Predicting the Evolution of Pain Relief: Ensemble Learning by Diversifying Model Explanations

ANDERSON BESSA DA COSTA, Computer Science Department, Universidade Federal de Minas Gerais
LARISSA MOREIRA, Pain Center, Department of Neurology, Universidade de São Paulo
DANIEL CIAMPI DE ANDRADE, Pain Center, Instituto do Câncer do Estado de São Paulo
ADRIANO VELOSO, Computer Science Department, Universidade Federal de Minas Gerais
NIVIO ZIVIANI, Computer Science Department, Universidade Federal de Minas Gerais, and Kunumi

Modeling from data usually has two distinct facets: building sound explanatory models or creating powerful predictive models for a system or phenomenon. Most of recent literature does not exploit the relationship between explanation and prediction while learning models from data. Recent algorithms are not taking advantage of the fact that many phenomena are actually defined by diverse sub-populations and local structures, and thus there are many possible predictive models providing contrasting interpretations or competing explanations for the same phenomenon. In this article, we propose to explore a complementary link between explanation and prediction. Our main intuition is that models having their decisions explained by the same factors are likely to perform better predictions for data points within the same local structures. We evaluate our methodology to model the evolution of pain relief in patients suffering from chronic pain under usual guideline-based treatment. The ensembles generated using our framework are compared with all-in-one approaches of robust algorithms to high-dimensional data, such as Random Forests and XGBoost. Chronic pain can be primary or secondary to diseases. Its symptomatology can be classified as nociceptive, nociplastic, or neuropathic, and is generally associated with many different causal structures, challenging typical modeling methodologies. Our data includes 631 patients receiving pain treatment. We considered 338 features providing information about pain sensation, socioeconomic status, and prescribed treatments. Our goal is to predict, using data from the first consultation only, if the patient will be successful in treatment for chronic pain relief. As a result of this work, we were able to build ensembles that are able to consistently improve performance by up to 33% when compared to models trained using all the available features. We also obtained relevant gains in interpretability, with resulting ensembles using only 15% of the total number of features. We show we can effectively generate ensembles from competing explanations, promoting diversity in ensemble learning and leading to significant gains in accuracy by enforcing a stable scenario in which models that are dissimilar in terms of their predictions are also dissimilar in terms of their explanation factors.

CCS Concepts: • **Computing methodologies** → **Ensemble methods**; • **Applied computing** → *Health informatics;*

This work was partially funded by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and by projects ATMOSPHERE (Horizon 2020 grant No 777154), RNP-MCTIC (grant No 51119), and MASWeb (grant FAPEMIG/PRONEX APQ-01400-14), and by the authors individual grants from CNPq, FAPEMIG, and Kunumi.

Authors' addresses: A. B. da Costa and A. Veloso, Computer Science Department, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil; emails: {anderson.bessa, adrianov}@dcc.ufmg.br; L. Moreira, Pain Center, Department of Neurology, Universidade de São Paulo, São Paulo, São Paulo, Brazil; email: larissaiulle@usp.br; D. C. de Andrade, Pain Center, Instituto do Câncer do Estado de São Paulo, São Paulo, São Paulo, Brazil; email: ciampi@usp.br; N. Ziviani, Computer Science Department, Universidade Federal de Minas Gerais, and Kunumi, Belo Horizonte, Minas Gerais, Brazil; email: nivio@dcc.ufmg.br.

2637-8051/2021/08-ART36 $15.00
https://doi.org/10.1145/3466781

Additional Key Words and Phrases: Explanatory modeling, predictive modeling, backbone structure learning, diversity, prediction-explanation stability

ACM Reference format:
Anderson Bessa da Costa, Larissa Moreira, Daniel Ciampi de Andrade, Adriano Veloso, and Nivio Ziviani. 2021. Predicting the Evolution of Pain Relief: Ensemble Learning by Diversifying Model Explanations. *ACM Trans. Comput. Healthcare* 2, 4, Article 36 (August 2021), 28 pages.
https://doi.org/10.1145/3466781

1 INTRODUCTION

In this article, we present models for the evolution of pain relief in patients suffering from a chronic pain condition under normal, usual medical management. It must be emphasized that chronic pain (pain lasting for longer than three months) ranks among the most common diseases affecting humans [51] and is the most common cause of years lived with disability worldwide [11]. Chronic pain is present in up to a fourth of the general population [48] and despite the existence of several guidelines and recommendations for its treatment, up to 40% of chronic pain patients may remain symptomatic despite best medical treatment. This is in part due to the heterogeneity of chronic pain mechanisms and individual variables that are not linearly related to the etiology of pain, but to the interplay of its pathophysiology, personal variables, and social context [14]. In this case, the data consists of attributes extracted from patients' self-reports obtained at the first appointment or consultation with the doctor, and our models must predict whether the patient will at the end of the treatment experience reduction in pain relief. In particular, the doctor uses a standardized pain questionnaire, asking the patient to choose the characteristics that best describe her pain (i.e., burning, tingling, sharp or dull). The patient is also asked how long the pain lasts, what makes the pain worse, and what relieves it (i.e., activities, medications, the weather). Predicting the evolution of pain relief is a hard task, because chronic pain can arise from many different conditions, such as fibromyalgia, cancer, arthritis, violent traumas, and many other possibilities [36]. It may be hard to detect these conditions at the first appointment, and the data presents many local structures so the factors contributing to the correct treatment decisions depend on a complex structure that emerges from specific characteristics reported by patients.

Intuitively, if different data points (i.e., patients) are associated with different local structures, then we would expect each structure to be better described by a different model, and then we can get a model for the entire data by combining (potentially simpler) models for all the local structures. In this case, a simple solution is to divide the original data space into biclusters, enabling concurrent feature and data point selection. Each bicluster may approximate a local structure from which a model is built [33]. Another possible solution is to estimate local structures in the data using the Expectation-Maximization algorithm to get maximum likelihood estimates [28]. More often, however, these many-structure phenomena are modeled using the simple *all-in-one* approach, which fits all the available factors (or features) into a single model. The all-in-one approach is clearly sub-optimal, since factors that are important for modeling one structure may become lurking or confounding variables influencing other structures in the data. A branch of a decision tree, for instance, may mix different local structures, or the same local structure may be fragmented in different branches of the tree. Also, parametric models that require combinatorial non-convex optimization, such as gradient descent, would be clearly benefited by decomposing the phenomenon into local structures and exponentially reducing the size of the search space [15].

The main result of this article is a new ensemble learning approach for modeling backbone-structure phenomena. We define a backbone structure as a type of local structure in which there exists a particular set of "backbone features" that, once set, causes the remainder of the features to decompose into independent subsets in the data space. Unlike previous attempts [2], we propose to cluster the explanation space[1] instead of (bi)clustering the

[1]We assume that explanations are given in terms of the central factors that unveil systemic pattern(s) within the model predictions.

data space [33]. By analyzing the sampled model space, we found a strong link between model predictions and model explanations, and we show evidence that models having their predictions explained by the same reasons (or factors) are likely to be suitable for modeling the same local structures in the data space. In summary, the main contributions of this article are:

- We investigate the evolution of pain relief in patients suffering from unknown chronic pain conditions. Specifically, given data from the patient's first consultation the model must predict the likelihood that pain will be significantly reduced at the end of the treatment. This is a particularly interesting problem, because it is defined by phenomena that exhibit the backbone structure. Thus, by learning simpler models that are likely associated with different local structures, we can achieve feature decompositions that algorithms like variable elimination cannot. Many other problems seem to exhibit the backbone structure (e.g., protein folding [15], Alzheimer's diagnosis [20]) and may benefit from our results.
- We propose to learn ensembles from local models (or base models) that present diversity in terms of their explanatory factors. While diversity is recognized as a central element to get significant performance improvements with the ensemble, measuring diversity is not straightforward, because there is no generally accepted formal definition [21]. To promote diversity while learning the ensemble, we select local models that are associated with different explanatory factors, thus our ensemble strategy is essentially a combination of competing explanations for the same phenomenon.
- We show that there is a multiplicity of performant models with diverse explanations, and learning the ensemble by forcing a prediction-explanation stability in the sense that models that are similar in terms of their predictions should have similar explanations, leads to gains in accuracy that are up to 12% when comparing against the best model with feature decomposition and up to 32% compared with all-in-one approaches.

The remainder of the article is organized as follows: Section 2 provides a discussion of relevant related work. Section 3 describes our proposed approach for modeling multi-structure phenomena. In Section 4, we describe our data, and in Section 5, we report the results. Section 6 concludes the article.

2 RELATED WORK

In this section is presented the relevant related work divided in two main subsections: Section 2.1 describes works that aim to model pain sensation and pain evolution, and Section 2.2 presents different perspectives for the problem of how to learn from data.

2.1 Methods for Modeling Pain

Data availability and quality is of crucial importance in chronic pain modeling. In the 1970s and 1980s, few hospitals collected structured data. Moreover, they followed their own nomenclature and definitions, making extremely difficult any attempt to algorithmically modeling pain [30]. Nowadays the amount of data available is huge, together with the advance in the standardization of medical nomenclatures.

Machine Learning. Recent works seek to model chronic pain to predict the evolution of the patient. Navani and Li [30] build a machine learning system for calculating dynamic changes to the weight-based chronic pain risk score on various aspects of health behavior. However, only three sources of information are used: depression, nutrition, and physical activity.

Machine learning approaches to predictive analysis of pain have some well-known limitations. Pieterse et al. [35] cited examples where the generated models are no better than human-analyzed regression models and in some cases are doomed to overfitting. In addition, it is corroborated by the quotation from Goldstein et al. [16] that clinicians are aware that in machine learning algorithms it is not possible to directly see or understand what

exactly influences model prediction. Besides being questionable to predict an event in which it is not possible to change its output. A systematic review in clinical decision support systems for pain management is presented by Pombo et al. [36]. In this review, it was found the great diversity of algorithms to be used in the **Clinical Decision Support Systems (CDSSs)**: rule-based algorithms such C4.5, CART, PRISM; artificial neural networks and statistical learning algorithms are common choices. Also, as reported by Abad-Grau et al. [1], it appears to be hard for medical experts to build valid models when too many variables affect the process. Pombo et al. [36] indicated that this limitation leads to the design of low-accuracy systems.

When choosing the learning algorithm, in general there are two contrasting options: interpretable models by their nature and black box approaches. As the name suggests, interpretable models by their nature are simpler models, such as tree models and linear regression, in which due to their simplicity it requires little effort to draw up an explanation from the learned model. For example, from a tree model it is possible to generate rule sets containing explicity conditions using the features provided, partitioning the instances based on the labels. The downside is that these approaches are generally associated with low-accuracy systems when compared with more complex models. Conversely, black box approaches are, in general, capable of generating more performant models. However, due to inerent associated complexity, they are difficult to interpret. Providing an unique human-readable interpretation from a small neural networks still is a challenge, despite recent advances in model agnostic interpretability techniques [23, 37, 38]. In critical domains such as healthcare sectors, interpretability is a desired characteristic as the system should help clinicians making decisions.

The use of machine learning in pain research is quite diverse [4]. There are works that perform analysis on self-reporting data [19, 41], but with different aims. Authors in Reference [19] presented a machine learning approach that analyzes self-reporting data collected from the integrated biopsychosocial treatment. However, their purpose is to identify an optimal set of features for supporting self-management. Reference [41] studied the question of self-report reliability. They compared self-reporting data with neuroimage to discriminate between individuals with and without chronic pain.

Model Explainability. The lack of interpretability has limited the use of powerful methods such as deep learning and ensemble models in medical decision support [31, 46]. Some of the recent works have focused on interpretability, as we can see in Reference [24], where it is presented an ensemble model-based machine learning method, *Prescience*, that predicts the near-term risk of hypoxemia during surgery and explains the patient and surgery specific factors that led to that risk.

Our Work. Our proposed work to model the evolution of chronic pain relief has as its main characteristics:

- The performance presented to predict the evolution of chronic pain is better than the typical performance associated with pain experts.
- The data analyzed is considerably high dimensional, with each patient containing almost 400 features.
- The generated model is simple and interpretable, allowing the results to be of great help to clinical doctors, avoiding a black box approach.
- The contribution of each feature to the model's decision is provided, allowing a focused and faster intervention.

2.2 Learning Models from Data

Learning models from high-dimensional data is a well-studied problem in various fields. Our work builds upon a wealth of previous research at the intersection of feature selection, feature decomposition, and ensemble learning.

Feature Selection. One common approach to address high-dimensional data is to reduce the number of features using only the most important ones. Methods for feature selection can be grouped in three types: filter, wrapper, and embedded. Filter-based methods select features regardless of the model, while wrapper methods rely on the

specific model being used. The problem with filter methods is that a subset of highly correlated features can dominate the selection. Wrapper methods' downside is that they typically need to test all possible combinations of features, being thus very expensive, especially if the feature set is large. Finally, embedded methods assume that the learning algorithm employs feature selection interleaved with learning. In tree-based ensemble algorithms [6, 8], for instance, each feature is evaluated as a potential splitting variable, which makes them robust to unimportant/irrelevant features, as such features that are not discriminative will not be selected as the splitting variable and hence will be associated with a low importance value. However, feature selection methods are not well suited for modeling phenomena defined by multiple local structures [26], because the model should not include features that do not influence the dependent variable, and correlations between features and the dependent variable are likely to vary in different local structures that exist in the data. That is, the dependent variable is actually affected by most of the features, but the strength of the observed correlations may vary, depending on the local structures in the data space. In this case, simply removing features will cause a significant loss of important information.

Feature Decomposition. While in feature selection the aim is to identify a representative set of features from which to build a model, in feature decomposition the aim is to decompose the original set of features into several subsets. Specifically, feature decomposition changes the representation of a learning problem, depending on the local structures in the data space. That is, instead of learning a single complex model, several sub-problems with different and smaller feature sets are defined [7]. Another feature decomposition approach is biclustering [9], which is a class of clustering algorithms that can group features and data points simultaneously. Formally, the goal is to find local structures in the data space defined as subsets of data points in which a specific subset of features are highly correlated. Thus, a subset of features that are highly correlated within a local structure may be independent features within other regions of the data space. Oliveira and Madeira [32] present a valuable survey on the subject.

Ensemble Learning. Ensemble methods are learning algorithms that construct a set of models and then classify new data points by taking a (weighted) vote of their predictions. A particular relevant work is Reference [33], where the authors presented BENCH, a method to construct an ensemble using biclusters. Their approach first divides the original data space into biclusters, and each bicluster becomes a candidate dataset for constructing a base model in the ensemble. Although biclustering is an unsupervised technique, BENCH takes into consideration labels and their correlation with the cluster, with the goal of forming candidate datasets that distinguish labels well.

Interpretability on tabular data. Several attempts to provide meaningful explanations of machine learning model decisions were proposed in past decades. However, these are mostly model-specific [3, 27, 40, 43]. More recently, methods that provide model-agnostic local explanations have received attention, including **Local Interpretable Model-agnostic Explanation (LIME)** [37], **SHapley Additive exPlanation (SHAP)** [23], and Anchors [39]. LIME is a model-agnostic method that induces a local model capable of locally explaining an instance while being simple enough to be interpretable. Local interpretable methods aim at explaining individual predictions, based on the intuition that while the entire model can be highly complex to be explained by a simpler model, it is possible explain a single prediction and its vicinity. In particular, a systematic comparison of these three interpretation methods in healthcare is presented in Reference [11], which concluded that no method emerged as a clear winner. In this work, we are interested in explanations that describe the local behavior using a linearly weighted combination of the input features. Anchor explanations are based on if-then rules, thus not fitting our objective. SHAP is an additive feature attribution method that determines additive feature attributions by finding a single unique solution with three desirable properties of local accuracy, missingness, and consistency.

Our Work. Intuitively, for the local models to be able to produce an improved ensemble, they must make correct predictions on different subsets of the data space. We exploit a different notion of diversity that is given in terms of the explanatory factors within each local model. Further, by learning local models composed of different feature sets, we can achieve feature decompositions that feature selection algorithms cannot. For instance, we show that our proposed approach is able to model phenomena that exhibit "backbone structures," which are a type of local structure induced by a specific feature that, once set, causes the remainder of the features to decompose into independent subsets in the data space.

3 MODELING BACKBONE-STRUCTURE PHENOMENA

In this section, we present a novel approach for modeling phenomena that are defined by multiple local structures in the data space, which we define as backbone structures. Formally, we want to learn a model from data that is a mixture of sub-populations where each sub-population is associated with a particular subset of features. The corresponding optimization problem has a non-convex error surface with no obvious global minimum, thus implying in a multiplicity of performant models, each of them providing a different explanation for the phenomenon. Therefore, there may be many contrasting interpretations or competing explanations for the same phenomenon. The modeling approach we will describe in this section is based on finding a model for the phenomenon that is coherent with all competing explanations. In particular, we propose to decompose the original set of features into several subsets, so a particular model is built for each subset of features. Then, the generated models are clustered according to their explanatory factors, promoting diversity in terms of possible explanations while learning the ensemble. We expect that the final ensemble model corresponds to a more general, global explanation for the phenomenon, leading to improved prediction accuracy.

3.1 Local Structures

A data space is defined as a set of n data points of the form $(\mathbf{x}, y)^n$, such that $\mathbf{x} \in \mathcal{R}^d$ is given as a feature vector $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d\}$ and y is the ground-truth output for $\mathbf{x}$. Often, in high-dimensional data spaces, there exist regions that show complex correlations among a specific set of features and the target label, and the same correlations are not necessarily so strongly observed in other regions of the data space. Thus, the theory–data relationship varies in different regions of the data space, hence forming local structures defined as subspaces spanned by a set of data points and a set of features [44, 45], as illustrated in Figure 1 (Left). Local structures can overlap and are often the result of mixing different sub-populations or distributions into the same data space, and thus one cannot easily separate them into multiple sub-spaces. A particular type of local structure resembles a backbone, in the sense that there is a set of features (a.k.a. backbone features) that show strong correlation with a specific set of features. Thus, forcing a backbone feature to appear in the same model with non-related features may incur in confounding situations.

3.2 Sampling the Model Space

Learning a model from the data space requires the minimization of an objective function $f(\mathbf{x})$. Instead of simply mixing multiple different structures into a single model $\mathbf{x}$ and minimizing $f(\mathbf{x})$, we sample the model space by minimizing different functions $f(\mathbf{x}')$, such that $\mathbf{x}' \subseteq \mathbf{x}$ and $|\mathbf{x}'| \ll |\mathbf{x}|$. Features that compose each model $\mathbf{x}'$ are randomly selected, and we used gradient boosted trees [8] and random forests [6] as learning algorithms (but other algorithms can be applied as well). After sampling the model space, each model $\mathbf{x}'$ is evaluated with respect to an error measure $\ell(x')$ on a validation set, so only minimally performant models for which $\ell(\mathbf{x}') \leq \epsilon$ are included in the final model space $\mathcal{H}'$. At this point, we expect that $\mathcal{H}'$ contains performant models corresponding to possible explanations for the phenomenon.

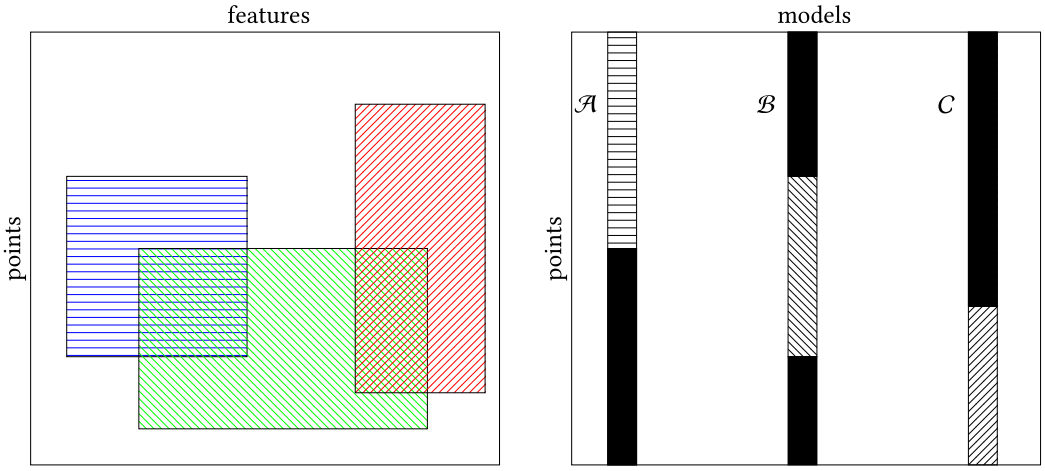

Fig. 1. (Color online) Left – An illustrative example of a data space with three local structures. For simplicity and to avoid clutter, local structures are shown as contiguous regions in the data space, but local structures may be non-contiguous in both axes. Right – An illustrative example of model preferences for three hypothetical models $\mathcal{A}$, $\mathcal{B}$, and C in $\mathcal{H}'$. Probabilities that a model assigns to points in the lighter region are closer to true label than the probabilities that the model assigns to points in the darker regions. Model $\mathcal{A}$ shows preference for points in the blue structure. Model $\mathcal{B}$ shows preference for points in the red structure. Model C shows preference for points in the green structure. A model may also show preference for points within multiple structures simultaneously.

3.3 Representing Model Preferences

We represent the model preference as a n-dimensional vector $\mathcal{P}(\mathbf{x}') = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n\}$, where $\mathbf{p}_i$ corresponds to the probability that model $\mathbf{x}'$ has assigned to data point i. We expect that models in $\mathcal{H}'$ are representative of the diverse local structures that exist in the data space, as illustrated in Figure 1 (Right). By filtering performant models for which $\ell(\mathbf{x}') \leq \epsilon$, we expect that the corresponding local structure is properly explained by the corresponding model $\mathbf{x}'$.

3.4 Representing Model Explanations

We represent how model $\mathbf{x}'$ explains the phenomenon as a d-dimensional vector $E(\mathbf{x}') = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_d\}$ showing which features are driving the model's prediction. Specifically, $\mathbf{e}_i$ takes a value that corresponds to the influence that the respective feature $\mathbf{x}_i$ had on the model decision. Since we do not assume feature independence while minimizing $f(\mathbf{x}')$, then correlated features within model $\mathbf{x}'$ should share credit or importance. For this reason, we employ the average **SHAP (SHapley Additive exPlanations)** [23] values for assessing feature importance.

SHapley Additive exPlanations. Given an instance x, SHAP provides us the weighted contribution of each feature in the outcome. As with LIME [37], SHAP focuses on *local methods*. It aims to learn how the model behaves in the vicinity of an instance x. This behavior, however, may not truly represent the original model f in entire data space. After all, only the model itself reliably describes the behavior globally. SHAP uses an *explanation model* g that is capable of locally represent f. The explanation model is described as:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i, \tag{1}$$

where $z' \in \{0,1\}^M$, M is the number of simplified input features, and $\phi_i \in \mathbb{R}$ is the feature contribution.

Lundberg and Lee (2017) describe SHAP as a permutation-based approach for feature importance attribution that defines a model as a cooperation of features, and it assigns a value for each feature in the cooperation based on its contribution to the model decisions. There are many other feature attribution methods [37, 38], but SHAP is the only method with the three desirable properties:

- Local accuracy: The explanations are truthfully explaining the model.
- Missingness: Missing features have no attributed impact to the model decisions.
- Consistency: If a model changes so some feature's contribution increases or stays the same regardless of the other features, then that feature's attribution should not decrease.

Options to calculate the explanation model g include LinearSHAP, KernelSHAP, DeepSHAP, and TreeSHAP [25]. In this work, we used TreeSHAP, since our learning algorithms are based on gradient boosted trees or random forests. TreeSHAP takes only "allowed" paths within the tree, meaning it does not include non-realistic combinations of features as in other permutation-based methods. Instead, it takes the weighted average of all the final nodes that were reachable by a certain coalition of features. Differently from the other options, TreeSHAP scales linearly with the number of data points and grows at a polynomial rate with the number of features.

3.5 Ensemble Learning

As $\mathcal{H}'$ may contain models with competing explanations, we want to build a synthetic model from $\mathcal{H}'$ by exploiting two concepts:

- The concept of diversity between individual models. Diversity is recognized as a central element to get significant performance improvements with the ensemble [21] by allowing the group to compensate individual errors and reach a better expected performance. However, measuring diversity is not straightforward, because there is no generally accepted formal definition. To promote diversity while learning the ensemble, we cluster models in $\mathcal{H}'$ based on the distance between their explanation vectors (i.e., SHAP values). Ideally, this creates a number of groups of models that are internally dense and also separated from the rest of the models in terms of their explanatory factors, that is, within each cluster the explanatory factors are similar, while factors within disjoint clusters are dissimilar.
- The concept of stability between model explanation and empirical predictions [42]. We define a configuration of clusters as stable if models within the same cluster are associated with the same explanatory factors and perform similar predictions. Achieving cluster stability is challenging, as models that perform similar predictions can be associated with different explanatory factors. To promote stability while learning the ensemble, we cluster the model space based on the distance between the explanation vector (i.e., SHAP values) associated with each model. However, we maximize cohesion and separation of the clusters based on the distance in terms of model preference. This enforces a stable configuration of clusters containing models that are similar both in terms of their predictions and explanatory factors.

Once clusters are found, we select a prototype model within each cluster, so we have as many prototype models as clusters. In particular, we select the most performant model within each cluster to maximize the performance of the ensemble. The intuition is that the phenomenon may have many explanations, and each prototype model is a possible explanation. To create the ensemble, we adopted the simplest combination in which each prototype model is given a weighted vote (i.e., validation error), and the label with the most votes is the prediction of the ensemble.

Figure 2 presents an overview of the framework proposed. The novelty presented in our approach is that we promote diversity in the creation of an ensemble based on competing explanations, showing superior performance when compared to other diversity metrics explored so far. The clusters formed proved to be coherent and concise, generating a stability in prediction-explanation.

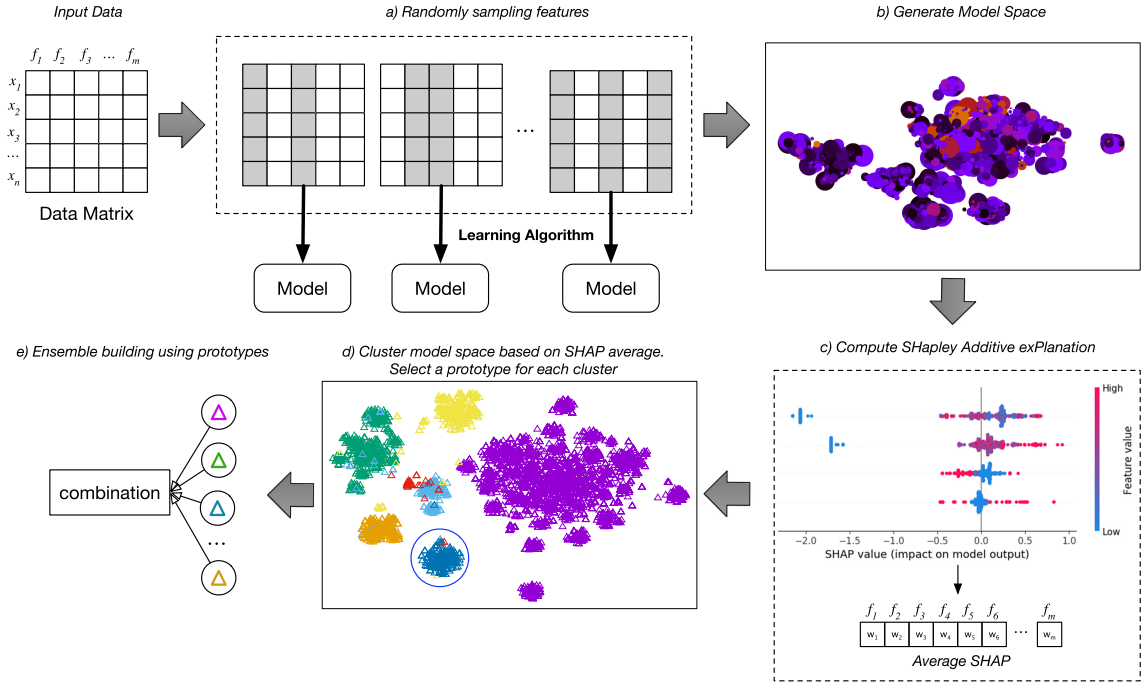

Fig. 2. (Color online) An overview of the framework. The framework starts with an input tabular matrix $n \times m$, being n the number of instances and m the number of features. (a) Randomly sample features with the size from 1 to f. For each set of features sampled use the learning algorithm to induce a model; (b) The set of all generated models will compose the model space; (c) For each model in the model space, compute the average of SHapley Additive exPlanations (SHAP) values; (d) Group the model space using as criterion the SHAP average. Select a prototype to represent each grouping. We will choose the model with the highest AUC of the grouping; (e) Finally, build an ensemble using as base models the prototypes previously selected.

4 PREDICTING THE EVOLUTION OF PAIN RELIEF

Pain makes us aware that something is wrong with our body. The **International Association for the Study of Pain (IASP)**[2] defines pain as "an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage." If the pain lasts beyond the time expected for healing following surgery, trauma or other condition, then it might be characterized as chronic. There is no accepted universal paradigm for chronic pain prevention and management [30]. Chronic pain is a public health concern that affects 20%–30% of the population of Western countries [5, 30]. Chronic pain usually falls into one of the following categories:

- Neuropathic pain occurs when there is actual nerve damage. Nerves connect the spinal cord to the rest of the body and allow the brain to communicate with the skin, muscles, and internal organs. Nutritional imbalance, alcoholism, toxins, infections, or auto-immunity can all damage this pathway and cause pain. Neuropathic pain can also be caused by a cancer tumor pressing on a nerve or a group of nerves. People often describe this pain as a burning or heavy sensation, or numbness along the path of the affected nerve.
- Nociceptive pain is caused by damage to body tissue and usually described as a sharp, aching, or throbbing pain. This kind of pain can be due to benign pathology; or by tumors or cancer cells that are growing

[2] http://www.iasp-pain.org.

larger and crowding other body parts near the cancer site. Nociceptive pain may also be caused by cancer spreading to the bones, muscles, or joints, or that causes the blockage of an organ or blood vessels.
- Nociplastic pain arises from altered nociceptive function despite no clear evidence of actual or threatened tissue damage causing the activation of peripheral nociceptors or evidence for disease lesion of the somatosensory system causing the pain.

Although there have been many scientific advances in the understanding of the neurophysiology of pain, precisely defining the best therapy for a patient is still a challenge. There do not currently exist validated classical frameworks for pain treatment. Chronic pain is an individualized experience with multifactorial aetiology, and understanding the biological, social, physical, and psychological contexts is vital to successful treatment. Standardized self-reported instruments and questionnaires to evaluate the patient's pain intensity, functional abilities, beliefs and expectations, and emotional distress are available and can be used to assist in treatment planning.

Pain Data. Pain is often assessed on a scale from "no pain" to "worst pain imaginable." The Visual Analogue Scale [18], or simply VAS, is a 10-cm line without markings from no pain to worst pain. Patients mark their pain score and a measurement in centimeters defines their level of pain. Six-hundred thirty-one participants self-completed the McGill Pain Questionnaire [29] and the **Visual Analogue Scale (VAS)**. The McGill Questionnaire assesses both quality and intensity of the pain. In summary, the questionnaire is composed of 78 words, of which respondents choose those that best describe their experience of pain (multiple markings are allowed). The words are organized in three dimensions:

- Sensory dimension: encompasses both the quality and severity of pain in terms of its temporal, spatial, pressure, and thermal properties.
- Affective dimension: refers to feelings and sentiments in the presence of pain, that is, how the patient feels emotionally as a result of pain.
- Evaluative dimension: refers to the global evaluation of the situation experienced by the patient and is strongly influenced by previous painful experiences. It is a subjective assessment of overall pain intensity.

As a result, our pain data includes variables regarding pain severity, change in pain relief over time, pain radiation, among others. Data was also collected via self-report on socioeconomic status (i.e., age, sex), global rating of overall health, known risk factors (i.e., age, smoking, alcohol intake) and concomitant illnesses. Finally, our pain data also includes the therapies prescribed by the doctor. In all, our pain data is composed of 338 variables about pain relief, socioeconomic status, and prescribed treatments. As a pre-processing step, we have opted to transform categorical features into sets of "dummy" features. We believe that this pre-processing has a positive impact on the generation of models.

Predicting the Evolution of Pain Relief. Satisfactory treatment can only come from comprehensive assessment of the biological aetiology of the pain in conjunction with the patient's specific psychosocial and behavioral presentation. The first consultation is potentially a pivotal event in a patient's pain history, affecting treatment adherence and engagement with longer term self-management [50]. The objective of our retrospective study is to predict, using data obtained only at the first consultation, if a certain treatment or therapy will be effective in reducing the patient's pain relief. We evaluated three distintic measures to identify a success in patient's pain relief:

- An overall reduction of pain intensity by 30% (a.k.a. VAS 30), which is formally considered to be a successful treatment outcome [10]. The ground truth labels are obtained by calculating the difference of pain intensities reported in the first and in the last consultation.
- An overall reduction of pain intensity by 50% (a.k.a. VAS 50). Again, the ground truth labels are obtained by calculating the difference of pain intensities reported in the first and in the last consultation.

- The Global Impact Change (a.k.a. GIC) as a discrete variance scale from −3 to 3 provided by the doctor indicating the degree of improvement in pain relief in the doctor's view. Success is given as a value of at least 2 in the last consultation.

We calculated the pair-wise correlation between GIC, VAS 30, and VAS 50. We observed that VAS 30 and VAS 50 are highly correlated, achieving a correlation value as high as 0.85. However, GIC is not highly correlated with both VAS 30 and VAS 50, showing correlation values of 0.1 and 0.097, respectively. This indicates that ratings from patients and assessment from clinicians may have discrepancies. Besides, when a patient achieve an overall reduction of pain intensity by 30%, most of the time it will also reach an overall reduction of 50%.

It is important to mention that while the type of pain has clear importance for diagnosis and treatment, the response to treatment is more a personal feature of the patients and it is less related to the type of pain itself. In our study, patients received treatment according their main pain syndrome, as part of their usual care, but the etiology of the pain was not a main factor and is not directly considered in our models.

Characterization based on VAS 30. As a successful treatment outcome can be formally given by VAS 30, the analysis presented throughout this section specifically refers to the VAS 30 label. Considering this label, we divided the 631 patients in our study into two populations:

- Population A: 277 patients for whom the treatment resulted in a significant reduction in pain relief, that is, these patients reported a significant +30% reduction in pain relief after the treatment is completed.
- Population B: 354 patients for whom the treatment was not effective.

Table 1 shows characteristics of the patients in our dataset. Pain is more prevalent in women, and it is harder to achieve significant pain reduction in patients that report low initial pain intensities. The table also shows the three dimensions of pain perception. Pain perceptions may overlap within the same dimension, and a total score for each dimension is given by summing up all types of pain perceptions. Similarly, the McGill score is given as the summation of the values associated with all words marked by the patient. Table 1 also shows the neuropathic pain scale that is used for assessing neuropathy pain and may be particularly useful for assessing response to therapies. The total neuropathy score is calculated as the sum of the possibilities and the cut-off value for the diagnosis of neuropathic pain is a total score of 4. The table also shows information about pain outbreaks and the time in which pain is worse. There are other variables that we omitted from the table to avoid clutter.

Finally, Figure 3 shows how often pain is reported in different areas of the human body. Interestingly, areas on the right side of the body are more frequently reported by patients in population B. The considered features enable myriad possibilities of combining diverse aspects about pain relief while learning predictive models.

5 EVALUATION

Next, we discuss our evaluation procedure and then we report our results. In particular, our experiments aim to answer the following research questions:

RQ1: Is there a relationship between model explanation and model preferences?
RQ2: Are prototype models diverse in terms of explanatory factors?
RQ3: Can we build effective ensembles by combining models that are associated with diverse explanatory factors?
RQ4: Is our explanation-diversifying ensemble approach superior to the biclustering ensemble approach?

Setup. While sampling the model space, we randomly set the number of features that compose each model, but we assure that no model has more than 15 features. There is a tradeoff when adjusting the maximum number of features. From our experiments, as we increase the maximum number of features allowed, we are able to obtain ensembles with better AUC values. Conversely, it is also expected to increase the computational cost and make

Table 1. Part of the Patient Data Obtained at the First Consultation

	Population A	Population B
N	277 (43.89%)	354 (56.11%)
Sex (male)	110 (39.71%)	151 (42.65%)
Age, y	54.86 (46–64)	56.66 (45–60)
0–15 McGill score	7.21 (4–10)	5.75 (3–9)
0–10 intial pain intensity	6.66 (5–8)	4.80 (2–8)
Sensory dimension	3.31 (1–5)	2.63 (1–4)
Burning	170 (61.4%)	188 (53.1%)
Painful	131 (47.3%)	139 (39.3%)
Slapped	113 (40.8%)	115 (32.5%)
Throbbing	111 (40.1%)	104 (29.4%)
Stabbings	104 (37.5%)	96 (27.1%)
Electric shocks	100 (36.1%)	99 (27.9%)
Sharp	95 (34.3%)	102 (28.8%)
Spreads	87 (31.4%)	86 (24.3%)
Affective dimension	2.59 (1–3)	2.13 (1–3)
Tiring	209 (75.4%)	227 (64.1%)
Nauseous	186 (67.1%)	191 (53.9%)
Annoying	157 (56.7%)	166 (46.9%)
Stifling	89 (32.1%)	91 (25.7%)
Scary	74 (26.7%)	79 (22.3%)
Evaluative dimension	1.30 (1–2)	0.99 (1–1)
Uncomfortable	260 (93.9%)	252 (71.2%)
Unbearable	100 (36.1%)	100 (28.2%)
Neuropathic pain scale		
Burning	193 (70.4%)	220 (62.7%)
Hypoesthesia to touch	143 (48.2%)	143 (40.7%)
Numbness	109 (39.8%)	101 (28.8%)
Pins and needles	107 (39.0%)	117 (33.3%)
Tingling	89 (32.5%)	97 (27.6%)
Electric shocks	85 (31.0%)	81 (23.1%)
Painful cold	46 (16.7%)	49 (14.0%)
Brushing	40 (14.6%)	37 (10.5%)
Duration of pain outbreaks		
Minutes	10 (3.6%)	17 (4.8%)
Hours	19 (6.9%)	16 (4.5%)
Days	2 (0.8%)	5 (1.4%)
Weeks	1 (0.4%)	3 (0.8%)
Months	6 (2.2%)	8 (2.3%)

Mean, first, and third quartiles within age, McGill score, initial pain intensity, and pain perception dimension scores. Population A refers to patients for whom the treatment was considered effective, that is, the patients reported a significant reduction in pain sensation after the treatment. Population B refers to patients for whom the treatment was not effective. Here, we consider the VAS 30 label. Pain characteristics are not mutually exclusive.

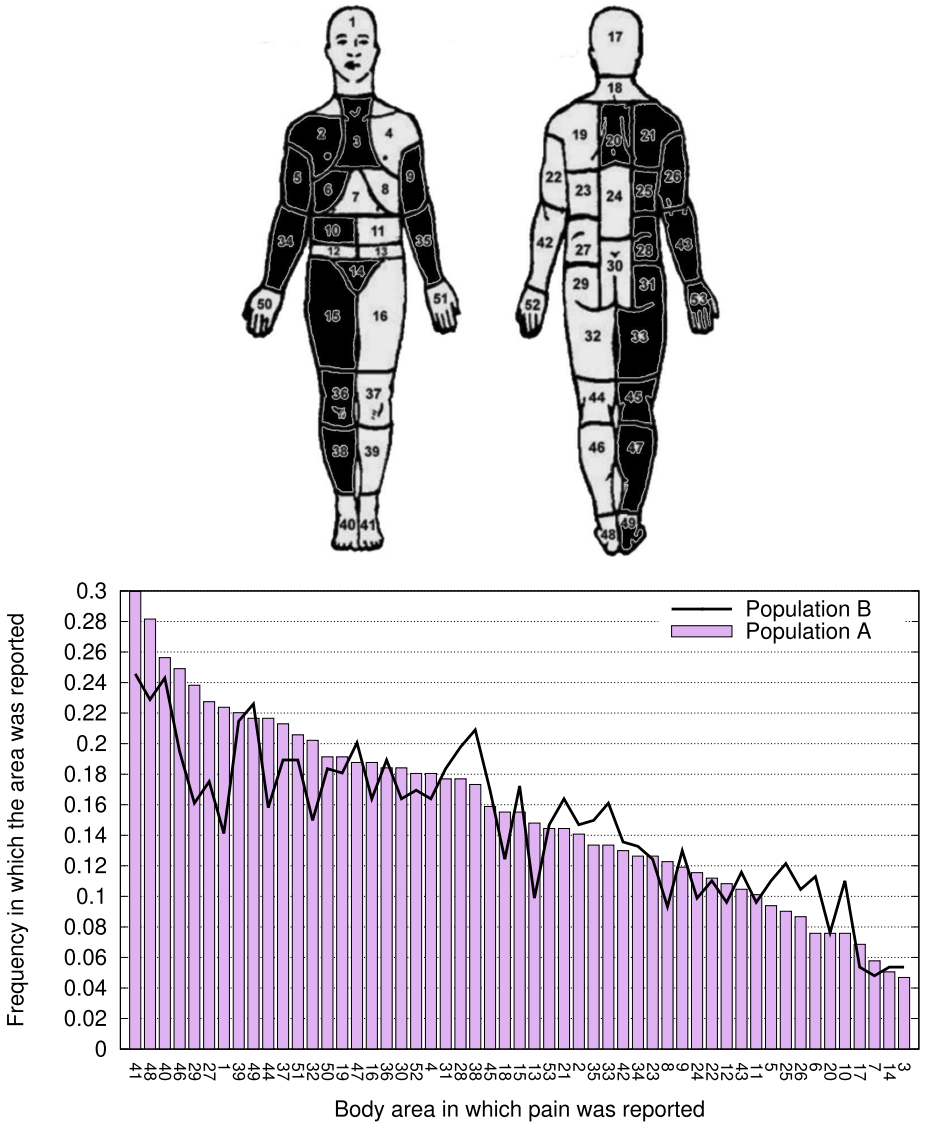

Fig. 3. (Color online) Areas associated with pain. Top – Areas in black are more frequently reported by patients in population B than by patients in population A. Bottom – Frequency in which each area is reported by patients in populations A and B.

the generated models harder to interpret. As interpretability is a crucial aspect of our work, we opted to setting the upper limit at 15 features as a good compromise between interpretability and performance. Using this upper limit, we are able to test validity and feasibility of our work. Further, it is intended to build a questionnaire to be applied by the doctor in the patient's first consultation and limiting the number of features is also one way to limit the number of questions.

The features that compose each model are also randomly selected. Models are built using SciKit-Learn implementations of XGBoost or Random Forests algorithms [34]. We sampled a total of 150,000 models using the

Table 2. Baselines, Separated by Label, with Average AUC Values Obtained by the All-in-one Approach and Carrying Out TPOT to Optimize Machine Learning Pipeline (Limiting Time by 24 Hours)

	XGBoost	Random Forests		TPOT
Label	AUC	AUC	Mean	AUC
VAS 30	0.648	0.652	0.650	0.632
VAS 50	0.634	0.597	0.615	0.598
GIC	0.564	0.575	0.569	0.568

XGBoost algorithm, and another 150,000 models using the Random Forests algorithm. To evaluate the performance of the models, we used the standard **AUC (area under the ROC curve)** measure [13, 17]. We conducted five-fold cross-validation, that is, data are arranged into five folds, and at each run, four folds are used as training set, and the remaining fold is used as test set. We also employed a separated validation set used to select the best models. We report the average AUC value over the five runs. This entire process was executed separately for each label, namely, VAS 30, VAS 50, and GIC.

Baseline Models. As baseline, we averaged AUC values by the all-in-one models and also carried out **Tree-based Pipeline Optimization Tool (TPOT).**[3] The first scenario represents the standard approach. The second scenario employs a tool that optimizes machine learning pipeline using genetic programming. We set up the time limit for optimization as 24 hours, once this is the approximate amount of time in our worst case to run our approach.

Feeding a XGBoost model with all features and using VAS 30 as label resulted in an average AUC of 0.648. With Random Forests, the value obtained was 0.652. Therefore, we consider a VAS 30 model as minimally performant if its average AUC value is at least equal to the AUC value of the all-in-one approach (in this case 0.648 for XGBoost model and 0.652 for Random Forests). While this performance threshold seems low, it greatly exceeds the estimated physician performance at the first consultation, which is no higher than 0.584. The close-to-random performance of physicians at the first consultation reveals how difficult is this predictive task. Table 2 presents the average values of AUC obtained using the all-in-one approach to all labels. It is worth mentioning the difficulty of predicting the GIC label. TPOT as a machine learning pipeline optimization tool automatically selects the learning algorithm. The average AUC obtained using VAS 30 as label was 0.632. In this case it was selected the XGBoost classifier. Using VAS 50 resulted in an average AUC of 0.598, through stacking up multiple estimators: Multinomial Naive Bayes, Gaussian Naive Bayes, and k-Nearest Neighbors Classifier. Finally, using GIC as label resulted in an average AUC of 0.568 through the stacking up of the following estimators: **Stochastic Gradient Descent (SGD)** and XGBoost.

The performance threshold resulted in a sampled VAS 30 model space $\mathcal{H}'$ for XGBoost and another space for Random Forests. The XGBoost VAS 30 model space is composed of 2,830 models out of the original 150,000 models (1.9% of the models perform better than the all-in-one model), while the Random Forests model space is composed of 2,507 models (1.7% of the models perform better than the all-in-one model). For the VAS 50 label, the number of sampled models for XGBoost was 1,408 models (0.94% of the models perform better than the all-in-one model), and 11,829 (7.89% of the models perform better than the all-in-one model) for Random Forests. Regarding GIC, 18,575 models (12.38% of the models perform better than the all-in-one models) are selected for XGBoost and 10,035 models for Random Forests (6.69% of the models perform better than the all-in-one model).

Figure 4 shows XGBoost and Random Forests model spaces for each label. In the figure, each point corresponds to a model, and the size of the point indicates the variance of the validation error. Thus, in the figure the best

[3] http://epistasislab.github.io/tpot/.

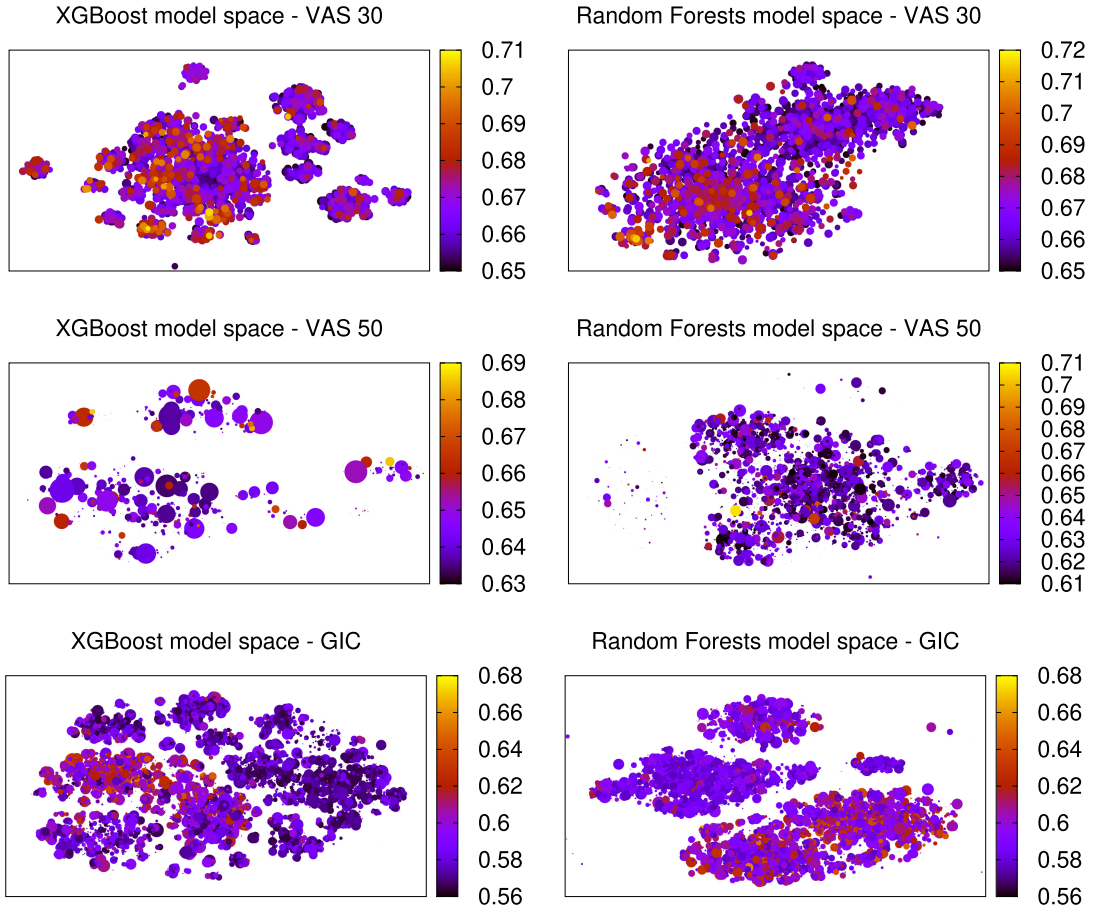

Fig. 4. (Color online) T-SNE visualization [47] of the sampled model space $\mathcal{H}'$. Each point represents a model $\mathbf{x}'$. Models are placed according to the probabilities of significant pain relief assigned to patients, so models that assign similar probabilities to the same patients are placed next to each other in the space (refer to Section 3.4). The color indicates the average (cross-validation) AUC value, and smaller points indicates that the corresponding model has less variance.

models are shown as clearer and smaller points. The figure shows that the best models are well scattered through the model space, indicating that there are models with different preferences but equally performant.

In our experiments, we observed that for the clustering algorithms to work correctly, at least 1,000 elements are needed in the explanation space. This number is obtained after filtering this space with an AUC threshold. There are two possible ways to change this amount: (a) decrease the threshold used for filtering or (b) increase the maximum number of features. The first alternative should be applied with caution, because a key feature for an ensemble to achieve good performance is that its base models must also necessarily have good performance [22]. The addition of models that do not have a good performance can have a negative impact on the ensemble performance. Therefore, it is more appealing to increase the maximum number of features.

5.1 Relating Model Preferences and Explanatory Factors

To answer RQ1, we embedded XGBoost and Random Forests models according to their model preferences (i.e., probabilities they assign to the data points), so models that assign similar probabilities to the same data points

are placed next to each other in the model space (as in Figure 4, but using the original high-dimensional space). Then, we clustered the model space using different criteria and clustering algorithms. More specifically, we employed Hierarchical clustering [49] and DBScan [12] algorithms. These two clustering algorithms represent two distinct ways to cluster data. Dendrogram is a tree-based clustering algorithm that partitions the given data rather than the entire instance space. However, DBSCan is a density-based clustering connecting points within certain distances thresholds, only when satisfying a density criterion.

All hyper-parameters were set by maximizing the silhouette value considering the model preference space. The silhouette value is a measure of how similar a data point is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the data point is well matched to its own cluster and poorly matched to neighboring clusters. If most data points have a high value, then the clustering configuration is appropriate. This measure was chosen because it is a standard metric to detect the poor quality of clustering. The silhouette value considered in this work is the mean silhouette value over all samples.

Figure 5 shows 2D T-SNE visualization for the XGBoost VAS 30 model space after being clustered using different criteria.[4] First, for the sake of comparison, we clustered the model space using the predictions performed by each model so models that perform the same predictions for the same data points are more likely to be associated with the same cluster. In this case, explanatory factors were not used. While hierarchical clustering leads to cohesion, it lacks performance in terms of separation. The opposite trend is observed for DBScan clusters. Both Hierarchical clustering and DBScan achieved low silhouette values. Specifically, Hierarchical clustering achieved a 0.17 silhouette value, while DBScan was able to achieve only a 0.01 silhouette value. The low silhouette values, especially for DBScan, may be due to similar probabilities that are associated with opposite predictions. That is, small differences between probabilities that are close to the threshold may lead to opposite predictions, thus models may be evaluated as similar in terms of model preference but different in terms of their predictions.

The figure also shows the XGBoost model space clustered using the indexes of the features within each model. This is also for the sake of comparison, and the intuition is to evaluate to what extent a specific set of features may be associated with a particular local structure in the data space. The problem with this clustering criterion (i.e., the features within the model) is that it neglects that different features may be correlated, and thus models may have similar preferences even if they are completely different in terms of their features. Again, this leads to poor clustering performance. Specifically, Hierarchical clustering achieved a 0.05 silhouette value, while DBScan achieved only a 0.03 silhouette value.

Finally, we evaluate our proposed approach of clustering the model space based on the explanatory factors associated with each model. As detailed in Section 3.4, we represent each model as a vector composed of the SHAP values associated with the factors explaining the model decisions. Interestingly, clustering based on explanatory factors results in groups with very high values of cohesion and separation, suggesting a strong link between model preferences and model explanation. Another possible advantage of clustering models as vectors of SHAP values is that the importance of each factor is divided if the model contains correlated features. In particular, our approach avoids a systematic instability in which models that are similar in terms of their preferences can have very different explanations. As a consequence, silhouette values are as high as 0.83 for Hierarchical clustering, and 0.95 for DBScan. In fact, the figure shows minor differences in the configuration of groups obtained by both clustering algorithms. As a result, our answer to RQ1 is positive considering the XGBoost model space, as the high silhouette values for both clustering algorithms indicate a relationship between model preferences and model explanation.

Similarly, Figure 6 shows 2D T-SNE visualization for the Random Forests model space after being clustered using different criteria. The same trend was observed. Again, for the sake of comparison, we clustered the model

[4]It is important to emphasize that T-SNE was only used for the sake of visualization, and all clusters were calculated in the original $m-$dimensional model space.

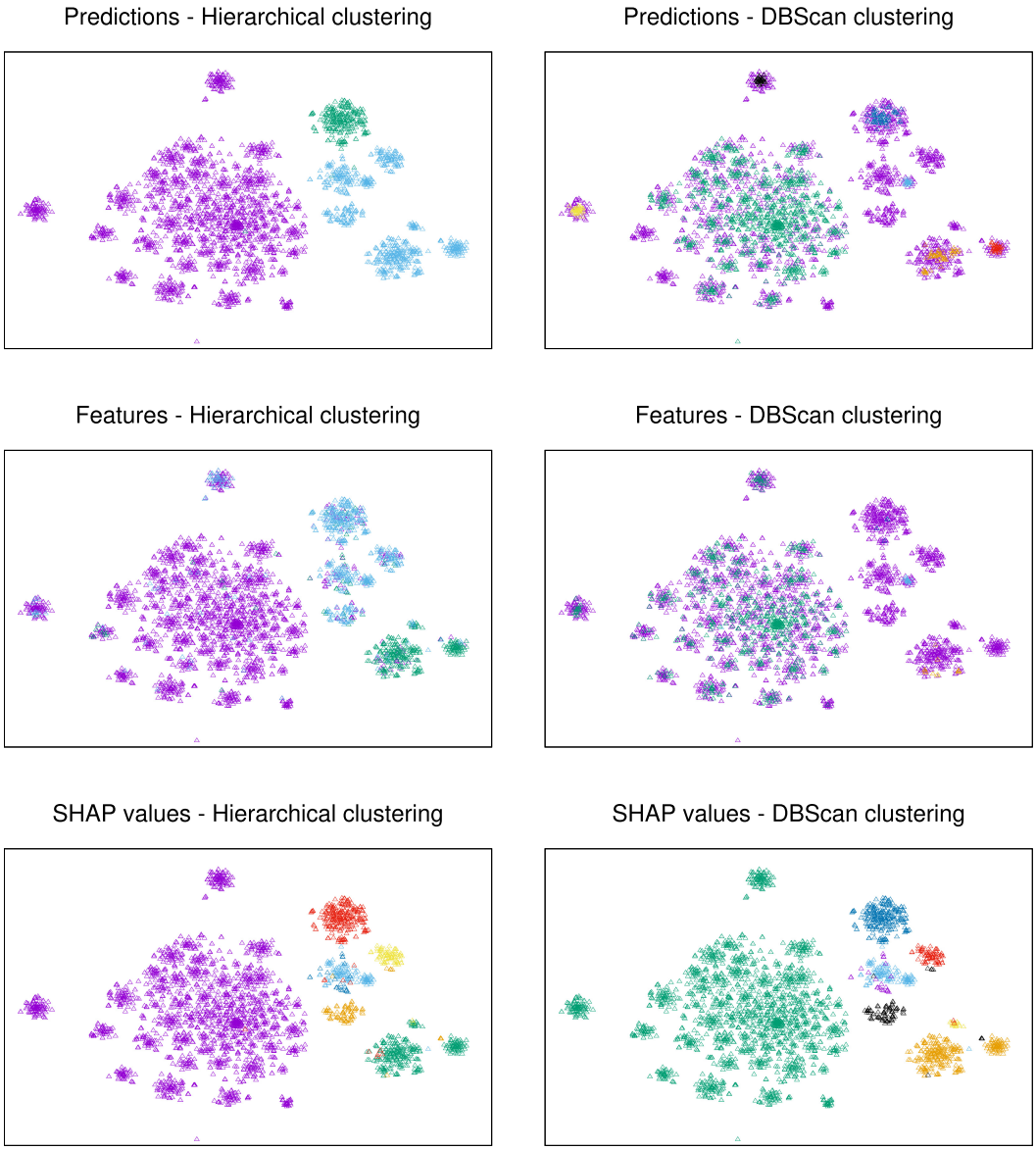

Fig. 5. (Color online) T-SNE visualization of the model space (VAS 30) after being clustered using different clustering algorithms and different criteria. Different colors mark different clusters. All clustering parameters were selected by maximizing the silhouette value in the model preference space. Models were built using Extreme Gradient Boosting (a.k.a. XGBoost). Silhouette values: Predictions = (0.17, 0.01); Features = (0.05, 0.03); SHAP values = (0.83, 0.95), respectively, for Dendrogram clustering and DBScan clustering.

space using the predictions performed by each model. In this case, again, both Hierarchical clustering and DBScan achieved low silhouette values. Specifically, Hierarchical clustering achieved a 0.06 silhouette value, while DBScan achieved a −0.33 silhouette value. Clustering the model space using the distance between the feature-sets within each model also leads to poor cohesion and poor separation. In this case, Hierarchical clustering achieved a 0.04 silhouette value, while DBScan achieved a 0.06 silhouette value. Again, clustering based on explanatory

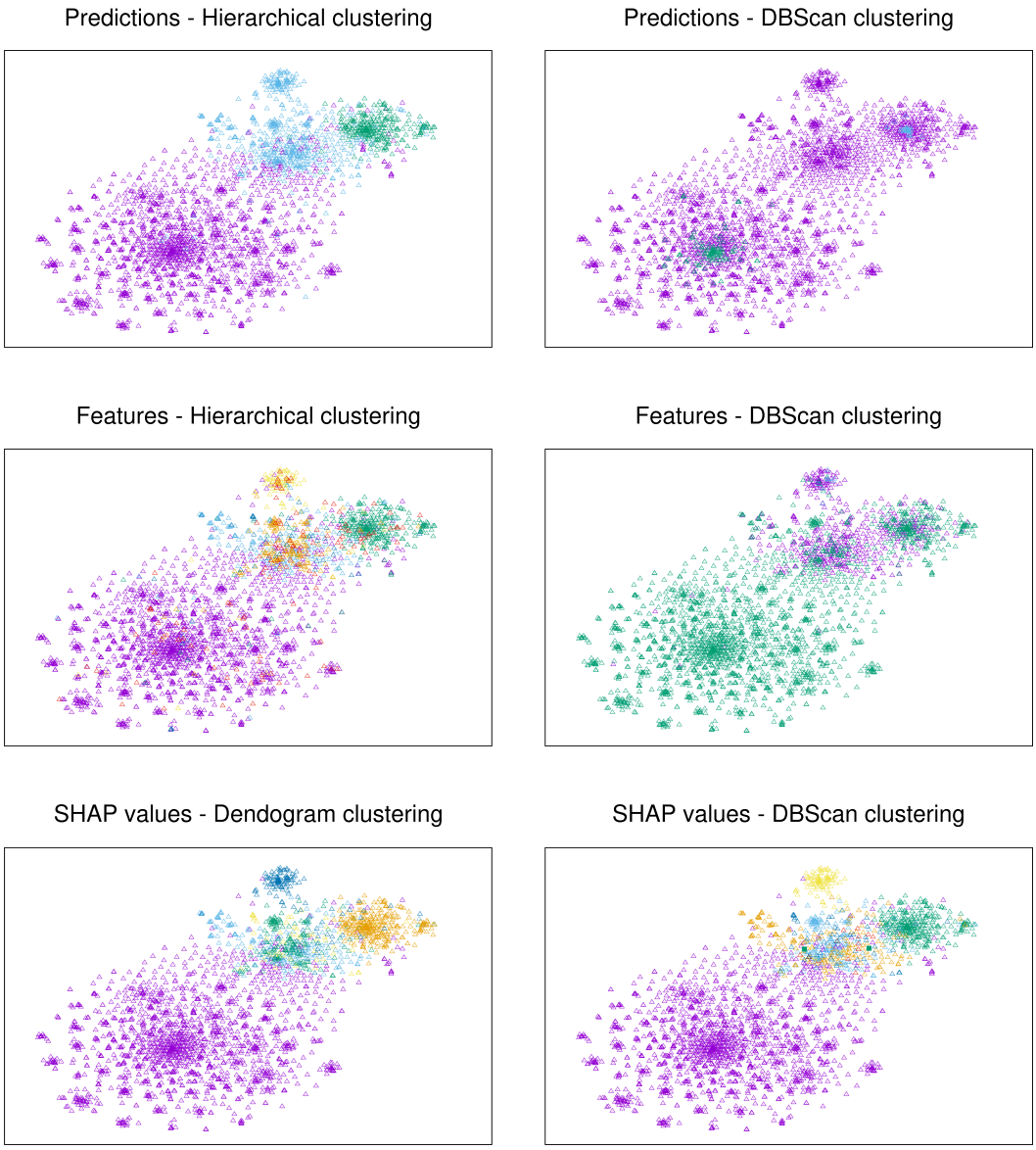

Fig. 6. (Color online) T-SNE visualization of the model space (VAS 30) after being clustered using different clustering algorithms and different criteria. Different colors mark different clusters. All clustering parameters were selected by maximizing the silhouette value in the model preference space. Models were built using Random Forests. Silhouette values: Predictions = (0.06, −0.33); Features = (0.04, 0.06); SHAP values = (0.87, 0.98), respectively, for Dendrogram clustering and DBScan clustering.

factors results in groups with very high values of cohesion and separation. Silhouette values are as high as 0.87 for Hierarchical clustering and 0.98 for DBScan. Therefore, our answer to RQ1 is also positive considering the Random Forests model space, as the high silhouette values for both clustering algorithms indicate a relationship between model preferences and model explanation.

Similar results were obtained also considering the VAS 50 and GIC labels. Specifically for VAS 50 with the XGBoost model, the silhouette score when clustering with feature criterion is 0.017 for hierarchical clustering and −0.021 for DBScan. For the probability criterion, the values of 0.089 and −0.288, respectively, are obtained with hierarchical clustering and DBScan. Finally, considering the SHAP criterion, good silhouetted values of 0.8115 and 0.95 are obtained, again showing the cohesion and separation obtained when we use explanatory factors as criterion. When we use the Random Forests model, the silhouette values follow the same pattern. For the features criterion, we get 0.0055 and −0.0076, for the probability criterion 0.0377 and −0.3741; last, the explanatory factor criterion, we get 0.7839 and 0.8895. Always the first value referring to the hierarchical clustering algorithm and the second to the DBScan clustering algorithm.

For GIC with XGBoost model, the silhouette score when clustering with feature criterion is 0.0055 for hierarchical clustering and −0.0175 for DBScan. For the probability criterion the values of 0.2311 and −0.4190, respectively, are obtained with hierarchical clustering and DBScan. Finally, considering the SHAP criterion, silhouette values of 0.3762 and 0.0064 are obtained. When we use the Random Forests model, the silhouette values follow the same pattern. For the features criterion, we get 0.01779 and 0.-0041, for the probability criterion 0.1592 and 0.2454; last, the explanatory factor criterion, we get 0.4027 and 0.2167.

Although silhouette values when grouping using explanatory factor criterion are at most 0.4027 (lower when compared to VAS 30 and VAS 50 labels), it is important to note that this exceeded other criteria for GIC. The lower silhouette value when compared to the other labels is probably due to the higher prediction difficulty for this label. However, notably, clusters with explanatory factors consistently show to be superior to the criterion of features and probabilities, regardless of the model and clustering algorithm used. Thus, we can conclude that RQ1 is also positive for the alternative labels VAS 50 and GIC.

5.2 Backbone Structure, Explanatory Factors, and Diversity

To answer RQ2, we inspected the prototype models within each cluster in the XGBoost and Random Forests model spaces. We focus on clusters based on explanation vectors produced by DBScan, as they will lead to the best ensemble results. Figure 7 shows two representative SHAP summary plots[5] associated with prototype models generated in VAS 30 model space, giving an overview of which features are most important for a model. We show only summary plots for XGBoost models to avoid clutter.

The first model in Figure 7 shows that a high initial pain intensity increases the chances of significant pain reduction at the end of the treatment. Typically, the most important feature within a model is a backbone feature, and then the model includes features that are somehow related to the backbone feature, such as a specific location or a particular medication.[6] As a result, models differ greatly in terms of their explanatory factors. Diversity becomes clear as we inspect the prototype models, as each model employs a set of features that is very different from the features used by the other prototype models. Specifically, within the eight XGBoost prototype models, there are a total of 42 distinct features, and only 5 features are present in two models. Thus, 37 features occur only in one prototype model. We also inspected the prototype models within each cluster in the Random Forests VAS 30 model space (although we do not show the corresponding summary plots). As with XGBoost, we focus on clusters based on explanation vectors produced by DBScan. Again, models differ greatly, depending on pain dimension, location of the pain, and predominance. Diversity is also observed in these prototype models. Specifically, there are a total of 46 distinct features within the 10 prototype models, from which 7 features are present in two models and only 2 features are present in three models.

[5]These summary plots show the SHAP values of every feature for every data point. In each plot, features are sorted by the sum of SHAP value magnitudes over all data points. The color represents the feature value.

[6]While a proper interpretation of model predictions is clearly important, our focus is on showing the diversity in terms of explanatory factors within each model, as this is the main intuition we will exploit for learning effective ensembles.

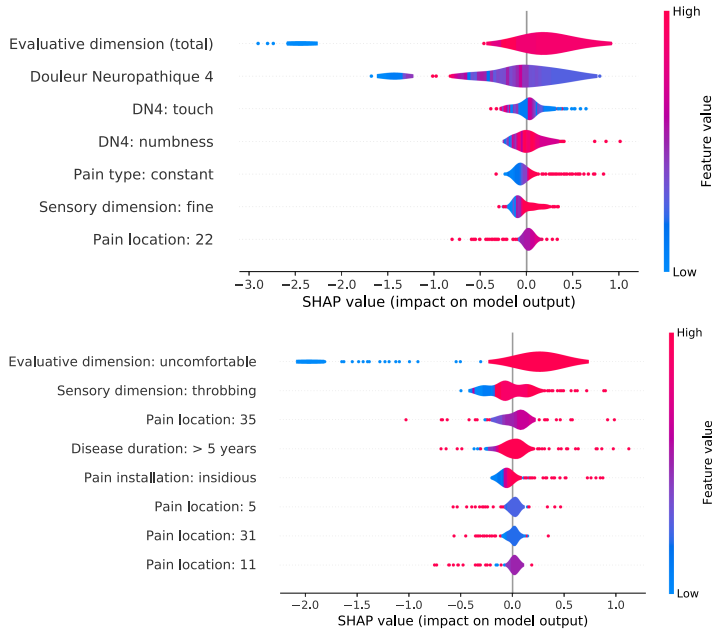

Fig. 7. (Color online) Explanation factors (viewed as SHAP summary plots) associated with prototype models for the VAS 30 label. Two of the eight prototype models built with XGBoost.

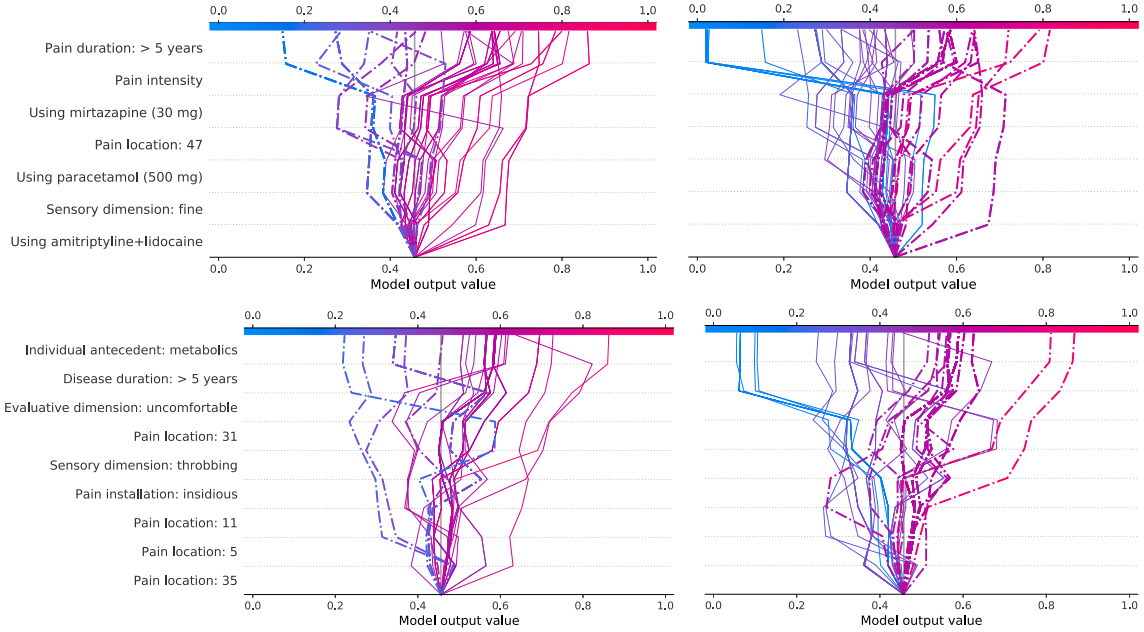

Fig. 8. (Color online) SHAP decision plots for the two models shown in Figure 7. Left – True positives (red) vs. false negatives (highlighted in blue). Right – True negatives (blue) vs. false positives (highlighted in red).

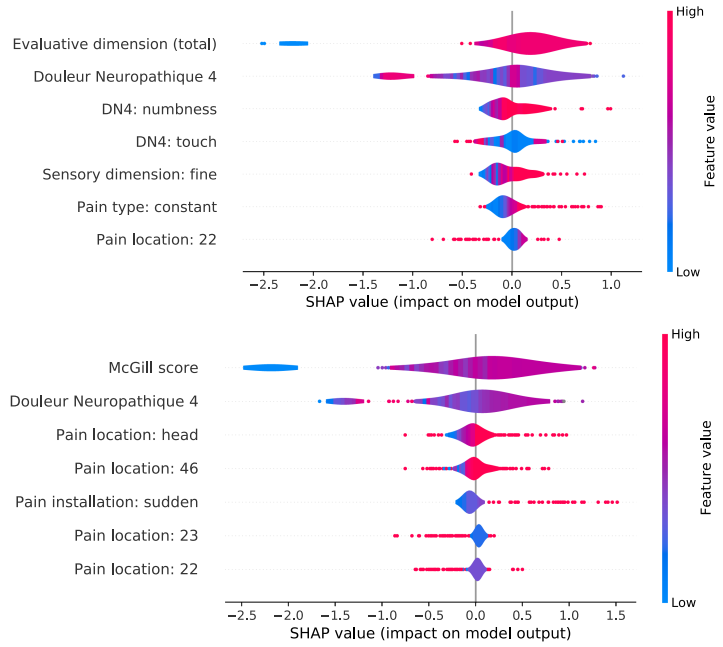

Fig. 9. (Color online) Explanation factors (viewed as SHAP summary plots) associated with prototype models for the VAS 50 label. Two of the nine prototype models built with XGBoost.

Figure 8 shows the SHAP decision plots of the prototype models previously shown in Figure 7. SHAP decision plots show how complex models arrive at their predictions. Each line presents a model decision path given an input. Each decision plot was generated for 50 inputs, to better represent the model behavior. This different view is effective mainly when many significants features are involved. As can be seen, models using different features are totally different ways to achieve the prediction.

Again, we observed the same trend for prototype models obtained for the VAS 50 and GIC labels. Figure 9 shows summary plots for two of the nine prototype XGBoost models using the VAS 50 label. In this case, the XGBoost prototypes comprise a total number of 91 features, of which 76 are unique features.

For Random Forests, the total number of features within the corresponding models is 155, from which 123 are unique and have occurred in only one model. Again, this is a strong indication that when grouped by explanatory factors, each prototype representing a cluster is diverse, a crucial strategy for building effective ensembles models. Figure 10 presents the SHAP decision plots for the two representatives models using VAS 50 label.

Figure 11 shows summary plots for the two prototype XGBoost models using the GIC label. In this case, the XGBoost prototypes comprise a total number of 13 features divided into two prototypes only. Interestingly, for Random Forests, 19 prototypes were generated, resulting in 203 features used, from which 143 were unique. Again, our ensemble strategy shows to employ diverse information while building the final model. Figure 12 shows the SHAP decision plots for the two representatives models using GIC label. Our answer to RQ2 is positive considering both the XGBoost and the Random Forests model spaces, as prototype models differ greatly in terms of the features being used.

As the models were selected by maximizing explanation diversity, the number of features that are shared among them is expected to be small. The most relevant features for each prototype can be obtained directly from their SHAP values (i.e., the higher the SHAP value, the more important is the feature). The most relevant set

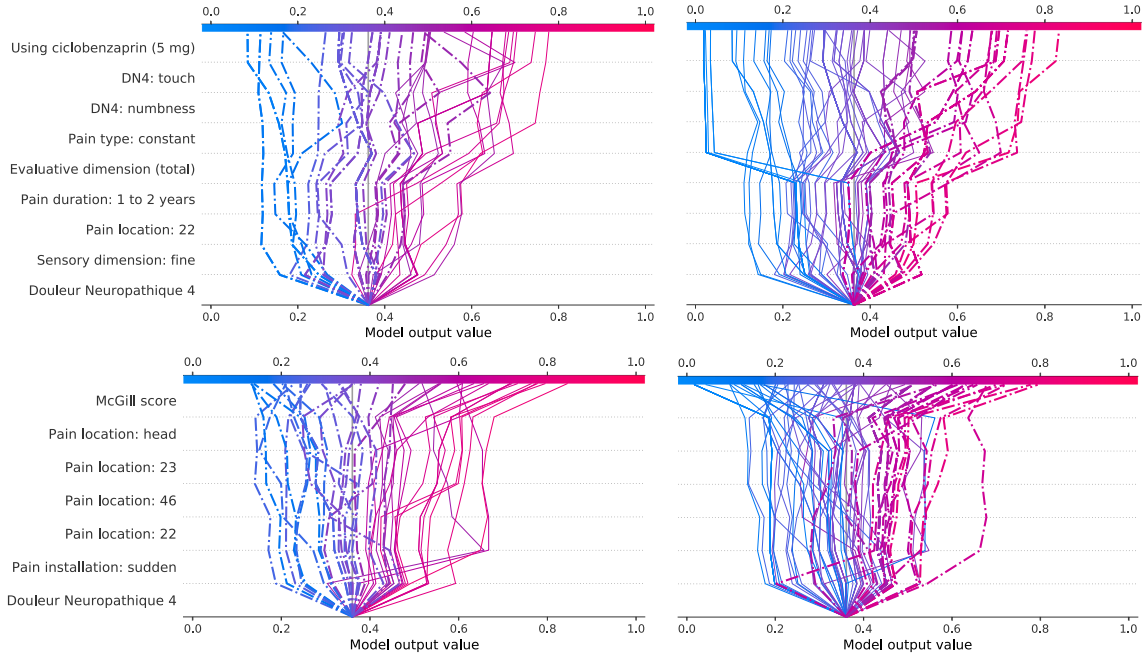

Fig. 10. (Color online) SHAP decision plots for the two models shown in Figure 9. Left – True positives (red) vs. false negatives (highlighted in blue). Right – True negatives (blue) vs. false positives (highlighted in red).

of features within the final model would be the combination of the most relevant features within its prototype models. For instance, using VAS30 as label, gradient boosted trees as learning algorithm, and DBScan as clustering algorithm, the set formed by the selection of the most relevant characteristic of each prototype model is formed by: pain intensity, evaluative dimension, affective dimension, DN4 quantitative, evaluative dimension uncomfortable, McGill, and sensitive dimension.

Next, we discuss how explanation diversity impacts the predictive performance of the ensemble.

5.3 Ensemble Performance

The next set of experiments is devoted to answer RQ3. Tables 3, 4, and 5 show AUC values for different ensemble configurations. The performance of the ensemble is compared with the performance of the best local model in the corresponding model space. Considering the VAS 30 label, different ensembles achieved AUC values that range from 0.68 to 0.78. Ensembles obtained from Random Forests prototype models provided gains that are up to 4.17%, although for some ensemble configurations the performance deteriorated. Ensembles obtained from XGBoost prototype models were more effective, providing gains that are up to 9.86% against the best model in the model space. If we compare the ensembles against all-in-one models, then the gains are up to 21%, a great improvement if we consider that fact that the models are also simpler.

For the VAS 50 and GIC labels, the obtained ensembles have always performed better than the best model within the corresponding model space. Although, for example, using the probability criterion and hierarchical clustering algorithm, it was possible to achieve a gain of up to 12.40% for the VAS 50 label using XGBoost (higher than the explanatory factors criterion), the result was not consistent. If we just change the model to Random Forests, instead of a big gain, then we have a considerable loss of −2.81 %. However, ensembles obtained by clustering according to model explanatory factors in all cases resulted in significant gains, even changing the

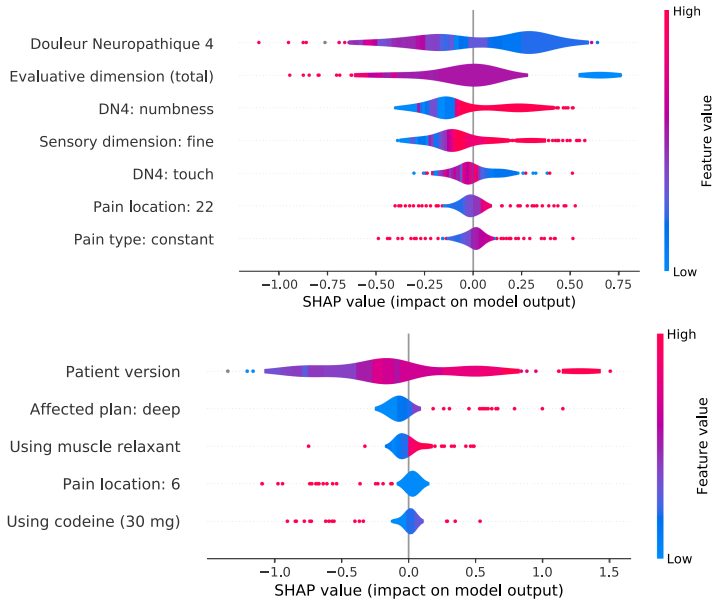

Fig. 11. (Color online) Explanation factors (viewed as SHAP summary plots) associated with prototype models for the GIC label. The two prototype models built with XGBoost.

Table 3. Ensemble Performance for Different Clustering Criteria and Clustering Algorithms Using VAS 30 as Label

		XGBoost			Random Forests		
Criterion	Clustering	AUC	Gain Best	Gain all-in-one	AUC	Gain Best	Gain all-in-one
Predictions	DBScan	0.73	2.82%	12.65%	0.68	−5.55%	4.29%
Predictions	Hierarchical	0.73	2.82%	12.65%	0.73	1.39%	11.96%
Feature values	DBScan	0.71	–	9.57%	0.70	−2.78%	7.36%
Feature values	Hierarchical	0.72	1.51%	11.11%	0.74	2.78%	13.50%
Explanations	DBScan	0.78	9.86%	20.37%	0.75	4.17%	15.03%
Explanations	Hierarchical	0.77	8.45%	18.83%	0.75	4.17%	15.03%

Baseline AUC values for XGBoost was 0.71 and for Random Forests 0.72.

label, the model, or the clustering algorithm. An improvement performance of up to 32% was achieved compared with all-in-one models.

Our proposed strategy of learning ensembles by clustering the model space using explanatory factors was always effective, providing significant gains despite the ensemble configuration. Thus, our answer to RQ3 is definitely positive.

5.4 Comparison with Physician and Biclustering Performance

Our last set of experiments is devoted to answer RQ4. For this, we consider the physician performance as being the known outcome of the last consultation. Further, as a strong baseline, we considered the **BENCH (Biclustering-driven ENsemble of Classifiers)** method proposed in Reference [33], which constructs an ensemble of classifiers through concurrent feature and data point selection guided by biclustering. Figure 13 shows

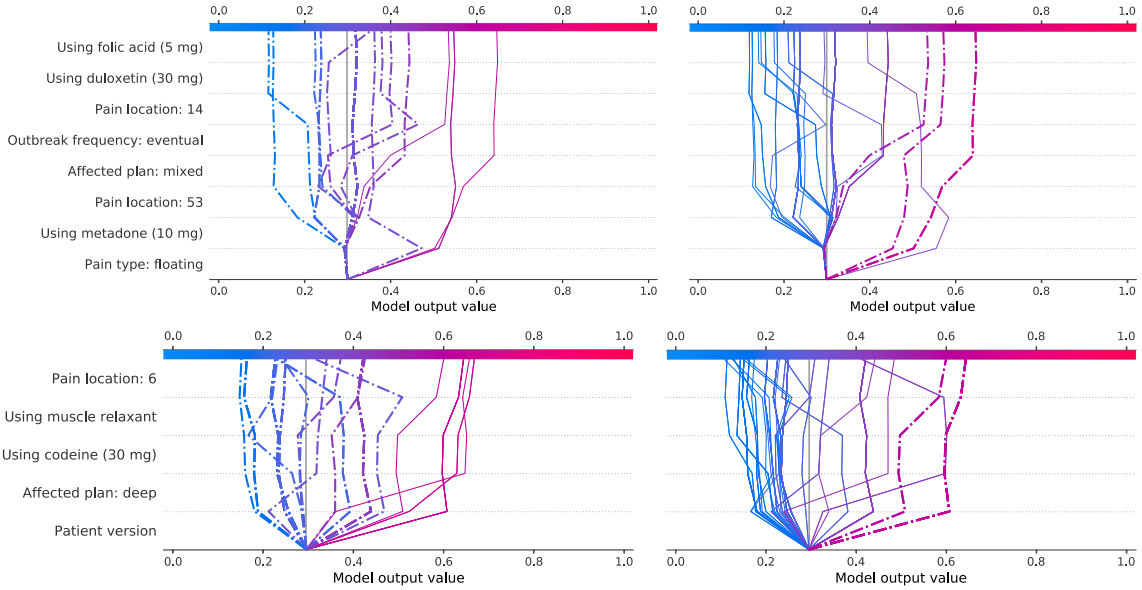

Fig. 12. (Color online) SHAP decision plots for the two models shown in Figure 11. Left – True positives (red) vs. false negatives (highlighted in blue). Right – True negatives (blue) vs. false positives (highlighted in red).

Table 4. Ensemble Performance for Different Clustering Criteria and Clustering Algorithms Using VAS 50 as Label

Criterion	Clustering	XGBoost AUC	XGBoost Gain Best	XGBoost Gain all-in-one	Random Forests AUC	Random Forests Gain Best	Random Forests Gain all-in-one
Predictions	DBScan	0.69	0.73%	8.83%	0.66	−7.04%	10.55%
Predictions	Hierarchical	0.79	15.33%	24.60%	0.72	1.41%	20.60%
Feature values	DBScan	0.73	6.57%	15.14%	0.69	−2.82%	15.58%
Feature values	Hierarchical	0.77	12.41%	21.45%	0.79	11.26%	32.33%
Explanations	DBScan	0.77	12.41%	21.45%	0.79	11.26%	32.33%
Explanations	Hierarchical	0.77	12.41%	21.45%	0.78	9.86%	30.65%

Baseline AUC values for XGBoost was 0.68 and for Random Forests 0.71.

ROC curves for BENCH, XGBoost+Explanations (i.e., ensemble of XGBoost models grouped by SHAP explanation factors) with DBScan, and Random Forests+Explanations with DBScan. Explanation-diversifying ensembles outperform BENCH in all ranges of false positive and true positive rates. We performed Welch's t-tests with $p = 0.01$, and both ensemble configurations are statistically different from BENCH, and thus our answer to RQ4 is also positive.

Limitations. Current limitation in framework is presented mainly in computing Shapley values. The two tree-based learning algorithms used in this work (Random Forests and Gradient Boosted Trees) rely in TreeSHAP. While KernelSHAP is a model-agnostic method to compute Shapley values, it can be slow and suffer from sampling variability. TreeSHAP, by focusing specifically on trees, computes local explanations based on exact Shapley values in polynomial time [25]. Using models that are KernelSHAP-dependent to computing local explanations are currently not viable. KernelSHAP is two orders of magnitude slower than TreeSHAP, making

Table 5. Ensemble Performance for Different Clustering Criteria and Clustering Algorithms Using GIC as Label

Criterion	Clustering	XGBoost AUC	XGBoost Gain Best	XGBoost Gain all-in-one	Random Forests AUC	Random Forests Gain Best	Random Forests Gain all-in-one
Predictions	DBScan	0.70	4.48%	24.11%	0.66	−2.97%	14.78%
Predictions	Hierarchical	0.70	4.48%	24.11%	0.72	5.88%	25.22%
Feature values	DBScan	0.65	−2.98%	15.24%	0.67	−1.47%	16.52%
Feature values	Hierarchical	0.68	1.49%	20.57%	0.69	1.47%	20.00%
Explanations	DBScan	0.68	1.49%	20.57%	0.76	11.76%	32.17%
Explanations	Hierarchical	0.71	5.97%	25.89%	0.74	8.82%	28.70%

Baseline AUC values for XGBoost was 0.67 and for Random Forests 0.68.

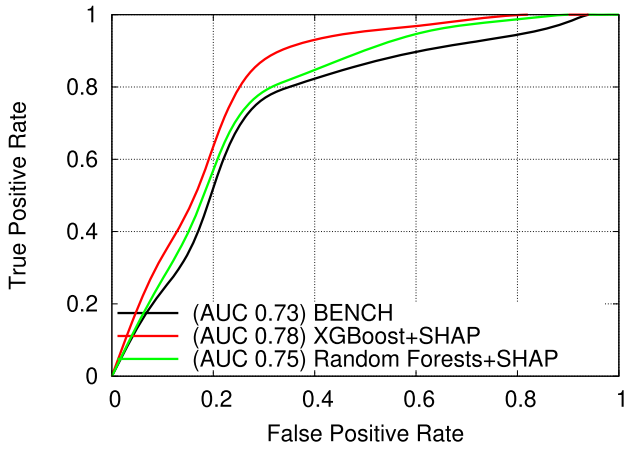

Fig. 13. (Color online) ROC curve comparing different ensemble approaches.

its use in the framework unfeasible. In addition, computational costs are also inherent to the framework, since approximately 150,000 models are generated in total. This requires the learning algorithm chosen should be capable of generating a trained model in a very short period of time.

A Note on Time Requirements. Although the time spent to generate the ensembles has no influence on the benefits already shown throughout the work such as increase of the AUC and reduced feature set, it can be a critical factor in the feasibility of applying the technique in miscellaneous use cases. Almost the entire runtime contribution comes from generating the model space with 150,000 sampled models. As we increase the maximum number of features allowed, we are able to obtain ensembles with improved performance. Conversely, it is also expected to increase the computational cost. As interpretability is a crucial aspect of our work, we opted to setting the upper limit to 15 features. The total time spent to sample the entire model space with upper limit of 15 features was 1,353.78 minutes with *XGBoost* and only 249.01 minutes with *Random Forests*. Both cases using VAS 30 as label. Generating the ensemble from the sampled model space takes less than 60 minutes for all configurations and learning algorithms. The specifications related to the hardware in which were executed the work are: Intel® Core™ i3-6100 CPU @ 3.70 GHz, 16 GB DDR3 1,600 MT/s and 256 GB SSD. As the algorithms used did not make use of the graphics card, we will omit the information.

6 CONCLUSIONS

In this article, we studied an underexplored link between explanatory modeling and predictive modeling, which leads to a novel approach for ensemble learning. Our proposed approach exploits two concepts: (i) local models that compose the ensemble should be diverse in terms of their explanatory factors, and (ii) candidate models should be organized by seeking stability in the sense that models that perform similar predictions should be also similar in terms of their explanatory factors. Another important contribution of our work is the evaluation of our ensemble learning approach in the task of predicting the evolution of pain relief in patients with unknown chronic pain conditions. This is a problem where typically exists a particular set of "backbone features" that, once set, causes the remainder of the features to decompose into different subsets in the data space. The backbone structure suggests that the problem is defined by multiple local structures, being thus a motivating example for our ensemble learning approach, which relates local structures and model explanations. Our experiments revealed that our proposed ensemble approach provides performance gains that are up to 11% when compared with the best local models, and it also significantly outperforms a biclustering approach, providing gains of 6.8%. When comparing with all-in-one approaches, the gains are up to 32%.

REFERENCES

[1] M. Abad-Grau, J. Ierache, C. Cervino, and P. Sebastiani. 2008. Evolution and challenges in the design of computational systems for triage assistance. *J. Biomed. Inform.* 41, 3 (2008), 432–441.

[2] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. 2005. Automatic subspace clustering of high dimensional data. *Data Mining Knowl. Discov.* 11, 1 (2005), 5–33.

[3] J. M. Benitez, J. L. Castro, and I. Requena. 1997. Are artificial neural networks black boxes? *IEEE Trans. Neural Netw.* 8, 5 (Sept. 1997), 1156–1164. DOI: https://doi.org/10.1109/72.623216

[4] L. Bernardes, M. Carvalho, S. Harnik, M. Teixeira, J. Ottolia, D. Castro, A. Veloso, R. Francisco, C. Listik, R. Galhardoni, V. da Silva, L. Moreira, A. de Amorim Filho, A. Fernandes, and D. Ciampi de Andrade. 2021. Sorting pain out of salience: Assessment of pain facial expressions in the human fetus. *Pain Rep.* 6, 1 (2021), e882.

[5] F. Blyth, L. March, A. Brnabic, L. Jorm, M. Williamson, and M. Cousins. 2001. Chronic pain in Australia: A prevalence study. *Pain* 89, 2-3 (2001), 127–134. DOI: https://doi.org/10.1016/S0304-3959(00)00355-9

[6] L. Breiman. 2001. Random forests. *Mach. Learn.* 45, 1 (2001), 5–32.

[7] M. Chen, K. Weinberger, and Y. Chen. 2011. Automatic feature decomposition for single view co-training. In *International Conference on Machine Learning*. 953–960.

[8] T. Chen and C. Guestrin. 2016. XGBoost: A scalable tree boosting system. In *International Conference on Knowledge Discovery and Data Mining*. 785–794.

[9] Y. Cheng and G. Church. 2000. Biclustering of expression data. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*. 93–103.

[10] R. Dworkin, D. Turk, K. Wyrwich, D. Beaton, D. Cleeland, J. Farrar, J. Haythornthwaite, M. Jensen, R. Kerns, D. Ader, N. Brandenburg, L. Burke, D. Cella, J. Chandler, P. Cowan P., R. Dimitrova, R. Dionne, S. Hertz, A. Jadad, N. Katz, H. Kehlet, L. Kramer, D. Manning, C. McCormick, M. McDermott, H. McQuay, S. Patel, L. Porter, S. Quessy, B. Rappaport, C. Rauschkolb, D. Revicki, and M. Rothman, 2008. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. *J. Pain* 9, 2 (2008), 105–121.

[11] Radwa Elshawi, Youssef Sherif, Mouaz Al-Mallah, and Sherif Sakr. 2019. Interpretability in healthcare a comparative study of local machine learning interpretability techniques. In *IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS'19)*. IEEE, 275–280.

[12] M. Ester, H. Kriegel, J. Sander, and X. Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *International Conference on Knowledge Discovery and Data Mining*. 226–231.

[13] T. Fawcett. 2006. An introduction to ROC analysis. *Pattern Recog. Lett.* 27, 8 (2006), 861–874.

[14] K. Ferreira, T. Bastos, D. Ciampi, A. Silva, J. Appolinario, M. Jacobsen, and M. Latorre. 2016. Prevalence of chronic pain in a metropolitan area of a developing country: A population-based study. *Arquivos de Neuro-psiquiatria* 74, 12 (2016), 990–998.

[15] A. Friesen and P. Domingos. 2015. Recursive decomposition for nonconvex optimization. In *International Joint Conference on Artificial Intelligence*. 253–259.

[16] B. Goldstein, A. Navar, and R. Carter. 2016. Moving beyond regression techniques in cardiovascular risk prediction: Applying machine learning to address analytic challenges. *Eur. Heart J.* 38, 23 (2016), 1805–1814.

[17] J. Hanley and B. McNeil. 1982. The meaning and use of the area under a receiver operating characteristic ROC curve. *Radiology* 143 (1982), 29–36.
[18] J. Hill, K. Dunn, M. Lewis, R. Mullis, C. Main, N. Foster, and E. Hay. 2008. A primary care back pain screening tool: Identifying patient subgroups for initial treatment. *Arthr. Rheum* 5, 59 (2008), 632–641.
[19] Yan Huang, Huiru Zheng, Chris Nugent, Paul McCullagh, Norman Black, Kevin E. Vowles, and Lance McCracken. 2010. Feature selection and classification in supporting report-based self-management for people with chronic pain. *IEEE Trans. Inf. Technol. Biomed.* 15, 1 (2010), 54–61. IEEE.
[20] D. Jha and G. Kwon. 2017. Diagnosis of Alzheimer's disease using a machine learning technique. *Alzh. Dement.* 13, 7 (2017), 1538.
[21] L. Kuncheva and C. Whitaker. 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach. Learn.* 51, 2 (2003), 181–207.
[22] Ludmila I. Kuncheva, Fabio Roli, Gian Luca Marcialis, and Catherine A. Shipp. 2001. Complexity of data subsets generated by the random subspace method: An experimental investigation. In *International Workshop on Multiple Classifier Systems*. Springer, 349–358.
[23] S. Lundberg and S. Lee. 2017. A unified approach to interpreting model predictions. In *Conference on Neural Information Processing Systems*. 4768–4777.
[24] S. Lundberg, B. Nair, M. Vavilala, M. Horibe, M. Eisses, T. Adams, D. Liston, D. Low, S. Newman, J. Kim, and S. Lee. 2018. Explainable machine learning predictions to help anesthesiologists prevent hypoxemia during surgery. *Nat. Biomed. Eng.* 2, 10 (2018), 749–760. DOI: https://doi.org/10.1101/206540
[25] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* 2, 1 (Jan. 2020), 56–67. DOI: https://doi.org/10.1038/s42256-019-0138-9
[26] O. Maimon and L. Rokach. 2002. Improving supervised learning by feature decomposition. In *International Symposium on the Foundations of Information and Knowledge Systems*. 178–196.
[27] Morteza Mashayekhi and Robin Gras. 2015. Rule extraction from random forest: The RF+HC methods. In *Advances in Artificial Intelligence*, Denilson Barbosa and Evangelos Milios (Eds.). Vol. 9091. Springer International Publishing, Cham, 223–237. DOI: https://doi.org/10.1007/978-3-319-18356-5_20
[28] G. McLachlan and D. Peel. 2000. *Finite Mixture Models*. Wiley Series in Probability and Statistics. John Wiley & Sons, New York.
[29] R. Melzack. 1975. The McGill pain questionnaire. major properties and scoring methods. *Pain* 1 (1975), 277–299.
[30] A. Navani and G. Li. 2016. Chronic pain challenge: A statistical machine-learning method for chronic pain assessment. *J. Rec. Adv. Pain* 2, 3 (2016), 82–86. DOI: https://doi.org/10.5005/jp-journals-10046-0048
[31] E. Nigri, N. Ziviani, F. Cappabianco, A. Antunes, and A. Veloso. 2020. Explainable deep CNNs for MRI-based diagnosis of Alzheimer's disease. In *International Joint Conference on Neural Networks*. IEEE, 1–8.
[32] A. Oliveira and S. Madeira. 2004. Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Trans. Computat. Biol. Bioinf.* 1, 1 (2004), 24–45.
[33] T. Pansombut, W. Hendrix, Z. Jacob Gao, B. Harrison, and N. Samatova. 2011. Biclustering-driven ensemble of Bayesian belief network classifiers for underdetermined problems. In *International Joint Conference on Artificial Intelligence*. 1439–1445.
[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. VanderPlas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12 (2011), 2825–2830.
[35] A. Pieterse, A. Stiggelbout, and V. Montori. 2019. Shared decision making and the importance of time. *JAMA - J. Amer. Med. Assoc.* 322, 1 (2019), 25–26. DOI: https://doi.org/10.1001/jama.2019.3785
[36] N. Pombo, P. Araújo, and J. Viana. 2014. Knowledge discovery in clinical decision support systems for pain management: A systematic review. *Arti. Intell. Med.* 60, 1 (2014), 1–11.
[37] M. Ribeiro, S. Singh, and C. Guestrin. 2016. "Why should i trust you?": Explaining the predictions of any classifier. In *International Conference on Knowledge Discovery and Data Mining*. 1135–1144.
[38] M. Ribeiro, S. Singh, and C. Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence*. 1527–1535.
[39] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence*, Vol. 18. 1527–1535.
[40] Greg Ridgeway, David Madigan, Thomas Richardson, and John O'Kane. 1998. Interpretable boosted naïve Bayes classification. In *KDD*. 101–104.
[41] Michael E. Robinson, Andrew M. O'Shea, Jason G. Craggs, Donald D. Price, Janelle E. Letzen, and Roland Staud. 2015. Comparison of machine classification algorithms for fibromyalgia: Neuroimages versus self-report. *J. Pain* 16, 5 (2015), 472–477. Elsevier.
[42] G. Shmueli. 2010. To explain or to predict? *Statist. Sci.* 25, 3 (2010), 289–310.
[43] Martin Tamajka, Wanda Benesova, and Matej Kompanek. 2019. Transforming convolutional neural network to an interpretable classifier. In *International Conference on Systems, Signals and Image Processing (IWSSIP'19)*. IEEE, 255–259. ZSCC: 0000002.

[44] A. Tanay, R. Sharan, M. Kupiec, and R. Shamir. 2004. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc. Natl. Acad. Sci.* 101, 9 (2004), 2981–2986.
[45] A. Tanay, R. Sharan, and R. Shamir. 2005. Biclustering algorithms: A survey. *Handb. Computat. Molec. Biol.* 9 (2005), 26–1.
[46] D. Valle, T. Pimentel, and A. Veloso. 2020. Assessing the reliability of visual explanations of deep models with adversarial perturbations. In *International Joint Conference on Neural Networks*. IEEE, 1–8.
[47] L. van der Maaten. 2009. Learning a parametric embedding by preserving local structure. In *International Conference on Artificial Intelligence and Statistics*. 384–391.
[48] T. Vos, A. Flaxman, M. Naghavi, R. Lozano, C. Michaud, M. Ezzati, K. Shibuya, J. Salomon, S. Abdalla, and V. Aboyans. 2012. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: A systematic analysis for the Global Burden of Disease Study 2010. *Lancet* 380, 9859 (2012), 2163–2196.
[49] J. Ward. 1963. Hierarchical grouping to optimize an objective function. *J. Amer. Statist. Assoc.* 58 (1963), 236–244.
[50] K. White, J. Lee, and A. de C Williams. 2016. Are patients' and doctors' accounts of the first specialist consultation for chronic back pain in agreement? *J. Pain Res.* (2016), 1109–1120.
[51] A. Williams and K. Craig. 2016. Updating the definition of pain. *PAIN* 157 (05 2016), 1. DOI: https://doi.org/10.1097/j.pain.0000000000000613

Received December 2019; revised November 2020; accepted May 2021