the discovery of meso-level groups, and the generation of meso-level explanations by encoding relationships between instances. In particular, among several possible approaches, we propose using Piecewise Linear Approximation (PLA) [35] to encode each feature during the encoding step. This choice introduces an inductive bias into the subsequent clustering process, under the assumption that instances approximated by similar linear models should be grouped together. In addition to standard clustering algorithms [53, 19], we also adopt fair clustering methods [27, 44] to ensure a more balanced distribution of instances across different classes. Experiments conducted on benchmark tabular datasets for classification tasks reveal that, as a global inference model, GROUX exhibits high variance. Although it generally underperforms compared to other interpretable global models, it can outperform them when hyperparameters are carefully tuned.

In Section 2 we review the existing literature while highlighting relevant background notions useful for the definition of our proposal. In Section 3 we formalize the problem and we introduce GROUX and its hyperparameters. In Section 4 we compare GROUX with a set of baselines and competitors, and analyze its performance as both an inference and explanation model. Finally, we conclude in Section 5.

2 Related Work and Background

Our proposal relates to three main research areas: local and global explainability, linear separability, and linear representations. In this section we briefly illustrate these three aspects, and illustrate their state of the art approaches.

Local vs Global Explainability. We can place the roots of meso-level explainability in *local to global* explainability algorithms, which introduce the locality spectrum [36]. Such algorithms rely on generalizing local explanations to global ones, thus creating a spectrum of locality. Generalization is highly coupled with the family of local explanations. Feature importance and saliency implement vectorial merging of explanations, often averaging them [4, 43, 30]. Decision rules instead follow a more sophisticated approach, often requiring rule selection [2, 42, 45, 48, 57], or rule inference [1, 8, 20, 46]. Intermediate solutions of this process, of which the literature is currently lacking, yield meso-level explanations.

Local to global explainability is typically limited as it needs a set of local explanations, requiring a costly and unstable preprocessing step of local explanation generalization [46]. Instead, meso-level explainability aims to directly infer meso-level explanations the problem. On this line, there are several algorithms of interest.

MUSE [26] models the problem on two levels, and induces sets of decision rules on each: one on the "meso-level", and one on the local level. The first level is comprised of decision rules acting as *descriptors*, and defining a meso-level subspace. These define the meso groups by partitioning the space in an explainable by design fashion, thus providing an explanation of the meso-level groups out of the box. The second level is comprised of sets of decision rules, each set locally explaining a meso group. The two sets are learned jointly through submodular optimization.

ε-maps [9] instead applies a top-down approach, and induces feature importance scores, rather than decision rules. Starting from a set X of instances, it first induces an approximate convex hull, and then generates a set of local explanations for each of its vertices. If they have a variance below a predefined threshold ε^2, then the set is deemed consistent, and a "meso"-level group is identified. If instead the explanations are sufficiently different, X is bi-partitioned with the hyperplane yielding maximum explanation difference, identifying two candidate meso-level groups. Much like in top-down decision tree induction, the process is recursively repeated until meso-level groups of variance below ε^2 are found. Finally, local explanations of instances in each meso group are averaged, yielding a set of meso-explanations. Unlike MUSE, ε-maps only induce explanations at meso level, without any descriptors.

PARTREE [19] and FPARTREE [27] also follow the top-down approach. They induce an unsupervised decision tree whose nodes are optimized to yield cohesive and well-separated clusters. Notably, FPARTREE regularizes for balanced clusters. Like in MUSE The result is a set of "meso-level groups" defined by a set of descriptors.

Instead, APE [11] adopts bottom-up approach. Indeed, unlike the other approaches, it does not generate "meso-level groups" for a whole dataset, rather it identifies the largest meso group for a given instance x. APE first identifies x', an adversarial point to x, through Growing Spheres [28], then generates a neighborhood of x on such basis, and finally estimates the quality of a linear explanation built on such local dataset through linear separability tests. Then, either a linear or rule-base meso-level explanation is induced. Unlike MUSE and ε-maps, APE only induces one meso scope, and requires a starting instance to do so, thus only offering partial explanations.

Unlike MUSE, PARTREE and FPARTREE, our proposal does not provide meso-level descriptors. Also, unlike all methods in the literature, it provides meso-level explanations as feature importance scores. Then, unlike APE, GROUX automatically discovers meso-level groups, and provides meso-level, rather than local, explanations. Finally, GROUX also encodes the data to a suitable representation to boost the efficacy of meso-level explanations.

Linear Models and Linear Separability. When it comes to interpretability, linear models are of major interest. By design, linear models learn feature saliency, have convergence guarantees, and are highly flexible [38]. Thus, they are often employed both as standalone interpretable models, or as local surrogates for uninterpretable models [40]. Linear models are defined as

$$f(x) = \theta x + \theta_0,$$

and parametrized by $\theta \equiv \langle \theta_1, \dots, \theta_m \rangle$ and θ_0. In such a family, the parameters offer an explanation of the impact of each feature in the model.

Linear models are a subset of a more abstract set of models, Generalized Additive Models (GAMs) [33], which define a model family

$$f(x) = \theta_0 + \sum^{m} \theta_m f_m(x_m)$$

parametrized by $\theta \equiv \langle \theta_1, \dots, \theta_m \rangle, \theta_0$, and arbitrary complex *shape functions* $f_1, \dots, f_{m-1}$. Trivially, for shape functions implementing the identity function, we have linear models. Additive models find ample use in interpretability, Shapley values [47] and their implementations and derivations [3, 15, 29, 50, 55] being the most noteworthy case.

Thanks to solid theoretical work and empirical results, linear separability is a solved task for two linearly separable sets. Yet, linear separability is a much more complex problem. Generally, given a set of non-linearly separable instances in a vector space, finding linearly separable subsets is nontrivial. If we then look to find maximally large subsets, this degenerates in an NP-complete problem [52]. A handful of approximation algorithms tackle this problem, but they lack theoretical guarantees [14], and are highly randomized [31, 32].

In our proposal, we tackle this issue by encoding the linear relationships between instances directly in the data representation.

Linear Relationships and Data Encoding. Linear models also find application in data approximation and transformation, wherein a dataset X in a domain $\mathcal{X}$, e.g., $\mathbb{R}^m$, is mapped to a dataset Z in a domain $\mathcal{Z}$, e.g., $\mathbb{R}^p$. Such transformation is, by design, linear, and thus a linear model. Principal Component Analysis (PCA) [22] achieves the transformation with an orthogonal basis that removes collinearity between features. Least Squares instead defines linear models globally approximating the data with a linear model. Both model families learn a single linear model for the whole dataset, and are thus *global*. Piecewise-Linear Approximation (PLA) [35] refines this approach for a finer grain modeling, and learns a set of *local* models. For an univariate dataset X, a PLA of X partitions X into a set of segments, each approximated by a linear model, ultimately yielding a set of linear models parametrized by $\{\theta^1, \dots, \theta^n\}$. Such parameters define the encodings $\{z_1, \dots, z_n\}$ of $\{x_1, \dots, x_n\}$. Scaling to multivariate datasets, either through univariate-PLA on each feature, or z-order approximation of multiple features [13], we achieve a linear encoding $Z \in \mathbb{R}^{n \times m}$ of the dataset $X \in \mathbb{R}^{n \times m}$. Notably, PLA and PLA-derived approximations have strong theoretical guarantees, and allow to bound both the number of linear models, and their approximation error [35]. We highlight that PLA is in stark contrast with complex nonlinearl encodings, e.g., deep autoencoders [25] on several fronts. Computationally, PLA is much more efficient. The learned representations, which are linear in nature, are of much easier interpretation, and preserve locality, unlike typical variational encoders which instead leverage a mean field, i.e., global, approach [25].

In GROUX, we rely on linear encodings of the data to inject an inductive bias in the learning algorithm, and favor linearly separable meso-level groups.

Algorithm 1 GROUX. First, it maps the given dataset X to a linear approximation Z (Line 1) in a space $\mathcal{Z}$. Then, it clusters instances in such space (Line 2). Finally, a linear model is learned on each cluster (Line 3).

Input: Data and labels X, Y, clustering hyperparameters θ
Output: Interpretable models $f_1, \dots, f_k$
1: **function** GROUX(X, Y)
2: $Z \leftarrow f^{\varepsilon}_{\Omega}(X)$ ▷ Linear data-encoding
3: $C \leftarrow f_C(Z, Y, \theta)$ ▷ Induce meso groups
4: $f_1, \dots, f_k \leftarrow f_M(c_1, \dots, c_k)$ ▷ Learn interpretable models
5: **return** $f_1, \dots, f_k$

3 Method

In this section we introduce GROUX, a meso-level explanation framework. Let X, Y be the feature matrix, and labels, respectively. We indicate with θ the hyperparameters of the training algorithm, with $f.$ a learned meso-level model, and with Z and C the encoded representation of X and its clustering. GROUX creates meso-level explanations with a three-steps pipeline, which we illustrate in Alg. 1.

First, the data is mapped to an *encoded* representation with a function f_Ω. f_Ω is designed to yield representations $\{z_1, \dots, z_n\}$ which already encode linear relationships *between instances*. This step provides a representation $Z = \{z_1, \dots, z_n\}$, which is then fed to the following clustering step. Secondly, the clustering step clusters Z, learning clusters $C_1, \dots, C_k$ which define the meso groups. Finally, a linear model f^i is learned on each meso-level group. To preserve the explainability of the model, the models are learned on the original data encoding.

At inference time, given a test instance x, GROUX follows a proximity-based approach, and assigns instances to the meso-level group at minimum distance, i.e., the one with the closest centroid. Then, the meso-level model of that group performs the prediction and the associated meso-level explanation. Formally, for an instance x, inference is implemented as:

$$f(x) = f^i(x) \quad \text{s.t. } \hat{c}_i \underset{c \in \{c_1, \dots, c_k\}}{\arg\min} ||c - x||^2 \tag{1}$$

Rather than a single algorithm, GROUX introduces a general algorithmic framework for meso-level explainability, wherein encodings, clustering algorithm, and meso-level learning algorithms can be tuned to specific applications. Here, we focus on a specific implementation, and provide some hyperparameter and ablation studies in the experimental section.

Encoding the Data: f_Ω. The encoding step aims to create a representation which is amenable to both clustering and linear separability. As such, we apply PLA encoding on each feature, resulting in a dataset Z of the same dimensionality of X (Line 1 of Algorithm 1). We bind the PLA to an error ε given by the standard deviation of each feature. Z thus encodes linear relationships between instances, and similar instances will end up having similar encodings. By constructing a linear representation, we aim to inject a simple inductive bias in the downstream clustering step: *instances approximated by similar linear models should be clustered together*.

An instance x_i is encoded by linear models parametrized by $\{\theta^1_a, \theta^1_b\}, \dots, \{\theta^m_a, \theta^m_b\}$, each pair of parameters $\theta^{\cdot}_a, \theta^{\cdot}_b$ encoding the slope and intercept of the linear model, and learned on each of x's

Table 1: Datasets used in the study.

Dataset	Size	Dimensionality	Dataset Balance
Adult	48842	7	.23
Arhythmia	68	279	.30
Credit	690	7	.33
Bank	45211	7	.11
Breast	683	10	.34
Compas	4534	10	.16
Heart failure	299	8	.32
Heloc	10459	24	.48
Magic	19020	11	.35
Nbfi	8308	19	.07
Pima	768	9	.34
Speeddating	1048	58	.17

features. The encoding $f_\omega(x_i)$ of x_i is thus comprised of a concatenation of such parameters:

$$f_\omega(x) \equiv \langle \theta_a^1, \theta_b^1, \dots, \theta_a^m, \theta_b^m \rangle.$$

Then, the dataset encoding is simply the set of its encoded instances:

$$f_\Omega(X) \equiv \langle f_\omega(x_1), \dots, f_\omega(x_n) \rangle.$$

When binding the PLA to a given error ε, we overload the notation with f_Ω^ε. This step aims to inject an inductive bias in the model, which is not encoded in other meso-level approaches in the literature.

Clustering: f_C. The clustering step is implemented through a *fair* clustering algorithm. Indeed, vanilla clustering is oblivious to the label of the data, thus clustering may yield clusters of homogeneous or highly unbalanced labels [53, 27]. In such a setting, any downstream linear models are bound to fail, thus we tackle the problem at its source, and leverage CLUSTERLETS [44], a state-of-the-art fair clustering algorithm. CLUSTERLETS provides clusters with label distribution close to the overall data distribution, thus reducing the impact of label unbalance, and label-homogeneity of the resulting clusters.

CLUSTERLETS first clusters instances on a class-by-class basis, creating single-label clusters named *clusterlets*. Then, it applies matching algorithms to aggregate clusterlets them into balanced clusters $\{c_1, \dots, c_k\}$. CLUSTERLETS relies on three hyperparameters:

- *Matcher:* The matching algorithm clustering the clusterlets.
- k*:* Defines the number of clusterlets per class.
- *hops:* Defines the maximum number of merging steps across sets of clusterlets.

The Matcher can follow different strategies, and cluster clusterlets according to the label distribution of the resulting clusters (Balance Pinball Matcher), their cohesion (Distance Pinball Matcher), or an approximate mix of the two (Centroid Matcher). k and *hops* allow us to tune the size of the GROUX model: the larger the set of initial clusters, the more the extracted clusters, and thus the more the meso-level groups. Much like in vanilla clustering, these parameters can also be tuned to optimize cluster cohesion, at the cost of label unbalance.

Meso-level Models. For the meso-level models, we rely on logistic regressors. While in principle any model can be used at this level, linear models are in line with the linear encoding that we developed in step 1 of GROUX.

4 Experiments

In the evaluation of GROUX, we focus on the following research questions (RQs):[1]

[1] A public implementation is available at https:github.com/msetzu/groux

Table 2: Hyperparameter space for the tested models.

	Description	Domain
Groux		
k	Number of initial clusterlets	$\{3, 5, 20, 50\}$
Encoding	Encoding applied to the data	{ `PLA`, `Identity` }
Matchers	Clusterlet matching algorithms	{ `B-PB`, `D-PB`, `C` }
Decision Tree, ParTree, FParTree		
d	Maximum tree depth	$\{2, 3, 4, 8\}$

Table 3: Performance of meso-level models of GROUX across different matchers, encoders, and initial number of clusters (k). Includes number of meso-level models (Size), ROC AUC, balanced accuracy (BAC) and f1 score (F1).

	ROC AUC ↑	BAC ↑	F1 ↑	Size ↓
Matcher				
B-PB	$.896 \pm .120$	$.896 \pm .120$	$.900 \pm .118$	26.4 ± 28.7
C	$.742 \pm .150$	$.742 \pm .150$	$.746 \pm .155$	2.3 ± 0.8
D-PB	$.801 \pm .142$	$.801 \pm .142$	$.806 \pm .143$	23.4 ± 24.8
Encoder				
Identity	$.839 \pm .149$	$.839 \pm .149$	$.843 \pm .150$	24.4 ± 27.3
PLA	$.835 \pm .140$	$.835 \pm .140$	$.839 \pm .140$	20.4 ± 25.0
k				
3	$.784 \pm .149$	$.784 \pm .149$	$.790 \pm .150$	$3.4 \pm .725$
5	$.810 \pm .147$	$.810 \pm .147$	$.810 \pm .148$	5.6 ± 1.6
20	$.862 \pm .132$	$.862 \pm .132$	$.867 \pm .131$	22.8 ± 10.5
50	$.891 \pm .125$	$.891 \pm .125$	$.895 \pm .124$	57.8 ± 27.8

- *RQ 1.* What hyperparameters impact GROUX's performance?
- *RQ 2.* How effective are the learned meso-level models?
- *RQ 3.* How complex are the learned meso-level models?
- *RQ 4.* Can meso-level models predict on a global-level scale?

We separately analyze meso-level (RQ 1, 2, 3) and global-level (RQ 4) inference. The former addresses the performance of meso-level models in *their own* meso-level groups,while the latter addresses the performance of the whole model, following the inference schema of Eq. 1. Experiments to answer these research questions have been run on 10 randomized runs, and 11 benchmark datasets, indicated in Table 1, on a $80 - 20$ hold-out train-test split. Grid searches have explored a wide range of hyperparameters for each algorithm. The full list can be found in Table 2. Unless otherwise stated, results are averaged across runs.

Competitor algorithms include three model families:

- *Clustering models.* Provide a baseline for meso-level models by identifying a meso-level group per cluster, allowing us to address RQ 2. We first cluster the data, e.g., with k-means, which provides a set of clusters $\{c_1, \dots, c_k\}$. Each cluster is treated as a meso-level group. Then, we simply follow the third step of GROUX, and learn a linear meso-level model on each such cluster. Inference then follows the same process of GROUX instances are associated to the closest cluster, and the meso-level model associated is used as a meso-level model. For this family, we leverage k-means.
- *Explainable clustering models.* Provide meso-level and global-level models, allowing us to address RQ 2, RQ 3, and RQ 4. Specifically, we leverage PARTREE [19] and FPARTREE [27]. Both algorithms learn decision tree-induced clusterings. As such, we can again leverage the same inference schema as for clustering models: each cluster defines a meso-level group, and a meso-level model is learned on each such group.
- *Interpretable global models.* Provide an upper-bound on the global-level inference of the meso-level models, thus allowing us to address RQ 4. We leverage Logistic Regression, Decision Trees

Table 4: Performance of meso-level models of GROUX and competitors. Model selection by f1-score on each dataset. Results averaged across datasets.

	ROC AUC ↑	BAC ↑	F1 ↑	Size ↓
k-means	$.701 \pm .129$	$.701 \pm .129$	$.703 \pm .139$	7.0 ± 6.1
PARTREE	$.695 \pm .151$	$.695 \pm .151$	$.693 \pm 160$	5.0 ± 3.6
FPARTREE	$.759 \pm .173$	$.759 \pm .173$	$.762 \pm .178$	21.1 ± 54.7
GROUX	$.837 \pm .173$	$.837 \pm .173$	$.841 \pm .178$	22.4 ± 54.7

(DT) [6], and Linear Trees (LinearDT) [58]. Linear Trees are decision trees whose splits are univaritate but in the last node before a leaf: in this case, they are multivariate instead.

Neither MUSE [26] nor ε-maps [9] provide an implementation of their algorithm, thus they are not compared against.

How do hyperparameters impact GROUX? We start by studying how different hyperparameters impact meso-level model performance, addressing RQ 1 and RQ 3. Table 3 reports the impact of different hyperparameters on the performance and size, i.e., the number of meso-level models.

Concerning matchers, meso-level balance appears to be a determining factor: the Balance Pinball Matcher (B-PB), which optimizes meso-level groups per label balance, yields by far the best results, followed by the Distance Pinball Matcher (D-PB), and the Centroid Matcher (C). The results are consistent across metrics. Still, both B-PB and D-PB yield larger models, on average with 25.9 ± 29.3 and 22.7 ± 23.8 meso-level models, respectively. The Centroid Matcher instead yields far more compact models, with on average 2.3 ± 0.6 meso-level models. As expected, increasing the model capacity yields better performance.

A less marked performance difference appears when considering the data encoding (f_Ω). While a base identity encoding yields better performances ($\approx +.13$ on all metrics), the PLA encoding yields more compact models (≈ -3.7 meso-level models). Thus, higher-capacity models once again yield better performances.

The number of starting clusterlets (k) also confirms the trend. As k grows, so does the model size, and thus capacity, and performance of the model, which grows of $\approx 10\%$ points across all metrics.

How do meso-level models perform? Table 4 reports performances of a F1-score based model selection of GROUX against its meso-level competitors k-means, PARTREE and FPARTREE, and addresses RQ 2. Results are averaged across dataset.

For k-means, we have leveraged the identified clustering, and trained a meso-level logistic regressor on each cluster. Notably, k-means and PARTREE are not explicitly designed to identify balanced meso-level groups, and thus may yield highly unbalanced – or single-class – meso-level groups, which would impair learning a meso-level model. Thus, we have excluded such unsuccessful "single-class" failures from the results.

As a baseline, k-means yields meso-level models with relatively low performance, but of small size. This is due to the large number of single-class failures, which tend to happen with a large k. A similar behavior is found in PARTREE, which has similar performances and model size. FPARTREE and GROUX, which instead balance classes, fare far better. GROUX outperforms FPARTREE of $\approx 7\%$ across metrics, at a moderate cost of ≈ 1.3 more meso-level models. Thus, GROUX is able to yield more accurate meso-level explanations than FPARTREE, at a similar complexity cost.

Can meso-level models infer at a global-level? Finally, we study how meso-level models perform in global-level inference, addressing RQ 4. Table 5 reports performance of meso-level and global-level models in global-level inference. Scores are averaged across datasets, randomized runs, and hyperparameters configuration. As expected, on average global models retain higher performance, with a marked difference separating meso-level and global-level models.

This difference significantly reduces when performing model selection. Table 6 reports scores averaged across datasets of models of each family, selected by best F1-score. Global-level models hardly have any improvement, with Decision Trees gaining $\approx .03$ across metrics. Meso-level models on the other hand varied improvements across metrics: PARTREE gains $\approx .15$, FPARTREE $\approx .35$, and GROUX $\approx .1$. Thus, meso-level algorithms show higher variance, and a high potential for global-level inference. This is to be expected particularly for GROUX, which is designed to provide meso-level explanations, rather than global-level inference.

Qualitative example. We report an example from a run on the Adult dataset, in which GROUX has discovered 3 meso-level groups. The dataset comprises of 7 features: `Age`, `Capital Gain`, `Capital Loss`, `Education Level`, `Adjustment Weight`, and `Weekly Hours`. The task is binary classification: instances are classified according to their expected future income, either below or above a predefined threshold of $50k\$$. The groups differ, among others, by average `age` $(35.8, 38.7, 44.8)$, `capital gain` $(4440.0, 153.3, 692.7)$, and `education level` $(11.6, 9.4, 10.0)$. Table 7 reports importance scores of each group – features not reported have negligible importance or do not sufficiently discriminate between groups.

Interestingly, features importance vary significantly across groups. Group 0, with lower mean age and higher capital gain, shows high importance in age and education level. This indicates that for younger people, age and education are strong determining factors in predicting future income. It stands to reason that this is a key factor: among young workers, work experience is scarce, thus predetermined factors such as education may well play a strong role. The same goes for education, as highly educated young workers are unlikely to have garnered enough experience to climb the ladder, thus their starting job, which varies a lot according to their education, is a factor as well. Group 1 and 2 instead, show lower coefficients across the board. Still, there is a marked difference. Group 1 comprises of middle-age, lower-educated, and low-capital gain people for which age and workload appear not to impact their income prospects. For this group, high importance comes to education. Thus, this constitutes a group of workers which have stabilized their career, and are unlikely to climb the ladder and thus increase their earnings. Group 2, the oldest of the three, shows an increased importance on age and weekly workload, and near null importance of education level. This indicates a group of older workers which are likely to already be in managerial positions, thus their work contributions are likely to outweigh their education.

5 Conclusions

In this paper, we have introduced GROUX an interpretable model for meso-level explainability. Unlike local or global models, GROUX tackles explanation on a *meso* level: by providing feature importance on a *group*-level, GROUX fills a gap between local and global explainability approaches. Meso-level models show high accuracy, and outperform baseline and competitors. As a pure inference model, GROUX shows high variance, and while in general underperforms w.r.t. other interpretable global models, it is able to outperform them when hyperparameters are carefully tuned.

The algorithm has several avenues for improvement. The encoding function offers lower model complexity, but does not yield largely

Table 5: Inference performance of meso-level and global-level models. Results averaged across datasets, randomized runs, and hyperparameter configurations. Size indicates number of meso-level models, or number of nodes for global DT and LinearDT.

	ROC AUC ↑	BAC ↑	F1 ↑	Size ↓
Meso-level models				
k-means	$.643 \pm .097$	$.643 \pm .097$	$.645 \pm .109$	7.0 ± 6.1
PARTREE	$.640 \pm .087$	$.640 \pm .087$	$.645 \pm .097$	5.0 ± 3.6
FPARTREE	$.648 \pm .107$	$.648 \pm .107$	$.641 \pm .111$	21.1 ± 54.7
GROUX	$.628 \pm .128$	$.628 \pm .128$	$.616 \pm .142$	22.4 ± 26.3
Global-level models				
DT	$.670 \pm .119$	$.670 \pm .119$	$.669 \pm .122$	59.5 ± 97.4
LinearDT	$.688 \pm .129$	$.688 \pm .129$	$.690 \pm .133$	11.7 ± 3.54
Logistic	$.674 \pm .123$	$.674 \pm .123$	$.676 \pm .130$	

Table 6: Global-level inference performance of meso-level and global-level models. Model selection per dataset.

	ROC AUC	BAC	F1	Size
Meso-level models				
k-means	$.663 \pm .103$	$.663 \pm .103$	$.668 \pm .112$	9.4 ± 4.9
PARTREE	$.655 \pm .079$	$.655 \pm .079$	$.663 \pm .086$	4.2 ± 3.2
FPARTREE	$.683 \pm .118$	$.683 \pm 118$	$.683 \pm 119$	8.7 ± 6.4
GROUX	$.729 \pm .114$	$.729 \pm .114$	$.738 \pm .112$	16.2 ± 23.2
Global-level models				
DT	$.717 \pm .119$	$.717 \pm .119$	$.723 \pm .114$	133.5 ± 149.0
LinearDT	$.688 \pm .135$	$.688 \pm .135$	$.690 \pm .139$	11.7 ± 3.71
Logistic	$.674 \pm .128$	$.674 \pm .128$	$.676 \pm .135$	

improved model performance. The learning algorithm shows high variance, thus an automatic tuning of its parameters is also of interest as future research. Finally, although this work focuses on features importance, the flexibility of the framework also allows to use other meso-level models, which we plan on studying in further research.

Acknowledgements

This work has been partially supported by the Italian Project Fondo Italiano per la Scienza FIS00001966 "MIMOSA", by the European Community Horizon 2020 programme under the funding schemes ERC-2018-ADG G.A. 834756 "XAI", by the European Commission under the NextGeneration EU programme National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR) Project: "SoBigData.it Strengthening the Italian RI for Social Mining and Big Data Analytics" Prot. IR0000013 Av. n. 3264 del 28/12/2021, M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR" - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", and "FINDHR" that has received funding from the European Union's Horizon Europe research and innovation program under G.A. 101070212.

References

[1] A. Andrzejak, F. Langner, and S. Zabala. Interpretable models from distributed data via merging of decision trees. In *CIDM*, pages 1–9. IEEE, 2013.

[2] M. Aria, A. Gnasso, C. Iorio, and G. Pandolfo. Explainable ensemble trees. *Comput. Stat.*, 39(1):3–19, 2024.

[3] J. Bento, P. Saleiro, A. F. Cruz, M. A. T. Figueiredo, and P. Bizarro. Timeshap: Explaining recurrent models through sequence perturbations. In *KDD*, pages 2565–2573. ACM, 2021.

[4] U. Bhatt, P. Ravikumar, and J. M. F. Moura. Towards aggregating weighted feature attributions. *CoRR*, abs/1901.10040, 2019.

[5] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, and S. Rinzivillo. Benchmarking and survey of explanation methods for black box models. *Data Min. Knowl. Discov.*, 37(5):1719–1778, 2023.

[6] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

[7] H. Cheng, D. Lian, B. Deng, S. Gao, T. Tan, and Y. Geng. Local to global learning: Gradually adding classes for training deep neural networks. In *CVPR*, pages 4748–4756. Computer Vision Foundation / IEEE, 2019.

[8] H. A. Chipman, E. I. George, and R. McCulloh. Making sense of a forest of trees. *Computing Science and Statistics*, pages 84–92, 1998.

[9] T. Chowdhury, R. Rahimi, and J. Allan. Equi-explanation maps: Concise and informative global summary explanations. In *FAccT*, pages 464–472. ACM, 2022.

[10] M. W. Craven and J. W. Shavlik. Extracting tree-structured representations of trained networks. In *NIPS*, pages 24–30. MIT Press, 1995.

[11] J. Delaunay, L. Galárraga, and C. Largouët. When should we use linear explanations? In *CIKM*, pages 355–364. ACM, 2022.

[12] R. Dwivedi, D. Dave, H. Naik, S. Singhal, O. F. Rana, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, and R. Ranjan. Explainable AI (XAI): core ideas, techniques, and solutions. *ACM Comput. Surv.*, 55(9): 194:1–194:33, 2023.

[13] V. Gaede and O. Günther. Multidimensional access methods. *ACM Comput. Surv.*, 30(2):170–231, 1998.

[14] S. I. Gallant. Perceptron-based learning algorithms. *IEEE Trans. Neural Networks*, 1(2):179–191, 1990.

[15] A. Ghorbani and J. Y. Zou. Neuron shapley: Discovering the responsible neurons. In *NeurIPS*, 2020.

[16] J. W. Goodell, S. Kumar, W. M. Lim, and D. Pattnaik. Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *Journal of Behavioral and Experimental Finance*, 32, 2021.

[17] R. Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Min. Knowl. Discov.*, 38(5): 2770–2824, 2024.

[18] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, and F. Turini. Factual and counterfactual explanations for black box decision making. *IEEE Intell. Syst.*, 34(6):14–23, 2019.

[19] R. Guidotti, C. Landi, A. Beretta, D. Fadda, and M. Nanni. Interpretable data partitioning through tree-based clustering methods. In *DS*, volume 14276 of *Lecture Notes in Computer Science*, pages 492–507. Springer, 2023.

[20] C. Hocquette, A. Niskanen, R. Morel, M. Järvisalo, and A. Cropper. Learning big logical rules by joining small rules. In *IJCAI*, pages 3430–3438. ijcai.org, 2024.

[21] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman. Measures for explainable AI: explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance. *Frontiers Comput. Sci.*, 5, 2023.

[22] J. E. Jackson and F. T. Hearne. Relationships among coefficients of vectors used in principal components. *Technometrics*, 15(3):601–610, 1973.

[23] Z. Jan, F. Ahamed, W. Mayer, N. Patel, G. Grossmann, M. Stumptner, and A. Kuusk. Artificial intelligence for industry 4.0: Systematic review of applications, challenges, and opportunities. *Expert Syst. Appl.*, 216: 119456, 2023. doi: 10.1016/J.ESWA.2022.119456. URL https://doi.org/10.1016/j.eswa.2022.119456.

[24] E. M. Kenny, W. Huang, S. Aryal, and M. T. Keane. Explaining and au-

Table 7: Feature importances of GROUX with three meso-level models on the Adult dataset.

	Importance			Descriptors		
Group	**Age**	**Education Level**	**Weekly hours**	**Age**	**Capital Gain**	**Education Level**
0	.313	.471	.124	35.8	4440.0	11.6
1	.022	.156	.017	38.7	153.3	9.4
2	.058	.018	.057	44.8	692.7	10.0

diting with "even-if": Uses for semi-factual explanations in AI/ML. In *KES-IDT*, volume 411 of *Smart Innovation, Systems and Technologies*, pages 135–145. Springer, 2024.
[25] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2014.
[26] H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec. Faithful and customizable explanations of black box models. In *AIES*, pages 131–138. ACM, 2019.
[27] C. Landi, A. Cascione, M. Marchiori Manerba, and R. Guidotti. Balancing fairness and interpretability in clustering with fairpartree. In *World Conference on Explainable Artificial Intelligence*, pages 1–24. Springer, 2025.
[28] T. Laugel, M. Lesot, C. Marsala, X. Renard, and M. Detyniecki. Comparison-based inverse classification for interpretability in machine learning. In *IPMU (1)*, volume 853 of *Communications in Computer and Information Science*, pages 100–111. Springer, 2018.
[29] S. M. Lundberg and S. Lee. A unified approach to interpreting model predictions. In *NIPS*, pages 4765–4774, 2017.
[30] S. M. Lundberg, G. G. Erion, H. Chen, A. J. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S. Lee. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.*, 2(1):56–67, 2020.
[31] M. Marchand and M. Golea. An approximation algorithm to find the largest linearly separable subset of training examples. In *Proceedings of the 1993 Annual Meeting of the International Neural Network Society*, volume 3, pages 556–559. Erlbaum Associates Hillsdale, NJ, 1993.
[32] M. Marchand and M. Golea. On learning simple neural concepts: from halfspace intersections to neural decision lists. *Network: Computation in Neural Systems*, 4(1):67–85, 1993.
[33] D. F. McCaffrey. Generalized additive models (T. j. hastie and r. j. tibshirani). *SIAM Rev.*, 34(4):675–678, 1992.
[34] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38, 2019.
[35] J. O'Rourke. An on-line algorithm for fitting straight lines between data ranges. *Commun. ACM*, 24(9):574–578, 1981.
[36] D. Pedreschi, F. Giannotti, R. Guidotti, A. Monreale, S. Ruggieri, and F. Turini. Meaningful explanations of black box AI decision systems. In *AAAI*, pages 9780–9784. AAAI Press, 2019.
[37] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol. Ai in health and medicine. *Nature medicine*, 2022.
[38] C. R. Rao. *Linear models and generalizations*. Springer, 2008.
[39] P. Regulation. General data protection regulation. *Intouch*, 25:1–5, 2018.
[40] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *KDD*, pages 1135–1144. ACM, 2016.
[41] M. T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI*, pages 1527–1535. AAAI Press, 2018.
[42] O. Sagi and L. Rokach. Approximating xgboost with an interpretable decision tree. *Inf. Sci.*, 572:522–542, 2021.
[43] J. Schrouff, S. Baur, S. Hou, D. Mincu, E. Loreaux, R. Blanes, J. Wexler, A. Karthikesalingam, and B. Kim. Best of both worlds: local and global explanations with human-understandable concepts. *CoRR*, abs/2106.08641, 2021.
[44] M. Setzu and R. Guidotti. Fair clustering with clusterlets. *arXiv preprint arXiv:639.5516*, 2024.
[45] M. Setzu, R. Guidotti, A. Monreale, and F. Turini. Global explanations with local scoring. In *PKDD/ECML Workshops (1)*, volume 1167 of *Communications in Computer and Information Science*, pages 159–171. Springer, 2019.
[46] M. Setzu, R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, and F. Giannotti. Glocalx - from local to global explanations of black box AI models. *Artif. Intell.*, 294:103457, 2021.
[47] L. S. Shapley et al. A value for n-person games. 1953.
[48] R. Sharma, N. Reddy, V. Kamakshi, N. C. Krishnan, and S. Jain. MAIRE - A model-agnostic interpretable rule extraction procedure for explaining classifiers. In *CD-MAKE*, volume 12844 of *Lecture Notes in Computer Science*, pages 329–349. Springer, 2021.
[49] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 2017.
[50] M. Sundararajan, K. Dhamdhere, and A. Agarwal. The shapley taylor interaction index. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 9259–9268. PMLR, 2020.
[51] H. Surden. Artificial intelligence and law: An overview. *Georgia State University Law Review*, 35, 2019.
[52] M. Tajine and D. A. Elizondo. New methods for testing linear separability. *Neurocomputing*, 47(1-4):161–188, 2002.
[53] P.-N. Tan et al. Data mining introduction. *Peoples Posts and Telecommunications Publishing House, Beijing*, 2006.
[54] S. Wachter, B. D. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *CoRR*, abs/1711.00399, 2017.
[55] J. Wang, J. Wiens, and S. M. Lundberg. Shapley flow: A graph-based approach to interpreting model predictions. In *AISTATS*, volume 130 of *Proceedings of Machine Learning Research*, pages 721–729. PMLR, 2021.
[56] L. Wang, H. Lu, X. Ruan, and M. Yang. Deep networks for saliency detection via local estimation and global search. In *CVPR*, pages 3183–3192. IEEE Computer Society, 2015.
[57] A. I. Weinberg and M. Last. Selecting a representative decision tree from an ensemble of decision-tree models for fast big data classification. *J. Big Data*, 6:23, 2019.
[58] O. T. Yildiz and E. Alpaydin. Linear discriminant trees. *Int. J. Pattern Recognit. Artif. Intell.*, 19(3):323–353, 2005.