RESEARCH

Ada-WHIPS: Explaining AdaBoost Classification with Applications in the Health Sciences

Julian Hatwell*, Mohamed Medhat Gaber and R.M. Atif Azad

*Correspondence:
julian.hatwell@bcu.ac.uk
Birmingham City University,
Curzon Street, B5 5JU
Birmingham, UK
Full list of author information is available at the end of the article

Abstract

Background: Computer Aided Diagnostics (CAD) can support medical practitioners to make critical decisions about their patients' disease conditions. Practitioners require access to the chain of reasoning behind CAD to build trust in the CAD advice and to supplement their own expertise. Yet, CAD systems might be based on black box machine learning (ML) models and high dimensional data sources (electronic health records, MRI scans, cardiotocograms, etc). These foundations make interpretation and explanation of the CAD advice very challenging. This challenge is recognised throughout the machine learning research community. eXplainable Artificial Intelligence (XAI) is emerging as one of the most important research areas of recent years, because it addresses the interpretability and trust concerns of medical practitioners and other critical decision makers.

Method: In this work, we focus on AdaBoost, a black box model that has been widely adopted in the CAD literature. We address the challenge – to explain AdaBoost classification – with a novel algorithm that extracts simple, logical rules from AdaBoost models. Our algorithm, *Adaptive-Weighted High Importance Path Snippets* (Ada-WHIPS), makes use of AdaBoost's adaptive classifier weights; using a novel formulation, Ada-WHIPS uniquely redistributes the weights among individual decision nodes at the internals of the AdaBoost model. Then, a simple heuristic search of the weighted nodes finds a single rule that dominated the model's decision. We compare the explanations generated by our novel approach with the state of the art in an experimental study. We evaluate the derived explanations with simple statistical tests of well-known quality measures, precision and coverage, and a novel measure *stability* that is better suited to the XAI setting.

Results: In this paper, our experimental results demonstrate the benefits of using our novel algorithm for explaining AdaBoost classification. The simple rule-based explanations have better generalisation (mean coverage 15%-68%) while remaining competitive for specificity (mean precision 80%-99%). A very small trade-off in specificity is shown to guard against over-fitting.

Conclusions: This research demonstrates that interpretable, classification rule-based explanations can be generated for computer aided diagnostic tools based on AdaBoost, and that a tightly coupled, AdaBoost-specific approach can outperform model-agnostic methods.

Keywords: Explainable AI; Computer Aided Diagnostics; AdaBoost; Black Box Problem; Interpretability

Introduction

Medical diagnosis is a complex, knowledge intensive process. A medical expert must consider the symptoms of a patient, along with their medical and family history including complications and co-morbidities [1]. The expert may carry out physical examinations and order laboratory tests and combine the results with their prior knowledge. These activities are time intensive and suitably-experienced, available practitioners are a scarce resource in many healthcare systems. As healthcare needs grow and the sources of medical data increase in size and complexity, the diagnostic process must scale.

State of the art machine learning (ML) methods support computer aided diagnostics (CAD) which may address the scalability challenges and may improve patient outcomes [2, 3]. These methods demonstrate exceptional predictive and classification accuracy and can handle high dimensional data sets that often have very high rates of missing values. Examples of such challenging data sets include high throughput bioinformatics, microarray experiments, and complex electronic health records [4, 5], as well as unstructured, user-generated content (e.g. from social media feeds) that have been used to learn individuals' sub-health and mental health status outside of a clinical setting [6, 7]. Unfortunately, however, many state of the art ML models are so-called "black boxes", because they defy explanation. The complexity of black box models renders them opaque to human reasoning. Consequently, experts and medical practitioners are reluctant to accept black box models in practice, because it is difficult to relate the models to supporting theory; that is, to verify, approve or reason about the model decisions in light of expert clinical training and experience before those decisions are applied to patients [1, 8, 9, 10, 11, 12, 13, 14]. This barrier to adoption is evident, even when the black box models are demonstrably more accurate. There is also a legal right to explanation for high stakes decisions, which health status and medical conditions belong to [15, 2]. Some might argue that a black box model is no less transparent than a doctor [16]. Nevertheless, a doctor can be asked to justify their diagnosis and will do so from a position of domain understanding. In contrast, providing explanations for black box models is a very complex challenge. These models find patterns in data without domain understanding. Yet we wish to communicate explanations to a variety of levels of domain expertise: patient, practitioner, healthcare administrators and regulators. Additionally, we set higher standards of statistical robustness and completeness before granting our trust model derived decisions and explanations [17, 18].

Recent studies found that classification is the most widely implemented ML task in the medical sector and solutions using the AdaBoost algorithm [19] form a significant subset of the available research. Clinical applications include the diagnosis of Alzheimer's disease, diabetes, hypertension and various cancers [20, 21, 22, 23]. There are also non-clinical assessments of self-reported mental health and subhealth status (chronic fatigue and infirmity that can lead to future ill-health) using unstructured, user generated content from online health communities [6, 7]. AdaBoost has also been used as a preprocessing tool to automatically select the most important features from high dimensional data [24, 25]. Yet, AdaBoost is considered a typical black box as a consequence of its internal structure: an ensemble of typically 100s

to 1000s of shallow decision trees. The ensemble uses a weighted majority vote to classify data instances, which is difficult to analyse mathematically. The widespread adoption of AdaBoost in medical applications, coupled with its black box nature leads to the challenge; to make AdaBoost explainable.

We present Ada-WHIPS, a novel method for explaining multi-class AdaBoost classification. Ada-WHIPS is model-specific; it queries the internals of an AdaBoost model; a collection of adaptive weighted decision shallow decision trees. The method proceeds to extract the decision path from each stump that is specific to the data instance requiring an explanation (the explanandum). Only the paths that agree with the weighted majority vote are retained. These paths are disaggregated into individual decision nodes (which we call path snippets), and the weights are reassigned according to depth within the tree and frequency within the ensemble before filtering and sorting the most important nodes by weight. The adaptive-weighted, high importance path snippets are then greedily added to a classification rule. The final rule is tested for quality metrics and counter-factual conditions against the AdaBoost model training (or historical) data. An Ada-WHIPS explanation takes the form of a simple classification rule, presented alongside quality metrics coverage (generality) and precision (specificity). Additionally, counter-factual information is presented that shows how much precision decreases when any individual rule term is violated. The following examples of Ada-WHIPS explanations have been selected at random from our experimental study:

Table 1: Explanation of a classifier for foetal heart abnormalities.

Decision:	Explanation:	Contrast:	Confidence:
Suspect	ALTV $\leq 8.70 \wedge$	-21.1%	Covers 61.3% of historical
	ASTV ≤ 61.67	-24.5%	Matches 93.3% of covered
Prior 78.0%			Vote margin 10.4%

Table 2: Explanation of a non-clinical mental health assessment classifier.

Decision:	Explanation:	Contrast:	Confidence:
Has sought	work interfere $\leq 1.5 \wedge$	-50.6%	Covers 26.9% of historical
treatment	family history > 0.9	-28.3%	Matches 99.6% of covered
Prior 54.9%			Vote margin 0.7%

Table 3: Explanation of automated 30-day hospital readmission risk assessment.

Decision:	Explanation:	Contrast:	Confidence:
Risk: Low	number diagnoses $\leq 6.8 \wedge$	-42.4%	Covers 3.9% of historical
	number emergencies $\leq 1.8 \wedge$	-63.3%	Matches 98.2% of covered
	inpatient visits $\leq 1.5 \wedge$	-64.8%	Vote margin 0.2%
	outpatient visits $\leq 2.5 \wedge$	-40.9%	
Prior 65.0%	time in hospital ≤ 6.5	-36.2%	

In Table 1, ASTV is the percentage of time with abnormal short term variability, and ALTV is the percentage of time with abnormal long term variability. These are statistical features computed by processing fetal cardiotocograms. In Table 2, an online health community responded to a twenty-four question survey on their mental health. The classification model identifies those individuals who have actually sought

Table 4: Explanation of a classifier for thyroid condition.

Decision:	Explanation:	Contrast:	Confidence:
Abnormal	TSH > 6.07	-74.5%	Covers 10.0% of historical Matches 93.4% of covered
Prior 26.0%			Vote margin 6.0%

treatment. The explanation shown refers to the questions "If you have a mental health condition, do you feel that it interferes with your work?" (Answers: 0 = Often, 1 = Sometimes, 2 = Not Sure, 3 = Rarely, 4 = Never) and "Do you have a family history of mental illness?" (Answers: 0 = No, 1 = Not Sure, 2 = Yes). Table 3 shows attributes from an electronic health record that were critical in determining the risk of readmission for one particular patient. Table 4 shows the results of a classifier for thyroid condition. TSH is the level of thyroid stimulating hormone. Full details of the data sets used can be found in Table 5.

The rest of this paper is organised as follows: Related work is discussed in Section 2. The Multi-Class AdaBoost algorithm is described in Section 3. Our contribution is presented in Section 4. Section 5 describes our experimental methodology and Section 6 reports and discusses the results. We summarise and suggest future work in Section 7.

Related Work

Early work to simplify AdaBoost models involved pruning (reducing the number of trees in the model) through a variety of approaches including statistical methods, clustering, or optimisation by genetic algorithms [26, 27, 28]. In common, these approaches seek to trade off the smallest decrease in accuracy while keeping only a subset of decision trees. More recently, various methods take a decompositional approach [29] to interpretability, directly querying the set of all decision nodes within each decision stump. Examples in the literature include: Bayesian Rule Lists [30] DefragTrees [31], Forex++ [32], RF+HC [33], inTrees [34], RuleFit [35], BRUTE [36]. All these methods generate a cascading rule list (CRL) as a simpler, surrogate model. The prevalence of CRL as interpretable models indicates the importance of logical rules for explainability. Logical rules are intuitive to understand, being the standard language of reasoning [17, 37].

An alternative approach is explanation methods that generate post-hoc, per instance explanations [38]. Several such methods have been proposed as model-agnostic frameworks (also known as didactic methods [29]). The model-agnostic assumption is that any model's behaviour can be explained given access only to the model inputs and outputs but not the training data nor the model internals. Model-agnostic methods probe the model's behaviour with a synthetic input sample. Each explanation is inferred from the effect of different input attributes on the outputs. LIME [18] generates a sparse linear model, SHAP [39] uses a game theoretic approach for a similar result: a set of non-zero coefficients for the input attributes. The coefficients are additive and their magnitude is proportional to the importance in the classification of the attributes they represent. This property leads to the categorisation of these methods as Additive Feature Attribution Methods (AFAM) [39]. Anchors [37] and LORE [40] also use synthetic samples but generate a single

classification rule (CR) as an explanation (as opposed to the coefficients of AFAM). Anchors uses the same synthetic sampling technique used by LIME. In fact, the method was developed by the same research team to overcome the short-comings of AFAM methods. LORE uses a genetic algorithm to generate the sample. Model-agnostic techniques have been shown to be effective in image and text classification but, on tabular data set, require additional checks because the attributes included in the explanation might be unstable over repeated trials [41, 42]. Furthermore, for tabular data, a realistic synthetic distribution must be estimated from the training data set or a large i.i.d. sample, which violates the model-agnostic assumption of accessing only the inputs and outputs of the black box model. LIME, Anchors and SHAP sample from the marginal training distribution and are thought to put too much weight on unlikely examples. Moreover, LIME and Anchors require all features of tabular data to be categorical. Continuous features must be discretised in advance of training the classification model and using an arbitrary procedure such as quantile binning [37]. This significant compromise puts constraints on the model of choice and potentially loses important information from the continuous features. A better performing model might be trained without this constraint.

In a recent, comprehensive literature review [43] the following methods were categorised as model-agnostic when, in fact, they are model-specific: Saliency Maps, Activation Maximisation, Layer-wise Relevance Propagation. These models require access to the internals of neural networks. Such misclassification aligns with a prevailing view that model-agnostic methods are a very active research area while model-specific methods may be in decline but we suggest that the model-agnostic assumption should be taken with caution. We argue that model-agnostic methods are only required for a subset of ML problems, such as model auditing by an external third party. We have already mentioned that, for tabular data, access to the training set must be assumed. If there is access to both the training data and the model internals, decompositional methods can deliver explanations that are more representative of the model's internals [43]. Model-specific methods remain as an active and research area. These methods address specific scenarios, making assumptions about the black box model to be explained. Treeinterpreter [44] is possibly the earliest of these, applicable to regression problems with Random Forest models. TreeSHAP [45], based on the SHAP method, assumes an underlying XGBoost model and queries the internal decision nodes. This provides faster and more consistent results than the original SHAP algorithm for XGBoost models.

Our novel algorithm is a model-specific, CR-based method. Our method targets AdaBoost models and is an improvement on previous research: CRL-based methods decompose an entire AdaBoost model to generate a proxy model. The proxy models are a trade-off; increasing interpretability but also increasing classification error and giving no guarantees of fidelity with the original model. Any less than perfect fidelity means that, for some instances, proxy and model do not agree. Explanations that refer to a different class than the model's predicted class are of no use in a critical setting, such as CAD. It is trivial to measure the generality of CR-based explanations because the relevance to other instances is determined unambiguously by the coverage. This is an advantage over AFAM explanations which are are difficult to apply to other instances, even when there are only small differences [37]. Model

agnostic methods can also be slow to compute. For example, computing Shapley Values entails solving a large combinatorial problem which limits the scalability [45], while LORE's synthetic samples are generated by a genetic algorithm that is not parallelisable in the currently available version[1]. Finally, our method exploits the information-theoretic discretisation of the continuous features that occurs during the induction of individual decision trees in the AdaBoost ensemble. This is an advantage over methods that require discretisation as a preprocessing step.

Multi-Class AdaBoost

In this section, we describe multi-class AdaBoost, on which our method is based. Boosting is the process of sequentially combining weak base classifiers to build up a strong classifier. It is one of the most significant developments in Machine Learning [46, 47]. AdaBoost [48] was the first, widely used implementation of boosting and is still favoured for its accuracy, ease of deployment and fast training time [49, 50, 51]. It uses shallow decision trees as the weak classifiers. On each iteration, the training sample is re-weighted such that the next decision tree focuses on examples that were previously misclassified, while previously generated classifiers remain unchanged (the details of this iterative re-weighting are not central to this research and can be found in [48, 52]). AdaBoost also adaptively updates its base classifier weights based on their individual performance, which we discuss now in further detail. Two algorithms, SAMME and SAMME.R [52] have emerged as the standard [53] for extending the original AdaBoost algorithm to multi-class problems. Let $f : \mathcal{X} \longmapsto \mathcal{Y}$ be an unknown classification function that we would like to approximate, where $\mathcal{X}$ is an $\mathbb{R}^d$ input space and $\mathcal{Y} = \{C_1, \ldots, C_K\}$ is the set of possible classes. Let $\mathbf{X}$ be an input data set and our multi-class AdaBoost model be $g(\mathbf{X}) \approx f(\mathbf{X})$. To classify an instance $\mathbf{x}$ with SAMME, each base classifier emits a one dimensional vector indicating the position of the output class. The output of the whole model is the weighted majority vote of all the base classifiers.

$$\begin{aligned} g(\mathbf{x}) = C_k,\ k = arg\,\underset{k \in K}{max} \sum_{m=1}^{M} \alpha^{(m)} \cdot T^{(m)}(\mathbf{x}), \\ T^{(m)}(\mathbf{x}) = [c_1, \ldots, c_K],\ \sum T^{(m)}(\mathbf{x}) = 1 \end{aligned} \tag{1}$$

where $c_k = 1,\ c_j = 0,\ j \neq k$ indicates that C_k is the output class and $g = \{\{T^{(1)}, \ldots, T^{(M)}\}, \{\alpha^{(1)}, \ldots, \alpha^{(M)}\}\}$ a set of base decision tree classifiers and the set of classifier weights. These weights are calculated during the training phase as:

$$\alpha^{(m)} = \log \frac{1 - err^{(m)}}{err^{(m)}} + \log(K - 1),\ 0 < err^{(m)} \leq 1 - \frac{1}{K} \tag{2}$$

where $err^{(m)}$ is the error rate at iteration m.

To classify an instance $\mathbf{x}$ with SAMME.R, each base classifier returns a vector of the conditional probabilities that the class of $\mathbf{x}$ is C_k. This is the distribution

[1] https://tinyurl.com/qlyxzlv

of training instance weights in the terminal node of the decision path taken by $\mathbf{x}$ through each tree:

$$T^{(m)}(\mathbf{x}) = [\mathbb{P}_{T^{(m)}}(C_1|x),\ \ldots,\ \mathbb{P}_{T^{(m)}}(C_K|x)],\ \sum T^{(m)}(\mathbf{x}) = 1,\ y \in \mathcal{Y} \tag{3}$$

and confidence weights are calculated at run time as:

$$\alpha_k^{(m)}|x = (K-1)\big(\log \mathbb{P}_{T^{(m)}}(C_k|x) - \frac{1}{K}\sum_{j=1}^{K} \log \mathbb{P}_{T^{(m)}}(C_j|x)\big). \tag{4}$$

The output of the whole model is the majority vote based on the additive contribution of these confidence weights per class:

$$g(\mathbf{x}) = C_k,\ k = \underset{k}{argmax} \sum_{m=1}^{M} \alpha_k^{(m)}|x. \tag{5}$$

where $g = \{T^{(1)},\ \ldots,\ T^{(M)}\}$ (weights evaluated at run time).

Ada-WHIPS

We now present our algorithm for generating a CR based explanation for the classification of an explanandum instance $\mathbf{x}$ by a previously trained AdaBoost model g. The algorithm begins by initialising an empty rule and setting the consequent as the classification outcome $g(\mathbf{x})$. The target is to identify a small set of antecedent terms, or logical conditions that must be true of $\mathbf{x}$ and that exert the most influence on the classification. The source of the logical conditions is the ensemble of decision trees that make up g. The influence is determined by the classifier weights within the internals of g, which themselves are derived from the error rates (weights increase as errors decrease).

Extracting Decision Paths

An AdaBoost model typically comprises 100's-1000's of shallow decision trees, potentially resulting in a very large search space. We can reduce this space logarithmically by considering only decision paths of $\mathbf{x}$ in each decision tree and ignoring all other branches. The paths retain all the information about how $g(\mathbf{x})$ was determined. An example decision path is shown in Fig 1. Here, $\mathbf{x} = \{\ldots,\ x_i = 0.1,\ x_j = 10,\ \ldots\}$, where x_i is the attribute value of the i^{th} feature. The decision path starts from the root node Q_1, following the binary split conditions up to a leaf node. The decision path contains node detail triples of the following form (j, ν, τ), where j is a feature index and $\nu \in \mathbb{R}$ is the threshold for the inequality $x_j < \nu$ and $\tau \in \{0, 1\}$ is the binary truth of evaluating the inequality. Note that for this instance, all other nodes are irrelevant. For example, even though Q_7 applies to x_i, it cannot be reached by $\mathbf{x}$.

The search space can be further reduced by only considering trees that participated in the majority vote. The rationale for this is based on the application of

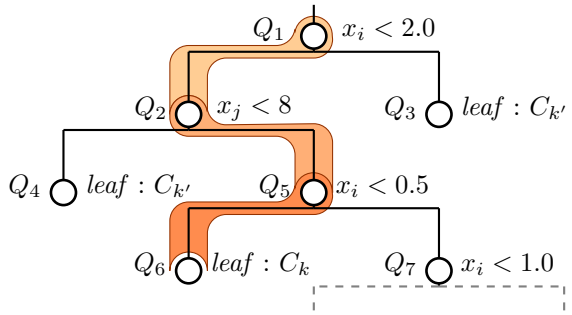


Figure 1: The decision path of an instance.

maximum margin theory to boosting [54]. If $\mathbf{x}$ is an unseen instance, the margin in SAMME is:

$$\begin{aligned} margin &= \frac{a^+ - a^-}{\sum_{m=1}^{T} \alpha^{(m)}},\ a^+ = \sum_{n=1}^{|\mathcal{T}^+|} \alpha^{(n)},\ a^- = \sum_{k=1}^{K} \frac{1}{K-1} \sum_{u=1}^{|\mathcal{T}^-|} \alpha^{(u)}, \\ \mathcal{T}^+ &= \{T : g(\mathbf{x}) = C_k,\ k = \underset{k \in K}{arg\,max} T(\mathbf{x})\}, \\ \mathcal{T}^- &= \{T : g(\mathbf{x}) = C_k,\ k \neq \underset{j \in K}{arg\,max} T(\mathbf{x})\},\ T^{(m)}, \alpha^{(m)} \in g. \end{aligned} \tag{6}$$

The quantity a^+, represents the sum of weights from the classifiers that voted for the majority class and $a^+ > a^-$ is always true for the majority class. The set $\mathcal{T}^+$ are the base classifiers voted in the majority and thus contributed their weight to a^+, and $\mathcal{T}^-$ are the remaining classifiers. $\mathcal{T}^+$ completely determines the ensemble's output for a given instance because an ensemble classifier formed from the union of $\mathcal{T}^+$ and any subset of $\mathcal{T}^-$ would return the same classification with a larger m because $a_*^- < a^-,\ \mathcal{T}_*^- \subset \mathcal{T}^-$. We found no margin formalisation for SAMME.R in the literature but we can define $\mathcal{T}^+ \coloneqq \{(T^{(m)}, \alpha_k^{(m)}) : \alpha_k^{(m)} \geq \alpha_j^{(m)},\ k, j \in \{1, \ldots, K\}\}$ and, as a convenience, we can substitute the α terms in equation (6) for the following Kullback-Leibler (KL) Divergence. The KL-Divergence (also known as "relative entropy") measures the information lost if a distribution P' is used, instead of another distribution P to encode a random variable and is defined as:

$$D_{KL}(P \parallel P') = -\sum_{x \in \mathcal{X}} P(\mathbf{x}) \log \left(\frac{P(\mathbf{x})}{P'(\mathbf{x})} \right) \tag{7}$$

and we set P, P' as the posterior class distribution of each $T^{(m)}(\mathbf{x})$ given in equation (3), and prior class distribution in the training data, respectively. The KL-Divergence will be larger for trees that classify with greater accuracy, relative to the prior distribution. The D_KL emulates the classifier weights yielded by equation (2), which allows the rest of the algorithm to proceed in an identical manner for SAMME and SAMME.R.

Redistributing the Weights

To avoid a combinatorial search of all the available decision nodes, we sort them, prior to rule merging, according to their ability to separate the classes. To do this, we disaggregate the whole ensemble into individual decision nodes and redistribute the classifier weights onto the nodes. This procedure is illustrated in Algorithm 1. The contribution of each node is conditional on the previous nodes in the path, so this sorting must take into account the nodes' depth in the originating tree. To do this, we apply equation (7) to determine the relative entropies at each point in a path. For each root node, we set P, P' as the class distribution when applying that decision to the training data, and the prior class distribution respectively. For subsequent nodes, P is the class distribution after applying all previous decision nodes and P' is the distribution up to but not including the current node. The relative entropy scores for nodes in a single path are normalised such that their total is equal to that of the classifier weight $\alpha^{(m)}$. The scores are grouped and summed for nodes that appear in multiple paths. We filter the nodes, keeping only those with the largest weights (e.g. top 20%). Finally, all nodes from all paths are sorted by this score in descending order.

Algorithm 1 Get Term Weights

procedure GET TERM WEIGHTS($\mathbf{x}, g, (\mathbf{X})$) ▷ instance, model and training set
  Terms Weights $\leftarrow \{< term >, < weight >\}$ ▷ initialise empty map of terms and weights
  $Y^{(\text{idx}_0)} \leftarrow g(\mathbf{X})$ ▷ training set classifications
  for $T^{(m)} \in \mathcal{T}^+,\ \mathcal{T}^+ \subset g$ **do**
    $\text{Path}^{(m)} \leftarrow$ Get Decision Path($\mathbf{x}, T^{(m)}$) ▷ See Fig 1
    $N \leftarrow$ length of $\text{Path}^{(m)}$
    for $n = 1, N, n++$ **do**
      $\text{idx}_n \leftarrow$ set of indices from $\mathbf{X}$ covered by $Q_n^{(m)} \wedge Q_{n-1}^{(m)} \wedge \cdots \wedge Q_1^{(m)}$
      $d_n \leftarrow D_{KL}(Y^{(\text{idx}_n)} \parallel Y^{(\text{idx}_{n-1})})$
    Normalise all d_n
    for $n = 1, N, n++$ **do**
      if $Q_n^{(m)} \notin$ Terms Weights $< term >$ **then**
        Append $Q_n^{(m)}$, $< d_n \cdot \alpha^{(m)} >$ to Terms Weights
      else
        Terms Weights $< weight > += d_n \cdot \alpha^{(m)}$ where Terms Weights $< term >= Q_n^{(m)}$
  Select top N (or top n%) Terms Weights
  Sort Terms Weights
  Return(Terms Weights)

Merging Decision Nodes into a Classification Rule

It is trivial to convert the node detail triples (j, ν, τ) into antecedent terms of a CR [55]. We use nodes and terms interchangeably from here on. The objective is to find a minimal set of terms that maximises both precision and coverage while mitigating the problem of over-fitting. Over-fitting can occur if we maximise precision as an objective function. We risk converging on a "tautological" rules that provide no generalisation. A tautological rule contains enough terms to uniquely identify a single instance. This is because precision is trivially maximised by single instances. In a noisy data set, there could be many such local maxima. Therefore, we propose *stability* as a novel objective function, defined as:

$$\zeta(\mathbf{x}, g, \mathbf{Z}) = \frac{|\{\mathbf{z} : g(\mathbf{z}) = g(\mathbf{x}),\ \mathbf{z} \in \mathbf{Z}\}|}{|\mathbf{Z}| + K} \tag{8}$$

where $\mathbf{Z}$ is the set of instances covered by the current rule and K the number of classes. The maximum achievable ζ is $\frac{1}{K}$ for a single instance but approaches precision asymptotically as $|\mathbf{Z}| \to \infty$. Stability, therefore acts as a brake on adding too many terms and over-fitting. We proceed with a breadth first search, iteratively adding terms to an initially empty rule. We always add the first term in the sorted list. Then, we work down the list, greedily adding further terms if they increase stability and discard them if they do not. The algorithm stops when a threshold stability (e.g. 0.95) is reached or the list is exhausted. These steps are illustrated in Algorithm 2

Algorithm 2 Merge Rule

procedure MERGE RULE($\mathbf{x}, g, \mathbf{X}, \mathbf{Y}, \rho$) ▷ instance, model, training set and target ζ
  Terms Weights ← Get Term Weights($\mathbf{x}, g, \mathbf{X}$)
  Consequent ← $g(\mathbf{x})$
  Initialise empty Antecedent
  $s \leftarrow \mathbb{P}(\mathbf{Y} = g(\mathbf{x}))$ ▷ prior class distribution
  while Terms Weights is not empty $\wedge$ $s <= \rho$ **do**
    Term ←$< term >$ from Terms Weights where $< weight > = max(< weight >)$
    Delete $< term >, < weight >$ from Terms Weights where $< term >$ = Term
    $\mathbf{Z}$ ← instances covered by Antecedent $\wedge$ Term
    if $\zeta(\mathbf{x}, g, \mathbf{Z}) > s$ **then**
      Append Term to Antecedent
      $s \leftarrow \zeta(\mathbf{x}, g, \mathbf{Z})$
  Return(Antecedent $\implies$ Consequent)

Generating Counterfactuals

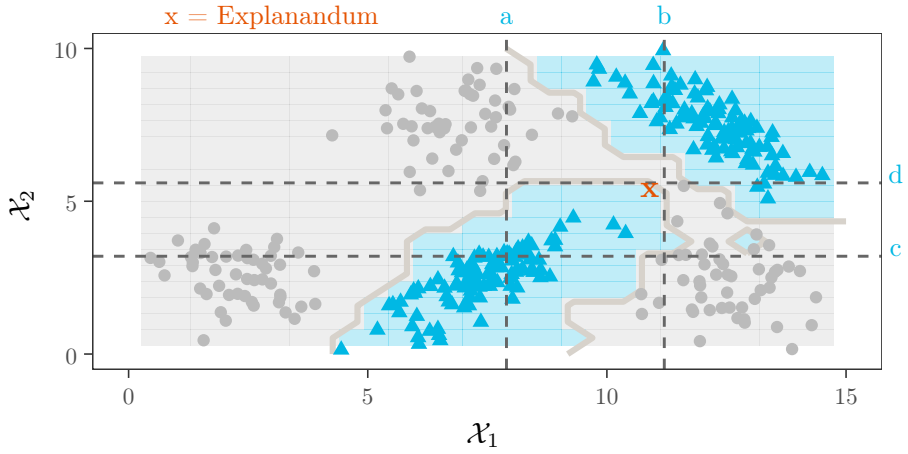
Counter-factuals answer the question "what would have happened?" They illustrate minimal changes in the inputs that would give different results. Some authors define counter-factual (sometimes called contrastive) explanations as a minimal change set on the inputs that would return a different result [40, 56, 2, 12]. However, illustrative examples are limited to a change of classification and do not include any room for uncertainty. We suggest a fuzzy definition for the change of output; namely, a drop in precision beyond a user-defined tolerance, rather than a discrete change of class. This definition accommodates more room for expert judgement in the interpretation. For example, a change from high to low confidence in a CAD or risk score. It is easy to derive such counter-factuals from CR and it is not necessary, nor even practical to provide every possible input scenario. Since the definition of counter-factuals is a minimal change set, it suffices to show the effect of a point change at each of the rule terms, one at a time. Any point changes that do not decrease the precision beyond the user-defined tolerance represent a non-counter-factual change and can be removed from the rule. This procedure provides an intuitive pruning mechanism for removing redundant terms that might have been added during the greedy rule merge algorithm. We illustrate this concept visually in Fig 2. Here a model with a complex decision boundary is trained on a synthetic data set (a Gaussian mixture model) which has two classes, shown as triangles and circles. The model classifies an explanandum instance $\mathbf{x}$ as a triangle. The explanation is found - the following CR: $\{\mathbf{z} : a \leq z_1 \leq b,\ c \leq z_2 \leq d,\ \mathbf{z} \in \mathcal{X}\} \implies$ triangle. The counter-factual

spaces are those spaces immediately adjacent to the four rule boundaries, derived by reversing one inequality at a time:

$$\begin{aligned}&\{\{\mathbf{z} : z_1 \leq a, c \leq z_2 \leq d\},\ \{\mathbf{z} : b \leq z_1, c \leq z_2 \leq d\},\\ &\{\mathbf{z} : a \leq z_1 \leq b, z_2 \leq c\},\ \{\mathbf{z} : a \leq z_1 \leq b, d \leq z_2\},\ \mathbf{z} \in \mathcal{X}\}\end{aligned} \tag{9}$$

Even though the triangle class is still predicted for parts of these spaces, a classification rule formed from any one of these counter-factual spaces as the antecedent would have a severely reduced expectation for precision with the same consequent. Thus, the original rule provides a crisp boundary, while the counter-factual rules are fuzzy.

Figure 2: Counter-factual spaces - conceptual diagram.



Experiments

To assess the performance of Ada-WHIPS, its effectiveness and efficiency are comparted to those of state-of-the-art methods. Three metrics are used to measure effectiveness, namely, coverage, precision and our new measure of stability. Efficiency is measured using the average time to generate an explanation.

We compared Ada-WHIPS in an experimental study with the state of the art in CR-based per instance explanation methods: Anchors [37] and LORE [40]. Both methods are model-agnostic. Readers who are familiar with XAI research may question the ommission of LIME [18] and SHAP [39], which are the most discussed per instance explanation methods. LIME and SHAP fall into a different class of methods, described as *additive feature attribution methods* (AFAM). AFAM comprise a local linear model (LM) with coefficients relating the importance of various attributes to the model's classification of the explanandum. There is no obvious way to apply the local LM for one instance to any other instances in order to calculate the quality measures such as precision and coverage, and comparison with CR-based methods is of limited value [37]. Fortunately, Anchors has been developed by the same research group that contributed LIME and uses the same synthetic

sampling technique. Anchors can be viewed as a rule-based extension of LIME and its inclusion into this experimental study provides a useful comparison to AFAM research.

Limitations of the Study

Unfortunately, we discovered that LORE was not scalable after finalising our experimental design. The time cost of generating a synthetic distribution by means of a genetic algorithm rendered the method unusable on some of the data sets. The time per instance was on average twenty-five to thirty minutes for the hospital readmission data set and more than two hours per instance on the understanding society data set. The method generated system errors on the mental health survey '14 data set and was not runnable at all. We thoroughly examined the source code to look for opportunities to parallelise the operation, but the presence of a dynamically genarated, non-serialisable distance function rendered this impossible. We have included the results where the method ran to completion.

As mentioned in the section on related work, Anchors requires all features of the data to be categorical. For our experiments, we generated a copy of each data set, and discretised them using Anchors' provided quartile binning function. A second AdaBoost model was generated from this discretised data set for Anchors to explain.

Data Sets

We used the nine data sets detailed in Table 5. These were sourced from the UCI Machine Learning repository [57] and represent specific disease diagnoses from clinical test results, except; the mental health surveys (Kaggle) which represent case studies in detecting mental health conditions from non-clinical online health community data; the hospital readmission data (Kaggle) which represents a large electronic health record, and understanding society [58] which is from the General Population Sample of the UK Household Longitudinal Study. We use the file from waves 2 and 3 where participants had a health visit carried out by a qualified nurse. At least one study [59] has shown that the biomarkers measured in the survey may be asociated with the results from self-completion instruments measuring mental health. We run a classification task for the SF-12 Mental Component Summary (PCS) which has been discretised into nominal values "poor," "neutral" and "good."

Table 5: Data sets used in the experiments.

data set	Target	Classes	Class balance	Features	Of which nominal	N
breast cancer	mb	2	0.63 : 0.37	31	1	569
cardiotocography	NSP	3	0.78 : 0.14 : 0.08	22	1	2126
diabetic retinopathy	dr	2	0.53 : 0.47	20	1	1151
cleveland heart	HDisease	2	0.54 : 0.46	14	8	303
mental health survey '16	mh2	2	0.50 : 0.50	46	44	1433
mental health survey '14	treatment	2	0.51 : 0.49	24	3	1259
hospital readmission	readmitted	2	0.54 : 0.46	65	1	25000
thyroid	diagnosis	2	0.74 : 0.26	30	3	9172
understanding society[2]	mh	3	0.22 : 0.62 : 0.16	330	246	11745

Hardware Setup

The experiments were conducted using Python 3.6.x running on a standalone Lenovo ThinkCentre with Intel i7-7600 CPU @ 3.4GHz and 32GB RAM using the Windows 10 operating system.

Method

Each data set was split into training and test sets (70%, 30%) by random sampling without stratification or other class imbalance correction. We trained AdaBoost models using ten-fold cross-validation of the training set on number of trees $ntrees \in \{200, 400, \ldots, 1600\}$. The tree's maximum depth *maxdepth* was always 4. We used the *ntree* setting that delivered the highest classification accuracy to train a final model on the whole training set. Anchors requires all features of the data to be categorical [37]. So, we generated a copy of each data set, and discretised them using Anchors' provided quartile binning function. Training and test splits used identical indices as the undiscretised versions. Using each discretised training set, a second AdaBoost model was generated to use with Anchors. Each test set was then used as the pool of unseen instances to be classified by the AdaBoost model and explained by Ada-WHIPS, Anchors and LORE. Thus, there are three comparable explanations for each test instance. Generating explanations is done instance by instance, not batch wise as in classification. So, for time constraints, the number of instances (test units) was limited to either the whole test set or the first one thousand test instances, whichever was the smaller. For each explanation, all the remaining instances from the entire test set were used to assess the standard quality measures, precision and coverage, along with the novel quality measure, stability (8), which is more sensitive to over-fitting. This leave-one-out procedure ensures that test scores are not biased by leakage of information from the explanation-generating instance. The entire procedure is repeated for SAMME and SAMME.R AdaBoost models.

We report differences between precision, stability and coverage among the algorithms using non-parametric hypothesis tests. The reason for using these tests is that these measures are proportions; from the interval $[0, 1]$ and very right-skewed by design as each method tries to generate very high precision explanations. We use the paired samples Wilcoxon signed rank test where we have results for just Ada-WHIPS and Anchors. The null hypothesis of this test is that the medians of the two samples are equal and the alternative is that the medians are unequal. We use the Friedman test where we have results for all three methods. The Friedman test is a non-parametric equivalent to ANOVA and an extension of the rank sum test for multiple comparisons. The original Friedman test produces an approximately χ^2 distributed statistic, but this is known to be very conservative. Therefore, we use the modified F-test given in [60], because we have very large values for N, i.e. the count of instances in the test set. The null hypothesis of this test is that there is no significant difference between the mean ranks R of all the groups and the alternative is that at least two mean ranks are different. The null hypothesis is rejected when F_F exceeds the critical value for an F distributed random variable with the first

degrees of freedom $df_1 = k-1$ and the second $df_2 = (k-1)(N-1)$, where k is the number of algorithms:

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}, \quad \chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4}\right] \tag{10}$$

Where this test returned a significant result, we ran the recommended pairwise, post-hoc comparison test with the Bonferroni correction (for three pairwise comparisons) proposed in [60]:

$$z = \text{diff}_{ij} \Big/ \sqrt{\frac{k(k+1)}{6N}}, \ \text{diff}_{ij} = R_i - R_j \tag{11}$$

where R_i and R_j are ranks of two algorithms and z is distributed as a standard normal under the null hypothesis that the pair of ranks are not significantly different. It is sufficient for this study to demonstrate whether the top scoring algorithm was significantly greater than the second place algorithm on our quality measures of interest.

AdaBoost Models

We present the performance scores of the trained models in Tables 6. It is important to note that the model training is part of the experimental setup. These scores are secondary to the current research and are not part of the research contribution. We only provide this level of detail to demonstrate that the trained models reasonably approximate the underlying data sets. An explanation method could just as well be applied to a poorly performing model. We show generalisation accuracy scores for the three resulting models and Cohen's κ for the two models on each data set. Cohen's κ is a useful measure in multi-class problems and class imbalanced data because this statistic corrects for chance agreement, which can be high in such cases. Values close to zero indicate a high degree of chance agreement. Cohen's κ is calculated as:

$$\kappa = \frac{N\sum_{i=1}^{K} N_{ii} - \sum_{i=1}^{K} N_{i+}N_{+i}}{N^2 - \sum_{i=1}^{K} N_{i+}N_{+i}}, \quad \begin{bmatrix} \mathrm{N}_{11} & \mathrm{N}_{12} & \dots & \mathrm{N}_{1K} \\ \mathrm{N}_{21} & \mathrm{N}_{22} & \dots & \mathrm{N}_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{N}_{K1} & \mathrm{N}_{K2} & \dots & \mathrm{N}_{KK} \end{bmatrix} \tag{12}$$

where K is the number of classes, N is the total number of instances, N_{ij} is the number of instances in cell ij of the confusion matrix of true vs. predicted class counts, and N_{i+}, N_{+j} are the i^{th} row and j^{th} column marginal totals respectively.

Discussion

Our approach for the experimental study is based on the simulated user study implemented in [37]. That study proposes that coverage represents the fraction of

Table 6: AdaBoost SAMME Model Performance Scores.

		Undiscretised: used by Ada-WHIPS & LORE		Discretised: used by Anchors	
data	ntree	Accuracy	κ	Accuracy	κ
	SAMME				
breast cancer	200	0.98	0.96	0.96	0.92
cardiotocography	800	0.94	0.84	0.89	0.70
diabetic retinopathy	1000	0.68	0.36	0.66	0.33
cleveland heart	200	0.77	0.52	0.80	0.59
mental health survey '16	200	0.88	0.76	0.88	0.75
mental health survey '14	200	0.83	0.65	0.81	0.62
hospital readmission	800	0.62	0.22	0.60	0.18
thyroid	1200	0.97	0.92	0.80	0.45
understanding society	600	0.64	0.13	0.61	0.14
	SAMME.R				
breast cancer	1000	0.98	0.96	0.95	0.90
cardiotocography	1600	0.94	0.82	0.88	0.67
diabetic retinopathy	200	0.69	0.38	0.65	0.30
cleveland heart	400	0.76	0.50	0.82	0.63
mental health survey '16	800	0.87	0.73	0.86	0.72
mental health survey '14	200	0.80	0.60	0.81	0.63
hospital readmission	200	0.62	0.22	0.63	0.23
thyroid	1600	0.97	0.92	0.76	0.37
understanding society	200	0.62	0.13	0.62	0.15

previously unseen instances a user could attempt to classify after seeing an explanation, showing how generally the rule applies to the whole population. Similarly, precision represents the fraction of those classifications that would be correct, indicating the specificity of the rule. Real users who were shown high coverage and precision rule-based explanations demonstrated significantly improved task completion scores over those who were shown AFAM explanations.

Coverage Analysis

We present a visual analysis of the raw data (see Appendix for results tables) and tabulate the results of our statistical tests. For all our three-way comparisons using the Friedman test, all p-values were vanishingly small ≈ 0. In the following reportage, we proceed directly to the post-hoc tests described. The critical value for a two-tailed test with the bonferroni correction is $\frac{0.025}{3} = 0.000833$. The two-way comparisons using the paired samples Wilcoxon signed rank tests are shown in separate tables to avoid drawing invalid comparisons. A significant result is indicated by ** and the winning algorithm formatted in boldface.

A cursory inspection of the mean coverage charts shown in Figs 3-4 indicates that Anchors has the lowest mean coverage over all the data sets but the comparison between Ada-WHIPS and LORE is less clear cut. The results of the hypothesis tests are given in Tables 7-8. The Wilcoxon tests showed that Ada-WHIPS always has significantly higher coverage than Anchors. Ada-WHIPS was the top algorithm in all but three of the post-hoc tests for three-way comparisons and in the top two alongside LORE for the remaining tests. There was no significant difference between the top two in these cases.

Figure 3: Mean Coverage for SAMME model explanations.

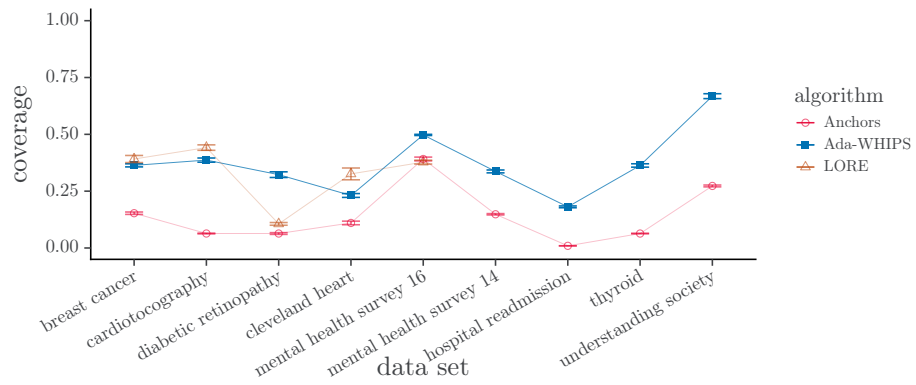
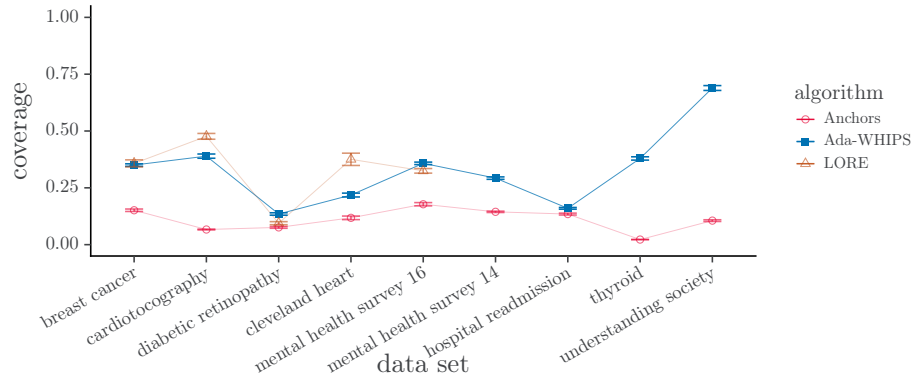


Figure 4: Mean Coverage for SAMME.R model explanations.



Precision Analysis

The mean precision chart, (Fig 5), shows that LORE has the lowest precision in all but one of the data sets where LORE results are available. It is harder to see if there is a definitive lead between Ada-WHIPS and Anchors.

However, the complete picture – and the cost to Anchors of implementing a precision guarantee – can be seen in the distribution chart in Fig 7. Here we see that a certain proportion of explanations have a precision of 0.0. The result shows that Anchors (and LORE to a lesser extent) is over-fitting. Some explanations are so specific that they only explain the explanandum and do not generalise to other instances in the test set. We present the proportion of 0.0 point explanations that were returned by each algorithm in Table 9.

The proportions vary from around $0.5\% - 28\%$. There are important consequences for methods that suffer this level of over-fitting. The most important consequence is that 0.0 precision rules are so specific that they uniquely identify the explanandum but cover no other instance from the held out test set. A unique identifier does not provide any useful new information that explains the model's classification. For the person requiring the explanation, this outcome represents a failure of the system. The lowest failure rates in this range may be tolerable, depending on the criticality

Table 7: Coverage: Top two by mean rank (mrnk) for three-way comparisons.

data	1st	mrnk	2nd	mrnk	N	z	p.value
SAMME							
breast	LORE	1.54	Ada-WHIPS	1.61	170	0.41	0.3412
cardiotocography	LORE	1.52	Ada-WHIPS	1.62	637	1.06	0.1442
diabetic retinography	**Ada-WHIPS**	1.39	LORE	2.20	344	6.76	≈ 0**
cleveland heart	LORE	1.63	Ada-WHIPS	1.82	90	0.8158	0.2072
mental health survey '16	**Ada-WHIPS**	1.51	Anchors	2.22	429	6.19	≈ 0**
SAMME.R							
breast	Ada-WHIPS	1.48	LORE	1.70	170	1.29	0.0980
cardiotocography	LORE	1.52	Ada-WHIPS	1.62	637	1.14	0.1269
diabetic retinography	**Ada-WHIPS**	1.57	Anchors	2.17	344	4.98	0.0000**
cleveland heart	LORE	1.50	Ada-WHIPS	1.86	90	1.52	0.0649
mental health survey '16	Ada-WHIPS	1.68	Anchors	1.80	429	1.04	0.1492

Table 8: Coverage: Mean rank (mrnk) for two-way comparisons.

data	1st	mrnk	2nd	mrnk	N	V	p.value
SAMME							
mental health survey '14	**Ada-WHIPS**	1.16	Anchors	1.84	377	66	≈ 0**
hospital readmission	**Ada-WHIPS**	1.01	Anchors	1.98	1000	782.5	≈ 0**
thyroid	**Ada-WHIPS**	1.10	Anchors	1.90	1000	14806	≈ 0**
understanding society	**Ada-WHIPS**	1.20	Anchors	1.80	1000	858	≈ 0**
SAMME.R							
mental health survey '14	**Ada-WHIPS**	1.13	Anchors	1.87	377	119	≈ 0**
hospital readmission	**Ada-WHIPS**	1.33	Anchors	1.67	1000	174990	≈ 0**
thyroid	**Ada-WHIPS**	1.02	Anchors	1.98	1000	1754	≈ 0**
understanding society	**Ada-WHIPS**	1.07	Anchors	1.93	1000	6417	≈ 0**

or compliance requirements of the application. We would not, however, expect a failure rate at the upper end of this range ever to be acceptable. Secondly, such over-fitting is symptomatic of an algorithm that generates rules that are overly long; having too many terms in the antecedent to be easily interpretable. To show the link between over-fitting and rule length we present the rule length distribution in Fig 9.

We present the results of the hypothesis tests in Tables 10-11. Clearly, Anchors dominates out of the three algorithms on a statistical test of median differences. However, we have shown that these results should be taken with caution. To begin with, Anchors required us to discretise the data as a preprocessing step, which resulted in alternative models that were less accurate classifiers. The difference was two or more percentage points in 7/9 for SAMME models and 5/9 for SAMME.R models. Moreover, Anchors has a long tail distribution of rule length, and sometimes a high proportion of critically over-fitting explanations. The tabulated means of precision do not show a clear difference between Ada-WHIPS and Anchors (see Appendix). Moreover, precision (specificity) is in a trade-off with coverage (generality). Rules that are too specific only apply to a small fraction of other instances. Ada-WHIPS makes a very small trade-off (just a percentage point or two in most cases), and delivers much more generalisable rules that rarely, if ever, over-fit. This behaviour is the result of optimising the novel objective function, stability (equation 8).

Stability Analysis

Stability can also be used as a quality measure in the XAI setting. A precision of 0.0 for an explanation on an held-out test set can be caused by sampling artefacts. There

Figure 5: Mean Precision SAMME

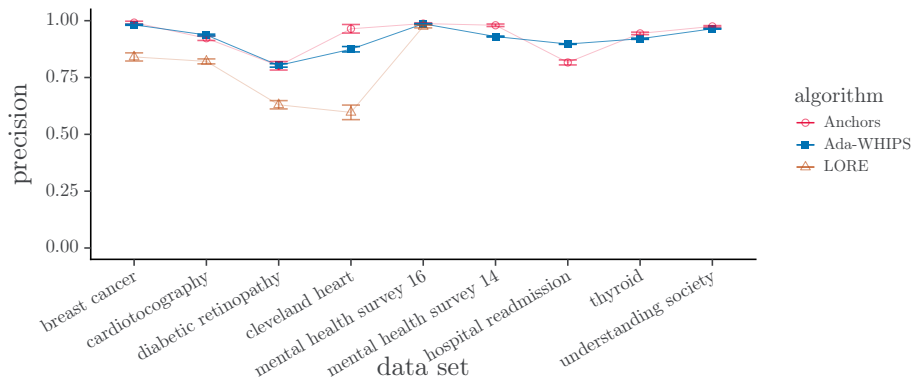
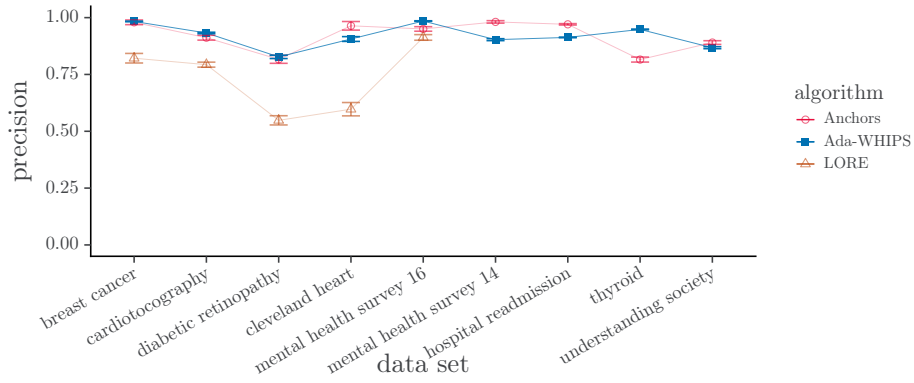


Figure 6: Mean Precision SAMME.R



is a non-zero probability that the key attribute(s) are under-represented in the test set. For this reason, it can be argued that a precision of 0.0 is a harsh penalty against the aggregate score. Yet, if the rule covers and is correct for just a single instance in the held out set, the precision will be 1.0, a discontinuity that masks over-fitting. Instead of precision, we can measure stability while including the explanandum in the held out set. This condition results in the formulation $\frac{n+1}{m+K}$ where n is the number of covered and correct instances, and m is the number of covered instances. Thus, stability is a classical additive smoothing function. The minimum/maximum are both $\frac{1}{1+K}$ for $N = 1$ but approach 0/1 asymptotically as $N \to \infty$. We present the visual analysis of stability in Figs 10=11 and the results of the hypothesis tests in Tables 12-13. The post-hoc tests for the three-way comparisons show that Ada-WHIPS is the top or in the top two with no statistical difference in all except mental health survey '16 for the SAMME model. For the two-way comparisons, Ada-WHIPS has a significantly higher rank for hospital readmission (SAMME) and thyroid (SAMME.R) but lower for the remaining results.

Figure 7: Distributions of Precision SAMME

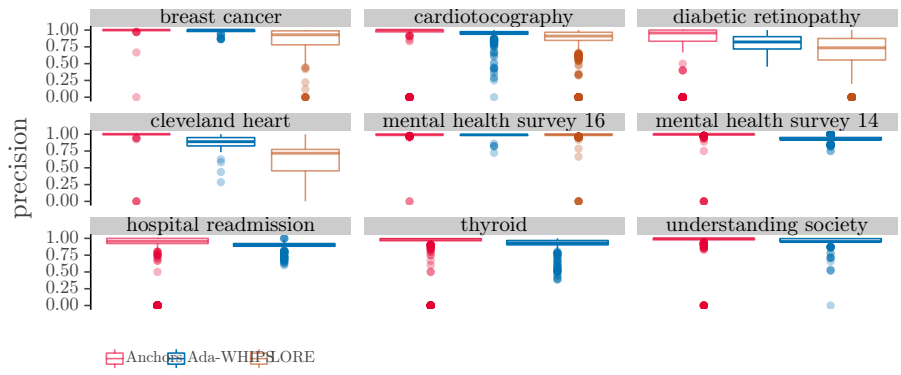
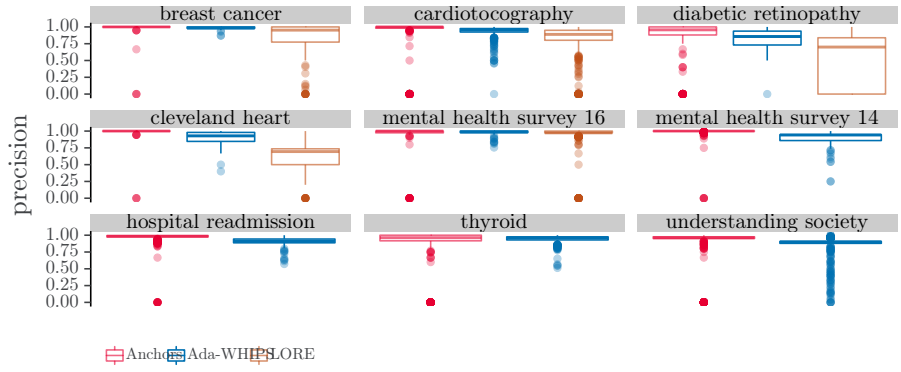


Figure 8: Distributions of Precision SAMME.R



Computation Time Analysis

Finally, we show the distribution of computation time per explanation in Fig 12. A brief visual inspection shows that Ada-WHIPS and Anchors are roughly comparable for all data sets. The shortest run-times are fractions of a second and the longest are two to three minutes. LORE runs at several orders of magnitude longer than this. As we discussed in previous sections, it was prohibitive to run LORE for the data sets mental health survey '14, hopital readmission, thyroid and understanding society with a single explanation taking over two hours to generate. We performed both static and dynamic analysis of the LORE source code and discovered that the bottle-neck was in a non-parallelisable, genetic-algorithmic step.

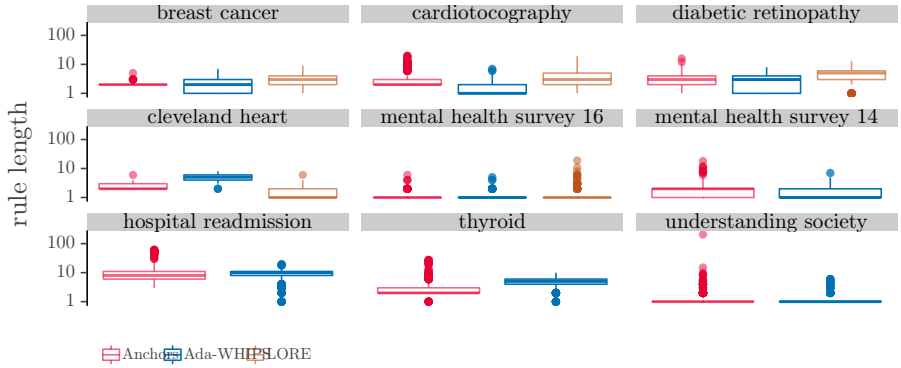
Conclusion & Future Work

Our contribution is the novel algorithm Ada-WHIPS for explaining the classification of AdaBoost models with simple classification rules. AdaBoost models are widely adopted as computer aided diagnostic tools and the non-clinical identification of sub-health and mental health conditions using unconventional data sources such as online health communities. Our method improves on prior research in that it delivers explanations that have high mean coverage (15%-68%). Ada-WHIPS

Table 9: Proportion of Over-fitting, 0.0 Precision Explanations.

	SAMME			SAMME.R		
data	Ada-WHIPS	Anchors	LORE	Ada-WHIPS	Anchors	LORE
breast cancer	0	0.01	0.04	0	0.18	0.06
cardiotocography	0.00	0.07	0.09	0.00	0.08	0.09
diabetic retinopathy	0	0.15	0.19	0.00	0.13	0.28
cleveland heart	0	0.03	0.14	0	0.03	0.12
mental health survey '16	0	0.00	0.01	0	0.04	0.06
mental health survey '14	0	0.01	N/A	0	0.01	N/A
hospital readmission	0	0.15	N/A	0	0.01	N/A
thyroid	0	0.03	N/A	0	0.15	N/A
understanding society	0.00	0.01	N/A	0.01	0.08	N/A

Figure 9: Distributions of Rule Length. Note the y-axis is $\log_{10}$ scaled.



explanations generalise well while making only a very small trade-off to keep precision/specificity competitive (80%-99%). At the same time, Ada-WHIPS is guarded against over-fitting while competing methods have the tendency to present critically over-fitting explanations, in 0.05%-28% of cases. A critically over-fitting explanation is defined as an explanation that uniquely identifies the explanandum in a held out test set. Ada-WHIPS does not make any assumptions about the underlying data distribution, while some competing methods require continuous features to be discretised prior to model training. This treatment of the data can result in a less accurate model, detracting from the main benefit of using AdaBoost at the outset. By design, Ada-WHIPS rules extract discrete, logical conditions from the base decision tree classifiers of the AdaBoost model. These logical conditions have an information-theoretic derivation and we speculate that this is what leads to Ada-WHIPS's favourable trade-off between precision and coverage. Ada-WHIPS is efficient. At its fastest, explanations are generated in fractions of seconds. On high dimensional data sets, we recorded times of up to three minutes per explanation. This is in line with competing methods and could still be considered real-time in the context of a medical consultation. As a minor contribution, we presented stability, a novel measure that is a regularised version of precision. It gives more informative results in the XAI setting as it penalises low coverage while correcting for sampling artefacts.

Table 10: Precision: Top two by mean rank (mrnk) for three-way comparisons.

data	1st	mrnk	2nd	mrnk	N	z	p.value
SAMME							
breast	**Anchors**	1.40	Ada-WHIPS	1.97	170	3.31	0.0004**
cardiotocography	**Anchors**	1.39	Ada-WHIPS	2.09	637	7.89	≈ 0**
diabetic retinography	**Anchors**	1.62	Ada-WHIPS	1.96	344	2.85	0.0022**
cleveland heart	**Anchors**	1.16	Ada-WHIPS	2.03	90	3.68	0.0001**
mental health survey '16	Anchors	1.83	LORE	1.95	429	1.02	0.1539
SAMME.R							
breast	**Anchors**	1.35	Ada-WHIPS	2.08	170	4.38	< 0.0001**
cardiotocography	**Anchors**	1.28	Ada-WHIPS	2.09	637	9.16	≈ 0**
diabetic retinography	**Anchors**	1.50	Ada-WHIPS	1.92	344	3.47	0.0002**
cleveland heart	**Anchors**	1.24	Ada-WHIPS	1.90	90	2.77	0.0028**
mental health survey '16	Anchors	1.83	Ada-WHIPS	1.84	429	0.08	0.4678

Table 11: Precision: Mean rank (mrnk) for two-way comparisons.

data	1st	mrnk	2nd	mrnk	N	V	p.value
SAMME							
mental health survey '14	**Anchors**	1.11	Ada-WHIPS	1.89	377	45074	≈ 0**
hospital readmission	**Anchors**	1.24	Ada-WHIPS	1.76	1000	333580	≈ 0**
thyroid	**Anchors**	1.19	Ada-WHIPS	1.81	1000	405600	≈ 0**
understanding society	**Anchors**	1.08	Ada-WHIPS	1.92	1000	458060	≈ 0**
SAMME.R							
mental health survey '14	**Anchors**	1.11	Ada-WHIPS	1.89	377	45281	≈ 0**
hospital readmission	**Anchors**	1.07	Ada-WHIPS	1.93	1000	480520	≈ 0**
thyroid	Anchors	1.47	Ada-WHIPS	1.53	1000	233670	0.1601
understanding society	**Anchors**	1.31	Ada-WHIPS	1.69	1000	266150	≈ 0**

Directions for future work include developing the method for Gradient Boosting Machines such as XGBoost that use decision trees as the base classifiers, and applying the proposed method on a variation of healthcare and medical data sets.

Ethics approval and consent to participate
Not applicable

Consent for publication

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
The source code and data sets analysed during the current study are available in our repository: https://tinyurl.com/yxuhfh4e.

Funding
Not applicable

Authors' contributions
Original concept was by JH and MMG. JH was the major contributor in developing the software, designing and executing the experiments, analysing the data and writing the manuscript. All authors read and approved the final manuscript.

Acknowledgements
Not applicable

References

[1] Shaker El-Sappagh et al. "An ontology-based interpretable fuzzy decision support system for diabetes diagnosis". In: *IEEE Access* 6 (2018), pp. 37371–37394.

[2] Sandra Wachter, Brent Mittelstadt, and Chris Russell. "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR". In: *Harvard Journal of Law & Technology* 31.2 (2017).

Figure 10: Mean Stability SAMME

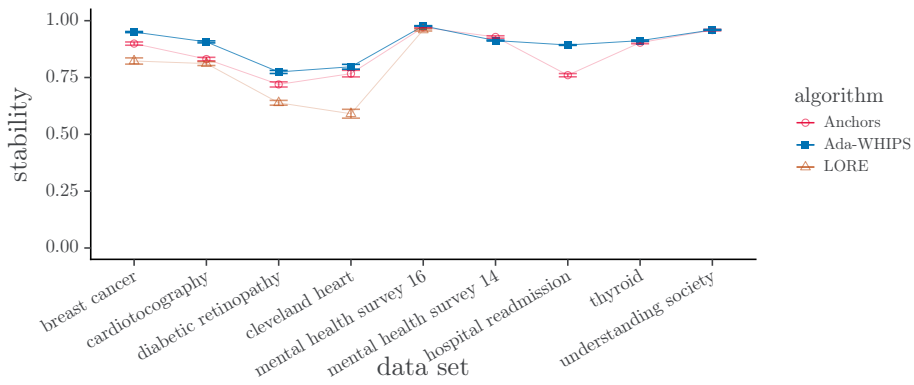
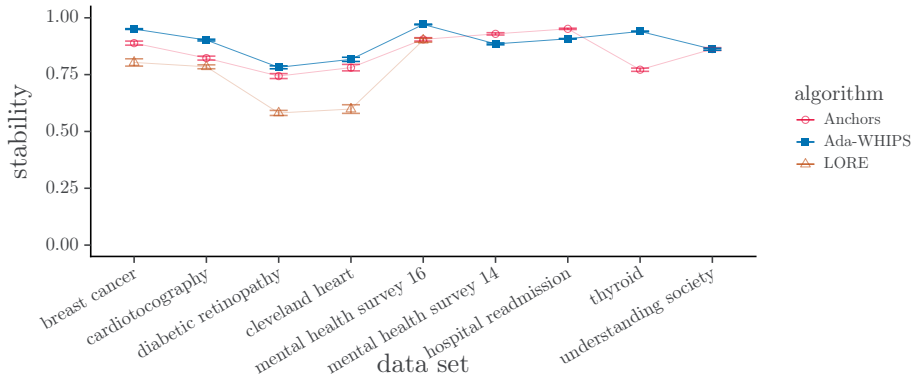


Figure 11: Mean Stability SAMME.R



[3] Rich Caruana et al. "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*. the 21th ACM SIGKDD International Conference. Sydney, NSW, Australia: ACM Press, 2015, pp. 1721–1730.

[4] Ioannis Kavakiotis et al. "Machine Learning and Data Mining Methods in Diabetes Research". In: *Computational and Structural Biotechnology Journal* 15 (2017), pp. 104–116.

[5] Adrin Jalali and Nico Pfeifer. "Interpretable per case weighted ensemble method for cancer associations". In: *BMC Genomics* 17.1 (Dec. 2016).

[6] Zhijun Yin, Lina M Sulieman, and Bradley A Malin. "A systematic literature review of machine learning in online personal health data". In: *Journal of the American Medical Informatics Association* 26.6 (June 1, 2019), pp. 561–576.

[7] Sheng Sun et al. "Subhealth state classification with AdaBoost learner". In: *International Journal of Functional Informatics and Personalised Medicine* 4.2 (2013), p. 167.

Table 12: Stability: Top two by mean rank (mrnk) for three-way comparisons.

data	1st	mrnk	2nd	mrnk	N	z	p.value
SAMME							
breast	**Ada-WHIPS**	1.48	Anchors	2.16	170	3.96	< 0.0001**
cardiotocography	**Ada-WHIPS**	1.48	Anchors	2.19	637	7.99	≈ 0**
diabetic retinography	Ada-WHIPS	1.70	Anchors	1.84	344	1.18	0.1198
cleveland heart	Ada-WHIPS	1.60	Anchors	1.70	90	0.42	0.3374
mental health survey '16	Anchors	1.87	LORE	2.00	429	1.14	0.1269
SAMME.R							
breast	**Ada-WHIPS**	1.38	Anchors	2.18	170	4.67	≈ 0**
cardiotocography	**Ada-WHIPS**	1.49	LORE	2.10	637	6.80	≈ 0**
diabetic retinography	Ada-WHIPS	1.64	Anchors	1.67	344	0.24	0.4050
cleveland heart	Ada-WHIPS	1.49	Anchors	1.73	90	0.98	0.1638
mental health survey '16	**Ada-WHIPS**	1.44	LORE	2.18	429	6.45	≈ 0**

Table 13: Stability: Mean rank (mrnk) for two-way comparisons.

data	1st	mrnk	2nd	mrnk	N	V	p.value
SAMME							
mental health survey '14	**Anchors**	1.19	Ada-WHIPS	1.81	377	39293	≈ 0**
hospital readmission	**Ada-WHIPS**	1.43	Anchors	1.57	1000	136050	≈ 0**
thyroid	**Anchors**	1.35	Ada-WHIPS	1.65	1000	307840	≈ 0**
understanding society	**Anchors**	1.14	Ada-WHIPS	1.86	1000	405340	≈ 0**
SAMME.R							
mental health survey '14	**Anchors**	1.19	Ada-WHIPS	1.81	377	40515	≈ 0**
hospital readmission	**Anchors**	1.14	Ada-WHIPS	1.86	1000	439750	≈ 0**
thyroid	**Ada-WHIPS**	1.18	Anchors	1.82	1000	50600	≈ 0**
understanding society	**Anchors**	1.39	Ada-WHIPS	1.61	1000	220150	≈ 0**

[8] Milos Jovanovic et al. "Building interpretable predictive models for pediatric hospital readmission using Tree-Lasso logistic regression". In: *Artificial Intelligence in Medicine* 72 (Sept. 2016), pp. 12–21.

[9] Lior Turgeman and Jerrold H. May. "A mixed-ensemble model for hospital readmission". In: *Artificial Intelligence in Medicine* 72 (Sept. 2016), pp. 72–82.

[10] Benjamin Letham et al. "Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model". In: *The Annals of Applied Statistics* 9.3 (Sept. 2015), pp. 1350–1371.

[11] Konstantina Kourou et al. "Machine learning applications in cancer prognosis and prediction". In: *Computational and Structural Biotechnology Journal* 13 (2015), pp. 8–17.

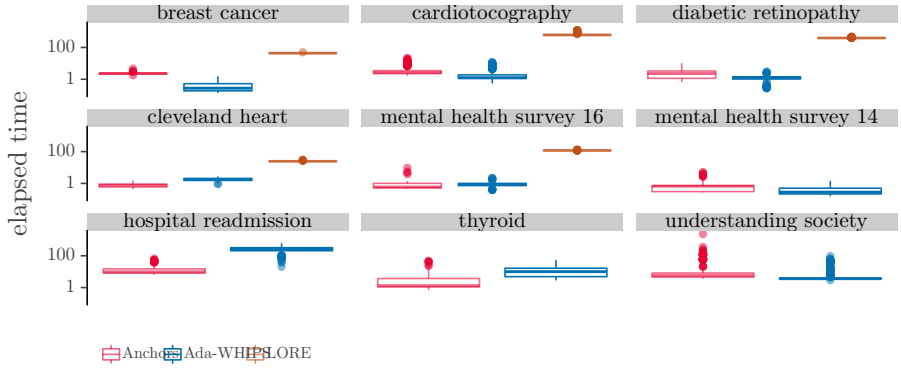
[12] Muhammad Subianto and Arno Siebes. "Understanding Discrete Classifiers with a Case Study in Gene Prediction". In: IEEE, Oct. 2007, pp. 661–666.

[13] Johan Huysmans, Bart Baesens, and Jan Vanthienen. "Using Rule Extraction to Improve the Comprehensibility of Predictive Models". In: *SSRN Electronic Journal* (2006). URL: https://tinyurl.com/y79jk4xx (visited on 11/16/2018).

[14] M. J. Pazzani, S. Mani, and W. R. Shankle. "Acceptance of Rules Generated by Machine Learning among Medical Experts". In: *Methods of Information in Medicine* 40.5 (2001), pp. 380–385.

[15] European Parliament {and} Council of the European Union. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of*

Figure 12: Distributions of Computation Time per Explanation. Note the y-axis is $\log_{10}$ scaled.



personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). May 25, 2018.

[16] Vijay Pande. "Artificial Intelligence's 'Black Box' Is Nothing to Fear". In: *The New York Times* (Jan. 2019). URL: https://tinyurl.com/y9gsu4u2 (visited on 08/14/2019).

[17] Dino Pedreschi et al. "Open the Black Box Data-Driven Explanation of Black Box Decision Systems". In: *arXiv:1806.09936 [cs]* (June 26, 2018).

[18] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?: Explaining the Predictions of Any Classifier". In: ACM Press, 2016, pp. 1135–1144.

[19] Yoav Freund. "An adaptive version of the boost by majority algorithm". In: *Proceedings of the twelfth annual conference on Computational learning theory - COLT '99*. the twelfth annual conference. Santa Cruz, California, United States: ACM Press, 1999, pp. 102–113. ISBN: 978-1-58113-167-3. DOI: 10.1145/307400.307419. URL: http://portal.acm.org/citation.cfm?doid=307400.307419 (visited on 04/16/2019).

[20] Shadnaz Asgari, Fabien Scalzo, and Magdalena Kasprowicz. "Pattern Recognition in Medical Decision Support". In: *BioMed Research International* 2019 (June 13, 2019), pp. 1–2.

[21] U. Rajendra Acharya et al. "Computer-aided diagnosis of diabetic subjects by heart rate variability signals using discrete wavelet transform method". In: *Knowledge-Based Systems* 81 (June 2015), pp. 56–64.

[22] Illhoi Yoo et al. "Data Mining in Healthcare and Biomedicine: A Survey of the Literature". In: *Journal of Medical Systems* 36.4 (Aug. 2012), pp. 2431–2448.

[23] Martin Dolejsi et al. "Reducing false positive responses in lung nodule detector system by asymmetric adaboost". In: *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. 2008 5th IEEE International Symposium on Biomedical Imaging (ISBI 2008). Paris, France: IEEE, May 2008, pp. 656–659.

[24] P. Mohamed Shakeel et al. "Automatic detection of lung cancer from biomedical data set using discrete AdaBoost optimized ensemble learning generalized neural networks". In: *Neural Computing and Applications* (Jan. 10, 2019).

[25] M Rangini and Dr G Wiselin Jiji. "Identification of Alzheimer's Disease Using Adaboost Classifier". In: *Proceedings of the International Conference on Applied Mathematics and Theoretical Computer Science*. 2013, pp. 229–234.

[26] Daniel Hernandez-Lobato et al. "Pruning Adaptive Boosting Ensembles by Means of a Genetic Algorithm". In: *Proceedings of the 7th International Conference Intelligent Data Engineering and Automated Learning*. IDEAL 2006. Vol. 4224. Lecture Notes In Computer Science. Burgos, Spain: Springer, 2006, pp. 322–329.

[27] Christino Tamon and Jie Xiang. "On the Boosting Pruning Problem". In: *Machine Learning: ECML 2000*. 11th European COnference on Machine Learning. Vol. 1180. Lecture Notes In Computer Science. Barcelona: Springer, 2000, pp. 404–412.

[28] Dragos D. Margineantu and Thomas G. Dietterich. "Pruning Adaptive Boosting". In: *ICML*. Vol. 97. 1997, pp. 211–218.

[29] Robert Andrews, Joachim Diederich, and Alan B. Tickle. "Survey and critique of techniques for extracting rules from trained artificial neural networks". In: *Knowledge-Based Systems* 8.6 (1995), pp. 373–389.

[30] Benjamin Letham. "Statistical Learning for Decision Making: Interpretability, Uncertainty, and Inference". PhD thesis. Massachusetts Institute of Technology, 2015. 196 pp.

[31] Satoshi Hara and Kohei Hayashi. "Making Tree Ensembles Interpretable: A Bayesian Model Selection Approach". In: *arXiv:1606.09066 [stat]* (June 29, 2016).

[32] Md Nasim Adnan and Md Zahidul Islam. "ForEx++: A New Framework for Knowledge Discovery from Decision Forests". In: *Australasian Journal of Information Systems* 21 (Nov. 8, 2017).

[33] Morteza Mashayekhi and Robin Gras. "Rule Extraction from Random Forest: the RF+HC Methods". In: *Advances in artificial intelligence 2015*. 28th Canadian Conference on Artificial Intelligence, Canadian AI. Vol. 9091. Lecture notes in computer science Artificial intelligence. Halifax, Nova Scotia, Canada: Springer, 2015, pp. 223–237.

[34] Houtao Deng. "Interpreting tree ensembles with intrees". In: *International Journal of Data Science and Analytics* 7.4 (2014), pp. 277–87.

[35] Jerome Friedman and Bogdan E Popescu. "Predictive Learning via Rule Ensembles". In: *The Annals of Applied Statistics* 2.3 (2008), pp. 916–954.

[36] Lemuel R. Waitman, Douglas H. Fisher, and Paul H. King. "Bootstrapping rule induction to achieve rule stability and reduction". In: *Journal of Intelligent Information Systems* 27.1 (July 2006), pp. 49–77.

[37] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Anchors: High-Precision Model-Agnostic Explanations". In: *Conference on Artificial Intelligence, 2018*. AAAI. New Orleans, 2018.

[38] Zachary Chase Lipton. "The mythos of model interpretability". In: *arXiv preprint arXiv:1606.03490*. 2016.

[39] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems* (2017), pp. 4768–4777.

[40] Riccardo Guidotti et al. "Local Rule-Based Explanations of Black Box Decision Systems". In: *arXiv:1805.10820* (May 28, 2018).

[41] Fracek Michal. *"Please, explain." Interpretability of black-box machine learning models*. Data Science Central. 2019. URL: https://tinyurl.com/y5qruqgf (visited on 04/19/2019).

[42] Hui Fen et al. "Why should you trust my interpretation? Understanding uncertainty in LIME predictions". In: *arXiv:1904.12991* (Apr. 29, 2019).

[43] Amina Adadi and Mohammed Berrada. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)". In: *IEEE Access* 6 (2018), pp. 52138–52160.

[44] Ando Sabaas. *Interpreting Random Forests*. Diving Into Data. Oct. 19, 2014. URL: http://blog.datadive.net/interpreting-random-forests/ (visited on 10/11/2017).

[45] Scott M. Lundberg and Su-In Lee. "Consistent feature attribution for tree ensembles". In: *arXiv:1706.06060 [cs, stat]*. ICML Workshop on Human Interpretability in Machine Learning. Sydney, Australia, June 19, 2017.

[46] Ron Appel et al. "Quickly Boosting Decision Trees–Pruning Underachieving Features Early". In: *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*. 2013, pp. 594–602.

[47] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. "Additive Logistic Regression A Statistical View of Boosting". In: *The Annals of Statistics* 28.2 (2000), pp. 337–407.

[48] Yoav Freund and Robert E Schapire. "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting". In: *Journal of Computer and System Sciences* 55.1 (Aug. 1997), pp. 119–139.

[49] Kevin W. Walker and Zhehan Jiang. "Application of adaptive boosting (AdaBoost) in demand-driven acquisition (DDA) prediction: A machine-learning approach". In: *The Journal of Academic Librarianship* 45.3 (May 2019), pp. 203–212.

[50] K Aravindh et al. "A Novel Data Mining approach for Personal Health Assistance." In: *International Journal of Pure and Applied Mathematics.* 119.15 (2018), pp. 415–426.

[51] Thongkam Jaree et al. "Breast cancer survivability via AdaBoost algorithms". In: *Proceedings of the second Australasian workshop on Health data and knowledge management*. HDKM '08. Vol. 80. Wollongong, NSW, Australia: Australian Computer Society, 2008, pp. 55–64.

[52] Trevor Hastie et al. "Multi-class AdaBoost". In: *Statistics and Its Interface* 2.3 (2009), pp. 349–360.

[53] Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[54] Yoav Freund and Robert E Schapire. "A Short Introduction to Boosting". In: *Journal of Japanese Society for Artificial Intelligence* 14.5 (1999), pp. 771–780.

[55] J R Quinlan. "Generating Production Rules From Decision Trees". In: *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*. IJCAI. Milan, Italy, 1987, pp. 304–307.

[56] Amit Dhurandhar et al. "Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives". In: *arXiv:1802.07623 [cs]* (Feb. 21, 2018).

[57] Dua Dheeru and Efi Karra Taniskidou. *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences, 2017. URL: https://archive.ics.uci.edu/ml/datasets/.

[58] *Understanding Society: Waves 2-3 Nurse Health Assessment, 2010-2012*. 3rd. Vol. 7251. University of Essex, Institute for Social and Economic Research: UK Data Service, 2019.

[59] Apostolos Davillas, Michaela Benzeval, and Meena Kumari. "Association of Adiposity and Mental Health Functioning across the Lifespan: Findings from Understanding Society (The UK Household Longitudinal Study)". In: *PLOS ONE* 11.2 (Feb. 5, 2016). Ed. by David Meyre.

[60] Janez Demsar. "Statistical Comparisons of Classifiers over Multiple Data Sets". In: *Journal of Machine learning research* (2006), pp. 1–30.

Appendix

Table 14: Coverage of Explanations of AdaBoost SAMME

data	Ada-WHIPS	Anchors	LORE
breast cancer	0.3635 ± 0.0068	0.1530 ± 0.0053	0.3914 ± 0.0156
cardiotocography	0.3867 ± 0.0092	0.0637 ± 0.0018	0.4417 ± 0.0120
diabetic retinopathy	0.3225 ± 0.0125	0.0636 ± 0.0039	0.1060 ± 0.0060
cleveland heart	0.2310 ± 0.0084	0.1101 ± 0.0079	0.3259 ± 0.0259
mental health survey '16	0.4974 ± 0.0026	0.3915 ± 0.0083	0.3777 ± 0.0086
mental health survey '14	0.3368 ± 0.0063	0.1483 ± 0.0030	N/A
hospital readmission	0.1809 ± 0.0040	0.0095 ± 0.0004	N/A
thyroid	0.3630 ± 0.0074	0.0636 ± 0.0015	N/A
understanding society	0.6679 ± 0.0108	0.2729 ± 0.0040	N/A

Table 15: Coverage of Explanations of AdaBoost SAMME.R

data	Ada-WHIPS	Anchors	LORE
breast cancer	0.33502 ± 0.0055	0.1513 ± 0.0054	0.3574 ± 0.0157
cardiotocography	0.3894 ± 0.0093	0.0667 ± 0.0019	0.4765 ± 0.0128
diabetic retinopathy	0.1349 ± 0.0053	0.0759 ± 0.0040	0.0945 ± 0.0068
cleveland heart	0.2182 ± 0.0085	0.1180 ± 0.0078	0.3754 ± 0.0271
mental health survey '16	0.3578 ± 0.0054	0.1778 ± 0.0072	0.3248 ± 0.0101
mental health survey '14	0.2927 ± 0.0053	0.1444 ± 0.0030	N/A
hospital readmission	0.1598 ± 0.0038	0.1345 ± 0.0042	N/A
thyroid	0.3793 ± 0.0073	0.0224 ± 0.0008	N/A
understanding society	0.6891 ± 0.0107	0.1057 ± 0.0038	N/A

Table 16: Precision of Explanations of AdaBoost SAMME

data	Ada-WHIPS	Anchors	LORE
breast cancer	0.9819 ± 0.0022	0.9915 ± 0.0062	0.8405 ± 0.0179
cardiotocography	0.9369 ± 0.0039	0.9915 ± 0.0097	0.8209 ± 0.0109
diabetic retinopathy	0.8031 ± 0.0075	0.8016 ± 0.0188	0.6300 ± 0.0182
cleveland heart	0.8744 ± 0.0118	0.9644 ± 0.0189	0.6300 ± 0.0321
mental health survey '16	0.9862 ± 0.0010	0.9873 ± 0.0035	0.9744 ± 0.0061
mental health survey '14	0.9301 ± 0.0021	0.9798 ± 0.0056	N/A
hospital readmission	0.8973 ± 0.0016	0.8163 ± 0.0110	N/A
thyroid	0.9205 ± 0.0026	0.9441 ± 0.0055	N/A
understanding society	0.9643 ± 0.0016	0.9749 ± 0.0035	N/A

Table 17: Precision of Explanations of AdaBoost SAMME.R

data	Ada-WHIPS	Anchors	LORE
breast cancer	0.9831 ± 0.0014	0.9793 ± 0.0103	0.8215 ± 0.0210
cardiotocography	0.9324 ± 0.0032	0.9117 ± 0.0107	0.7931 ± 0.0110
diabetic retinopathy	0.8272 ± 0.0073	0.8164 ± 0.0175	0.5481 ± 0.0203
cleveland heart	0.9059 ± 0.0105	0.9640 ± 0.0189	0.5971 ± 0.0293
mental health survey '16	0.9849 ± 0.0013	0.9502 ± 0.0100	0.9129 ± 0.0124
mental health survey '14	0.9030 ± 0.0043	0.9811 ± 0.0056	N/A
hospital readmission	0.9129 ± 0.0013	0.9811 ± 0.0032	N/A
thyroid	0.9481 ± 0.0015	0.8154 ± 0.0110	N/A
understanding society	0.8677 ± 0.0043	0.8903 ± 0.0081	N/A

Table 18: Stability of Explanations of AdaBoost SAMME

data	Ada-WHIPS	Anchors	LORE
breast cancer	0.9500 ± 0.0024	0.8992 ± 0.0072	0.8226 ± 0.0137
cardiotocography	0.9067 ± 0.0044	0.8311 ± 0.0078	0.8113 ± 0.0085
diabetic retinopathy	0.7745 ± 0.0067	0.7196 ± 0.0114	0.6388 ± 0.0106
cleveland heart	0.7973 ± 0.0106	0.7671 ± 0.0145	0.5906 ± 0.0195
mental health survey '16	0.9770 ± 0.0011	0.9706 ± 0.0053	0.9592 ± 0.0046
mental health survey '14	0.9125 ± 0.0021	0.9283 ± 0.0053	N/A
hospital readmission	0.8930 ± 0.0017	0.7306 ± 0.0071	N/A
thyroid	0.9121 ± 0.0028	0.9033 ± 0.0047	N/A
understanding society	0.9594 ± 0.0017	0.9586 ± 0.0035	N/A

Table 19: Stability of Explanations of AdaBoost SAMME.R

data	Ada-WHIPS	Anchors	LORE
breast cancer	0.9505 ± 0.0017	0.8885 ± 0.0089	0.8035 ± 0.161
cardiotocography	0.9020 ± 0.0038	0.8226 ± 0.0087	0.7844 ± 0.0086
diabetic retinopathy	0.7821 ± 0.0064	0.7436 ± 0.0109	0.5814 ± 0.0111
cleveland heart	0.8171 ± 0.0092	0.7807 ± 0.0143	0.5985 ± 0.0190
mental health survey '16	0.9707 ± 0.0015	0.9051 ± 0.0073	0.9013 ± 0.0088
mental health survey '14	0.8852 ± 0.0041	0.9293 ± 0.0051	N/A
hospital readmission	0.9075 ± 0.0029	0.9514 ± 0.0029	N/A
thyroid	0.9401 ± 0.0015	0.7716 ± 0.0071	N/A
understanding society	0.8616 ± 0.0043	0.8624 ± 0.0063	N/A