Transparency Promotion with Model-Agnostic Linear Competitors

Hassan Rafique [1] **Tong Wang** [2] **Qihang Lin** [2] **Arshia Sighani** [3]

Abstract

We propose a novel type of hybrid model for multi-class classification, which utilizes competing linear models to collaborate with an existing black-box model, promoting transparency in the decision-making process. Our proposed hybrid model, Model-Agnostic Linear Competitors (MALC), brings together the interpretable power of linear models and the good predictive performance of the state-of-the-art black-box models. We formulate the training of a MALC model as a convex optimization problem, optimizing the predictive accuracy and transparency (defined as the percentage of data captured by the linear models) in the objective function. Experiments show that MALC offers more model flexibility for users to balance transparency and accuracy, in contrast to the currently available choice of either a pure black-box model or a pure interpretable model. The human evaluation also shows that more users are likely to choose MALC for this model flexibility compared with interpretable models and black-box models.

1. Introduction

There has been an increasing need for modern machine learning models to provide accurate and interpretable predictions to assist humans in decision making, especially in high stakes applications such as healthcare, judiciaries, etc. (Letham et al., 2015; Yang et al., 2018; Caruana et al., 2015; Chen et al., 2018). Thus, many state-of-the-art machine learning models, such as neural networks and ensembles, stumble in these domains since they are *black-box* in nature. Black-box models have an opaque or highly complicated decision-making process that is hard for a human to understand and rationalize. Driven by the practical needs, researchers have shifted their focus to account for both transparency and predictive performance of models in recent years.

Due to the extended interest and effort, various forms of interpretable models have been proposed (Wang et al., 2017; Zeng et al., 2017; Richter & Weber, 2016). However, interpretability has multiple goals that are not always aligned with the production of the most generalizable model architecture (Lipton, 2018), especially when users have strict and domain-specific requirements for interpretability. Thus the performance loss is often prevalent when dealing with complicated predictive tasks, putting users in the dilemma of choosing between interpretability and predictive performance.

In this paper, we propose a new form of a model, Model-Agnostic Linear Competitors (MALC), for multi-class classification. MALC utilizes linear models to predict a subset of data while leaving the remaining possibly harder predictions to a black-box model. The model combines the intuitive power of interpretable models and the good predictive performance of black-box models to reach some controllable middle ground where both transparency and good predictive performance is possible.

To build a hybrid model for multi-class classification, we design a unique mechanism to utilize competition and collaboration among the participating models. Given a $K-$class classification problem, we design $K+1$ models, which we call *competitors*. K of the competitors are *interpretable*, capturing K classes, respectively. The remaining one is a pre-trained black-box model, called competitor $\mathcal{B}$. Given an input $\mathbf{x}$, all of the K interpretable competitors bid to claim the input by proposing a score. The input is then assigned to the highest bidder, with a significant margin over the other competitors' scores. If there does not exist a winner (not winning by a large margin), then none of the K competitors can claim the input, and it is then sent to competitor $\mathcal{B}$ by default. At competitor $\mathcal{B}$, the input will be classified, and this classification process is unknown to other competitors the whole time (during training and testing), i.e., *model-agnostic*. The competitor $\mathcal{B}$ can be *any* pre-trained multi-class classifier with high predictive accuracy. We let all interpretable competitors be linear models, which are of-

[1]Program in Applied Mathematical and Computational Sciences, The University of Iowa, Iowa City, Iowa, USA [2]Department of Business Analytics, The University of Iowa, Iowa City, Iowa, USA [3]BASIS Independent Silicon Valley, San Jose, California, USA. Correspondence to: Tong Wang <tong-wang@uiowa.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

ten used in high-stakes decision-making scenarios (Aubert et al., 2010). We call the proposed method Model-Agnostic Linear Competitors (MALC).

MALC partitions the feature space into $K+1$ regions, each claimed by a competitor. Competitor $k(1 \leq k \leq K)$ captures the most representative and confident characteristics of class k by claiming the most plausible area in the feature space for class k. Predictions for this area are inherently interpretable since the competitors are linear models with regularized numbers of non-zero coefficients. So we define the percentage of data in this area claimed by the linear models as *transparency* of MALC. The unclaimed area represents the subspace where none of the linear competitors can "convince" other linear competitors by proposing a high enough score; thus, it is left to the most competent black-box competitor $\mathcal{B}$ to determine. See Figure 1 for an illustration. Meanwhile, the coefficients of the K linear models also show the most distinctive characteristics of each class, providing an intuitive description of the classes.

To train MALC, we formulate a carefully designed convex optimization problem that considers the predictive performance, interpretability of the linear competitors (coefficients regularization), and model transparency. Then we use accelerated proximal gradient method (Nesterov, 2013) to train MALC. By tuning the parameters, MALC can decide to send more or less area to the linear competitors, at the possible cost of the predictive performance.

To evaluate the model, we conduct experiments on public datasets and compare MALC with interpretable baseline models. In addition, to study whether MALC is likely to be accepted by users and understand humans' preferences for transparency and accuracy, we conduct a human evaluation on a group of 72 subjects. Results show that most users can tolerate only up to 5% of accuracy loss if required to use an interpretable model. In contrast, results on the public data show that the interpretable models may lose more than 10% of accuracy. Thus, the majority of users are willing to use a hybrid model instead, especially when the hybrid model can effectively trade-off transparency and accuracy.

The rest of the paper is organized as follows. We review related work in Section 2. The MALC model is presented in Section 3, where we formulate the model and describe the training algorithm. We conduct an experimental evaluation in Section 4 on public datasets and compare them with interpretable baselines. We also present the human evaluation and discuss the findings.

2. Related Work

We have found a few works in the literature on the combination of multiple models (Kohavi, 1996; Towell & Shavlik, 1994). For example, (Kohavi, 1996) combined a decision tree with a Naive Bayes model, (Shin et al., 2000) proposed a system combining neural network and memory-based learning, (Hua & Zhang, 2006) combined SVM and logistic regression, etc. A recent work (Wang et al., 2015) divides feature spaces into regions with sparse oblique tree splitting and assign local sparse additive experts to individual regions. Besides these more isolated efforts, there has been a large body of continuous work on neural-symbolic or neural-expert systems (Garcez et al., 2015) pursued by a relatively small research community over the last two decades and has yielded several significant results (McGarry et al., 1999; Garcez et al., 2012; Taha & Ghosh, 1999; Towell & Shavlik, 1994). This line of research has been carried on to combine deep neural networks with expert systems to improve predictive performance (Hu et al., 2016).

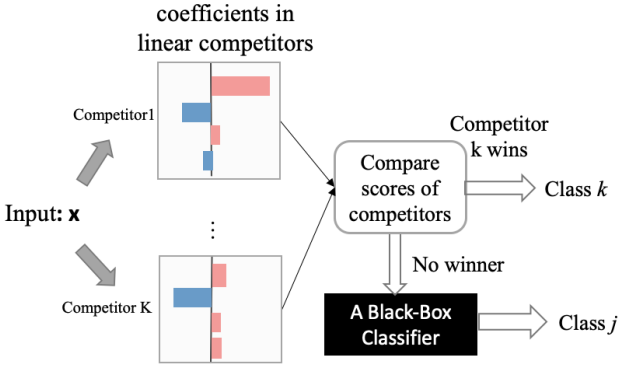


Figure 1. The decision-making process of MALC.

Compared to the models discussed above, our method is distinct in that it is model-agnostic and can work with ***any*** black-box classifier. The black-box can be a carefully calibrated, advanced model using confidential features or techniques. Our hybrid model only needs predictions from the black-box model and does not need to alter it during training or know any other information from it. This minimal requirement of information from the black-box collaborator renders much more flexibility in creating collaboration between different models, mainly preserving confidential information from the more advanced partner.

The idea of a hybrid model is also discussed in (Wang, 2019), which proposes a hybrid rule set (HyRS) to combine decision rules with a black-box model for binary classification. An input goes through a positive rule set, a negative rule set, and a black-box model sequentially until it is classified by the first model that captures it. MALC uses linear models instead, enriching users' choice of models, as some models are more popular than others in different domains. Also, MALC is designed to work with multi-class classification. The K interpretable competitors compete for an input simultaneously in a fair mechanism.

Distinctions from Black-box Explainers We make a clear distinction of our model from black-box explainers (Ribeiro et al., 2018; Lundberg & Lee, 2017). The main ideas of explainers include using simple models like linear models to approximate the predictions of black-box models, or attributing contributions to features and providing feature importance analysis (Lundberg & Lee, 2017), etc. Since the first paper of LIME (Ribeiro et al., 2016), a local linear explainer of any black-box model, various explainer models have been proposed (Ribeiro et al., 2018; Lundberg & Lee, 2017). However, some concerns have been brought up (Rudin, 2019; Aïvodji et al., 2019; Thibault et al., 2019; Slack et al., 2020) on potential issues of black-box explainers. For example, there exists ambiguity and inconsistency (Ross et al., 2017; Lissack, 2016) in the explanations of black-box explainers since there could be different explanations for the same prediction generated by different explainers, or by the same explainer with different parameters. (Alvarez-Melis & Jaakkola, 2018) showed LIME's explanation of two close points (similar instances) could vary greatly. This instability in the explanation demands a cautious and critical use of this type of explainer. Also, recent work demonstrates that explanations can sometimes be deceptive (Aïvodji et al., 2019). (Slack et al., 2020) shows that it is easily possible to fool post hoc explainers like LIME and SHAP (Lundberg & Lee, 2017), which rely on input perturbations, through an adversarial attack. (Slack et al., 2020) proposes a technique to effectively hide the biases of the given classifier and get an arbitrarily desired explanation from LIME and SHAP for that classifier. All of the issues with post hoc explainers result from the fact that the explainers only approximate in a post hoc way. They are not the decision-making process themselves.

We emphasize here that MALC is not a black-box explainer. MALC does not explain or approximate the behavior of a black-box model, but instead, collaborates with the black-box model and shares the prediction task. It is a predictive model, a decision-maker itself.

3. Model-Agnostic Linear Competitors

In this paper, we focus on the multi-class classification problem. Suppose there are K distinct classes. We consider an approach similar to one-vs-all linear classification and review it here. Given a linear classifier $f_i(\mathbf{x}) = \mathbf{w}_i^\top \mathbf{x}$, $i \in [K] := \{1, 2, ..., K\}$, if $f_i(\mathbf{x}) - f_j(\mathbf{x}) \geq 0$, for every j other than i, then $\mathbf{x}$ belongs to class i. For class i,

$$\mathcal{P}_i(\mathbf{x}) = \bigcap_{j \neq i} \Big\{ f_i(\mathbf{x}) - f_j(\mathbf{x}) = 0 \Big\}$$

is the decision boundary. Most mistakes made by a linear model happen around the decision boundary. Therefore, in a hybrid model, we exploit the high predictive power of a black-box model and leave this more difficult area to it while having the linear classifier classify the rest. Then the linear classifier produces a decision only when it is confident enough, this time comparing against thresholds $\{\theta_i \geq 0\}_{i=1}^K$: to predict class i when $f_i(\mathbf{x}) - f_j(\mathbf{x}) \geq \theta_i$ for every j other than i and unclassified otherwise. Thus the linear model generates K decision boundaries, creating a partition of a data space into $K + 1$ regions, a region for each of the K classes and an unclassified region. This unclassified region contains data that the linear model is not confident to decide so the black-box is activated to generate predictions on the unclassified region, see Figure 2. Thus we build K linear competitors, each advocating for a class, to collaborate with the black-box model. We call this classification method Model-Agnostic Linear Competitors (MALC) model.

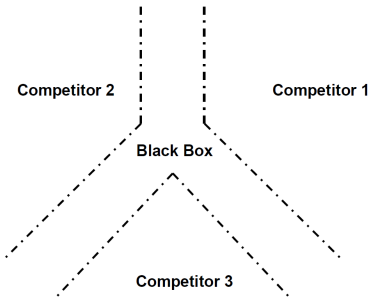


Figure 2. A simplified depiction of partitioning of the data space in the case of three classes, by linear and black-box competitors.

The goal of building such a collaborative linear model is to replace the black-box system with a transparent system on a subset of data at the minimum loss of predictive accuracy. Therefore a key determinant in the success of MALC is the partitioning of the data, which is determined by the coefficients $\mathbf{w}_i$ in the linear model and the thresholds θ_i, $i \in [K]$. In this paper, we formulate a convex optimization problem to learn the coefficients and thresholds. The objective function considers the fitness to the training data, captured by a convex loss function, the regularization term, and the sum of thresholds. As θ_i gets close to 0, more data can be decided by the linear model, increasing the transparency of the decision-making process, but at the cost of possible loss of predictive performance. Our formulation is compatible with various forms of the convex loss function and guarantees global optimality.

We work with a set of training examples $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ where $\mathbf{x}_i \in \mathbb{R}^d$ is a vector of d attributes and $y_i \in [K] := \{1, 2, ..., K\}$ is the corresponding class label. Let $f(\mathbf{x}) : \mathbb{R}^d \to [K]$ represent the MALC classification model that is constructed based on linear models $f_{l,i}(\mathbf{x}) = \mathbf{w}_i^\top \mathbf{x}$, $i \in [K]$ and a black-box model $f_b(\mathbf{x}) : \mathbb{R}^d \to [K]$. The black-box model is given, which can be *any* trained model. We need

its prediction on the training data $\mathcal{D}$, denoted as $\{y_i^b\}_{i=1}^n$ and $y_i^b = f_b(\mathbf{x}_i)$. Our goal is to learn the coefficients $\mathbf{w_i}$ in the linear models $f_{l,i}$ together with thresholds θ_i (≥ 0), $i \in [K]$, in order to form a hybrid decision model f as:

$$f(\mathbf{x}) = \begin{cases} k & \text{if } \mathbf{w}_k^\top \mathbf{x} - \mathbf{w}_j^\top \mathbf{x} \geq \theta_k,\ \forall\, j \in [K] \setminus \{k\} \\ f_b(\mathbf{x}) & \text{otherwise.} \end{cases} \quad (1)$$

Note the hybrid model uses K thresholds to partition the data space into $K+1$ regions, a region for each class, and an undetermined region left to the black-box model. Data falling into any of the K class's claimed regions is considered "transparent" by the linear model, and we refer to the percentage of this data subset as the **transparency** of the model.

3.1. Model Formulation

In this section, we formulate an optimization framework to build a MALC model. We consider three factors when building the model: *predictive performance*, *data transparency*, and *model regularization*. We elaborate each of them below.

The (in-sample) predictive performance characterizes the fitness of the model to the training data. Since f_b is pre-given, the predictive performance is determined by two factors, the accuracy of $f_l = [f_{l,1},\ f_{l,2}, \dots, f_{l,K}]$ on instances as described in (1) and the accuracy of f_b on the remaining examples. We wish to obtain a good partition of data $\mathcal{D}$ by assigning f_b and f_l to different regions of the data such that the strength of f_b and f_l are properly exploited. Second, we include the sum $\sum \theta_i$ as a penalty term in the objective to account for data transparency of the hybrid model. The smaller sum implies that the linear model classifies more data. In the most extreme case where $\sum \theta_i = 0$, all data is sent to the linear model, and the MALC model is reduced to a pure one-vs all linear classifier, i.e., transparency equals one. Finally, we also need to consider model regularization in the objective. As the weight for the sparsity enforcing regularization term increases, the model encourages using a smaller number of features, increasing the interpretability of the model and preventing overfitting.

Combining the three factors discussed above, we formulate the learning objective for MALC as:

$$\min_{\mathbf{w},\theta \geq 0} F(\mathbf{w}, \theta) := \mathcal{L}(\mathbf{w}, \theta; \mathcal{D}) + C_1 \sum_{i=1}^{K} \theta_i + C_2 r(\mathbf{w}), \quad (2)$$

where $\mathbf{w} = [\mathbf{w}_1,\ \mathbf{w}_2, \dots, \mathbf{w}_K]$, $\theta = [\theta_i,\ \theta_2, \dots, \theta_K]$, $\mathcal{L}(\mathbf{w}, \theta; \mathcal{D})$ is the loss function defined on the training set $\mathcal{D}$ associated to the decision rule f in (1), $\sum_{i=1}^K \theta_i$ is a penalty term to increase the transparency of f, r is a convex and closed regularization term (e.g. $\|\mathbf{w}\|_1$, $\frac{1}{2}\|\mathbf{w}\|_2^2$ or an indicator function of a constraint set), and C_1 and C_2 are non-negative coefficients which balance the importance of the three components in (2).

Let $I_k = \{i \mid y_i = k\}$, which is the index set of all the data points ($\mathbf{x}$) belonging to class k. Similarly, let $I_k^+ = \{i \in I_k \mid y_b^i = y_i\}$ and $I_k^- = \{i \in I_k \mid y_b^i \neq y_i\}$. The loss function in (2) over the dataset $\mathcal{D}$ is then defined as

$$\mathcal{L}(\mathbf{w}, \theta; \mathcal{D}) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in I_k^+} \sum_{\substack{j=1 \\ j \neq k}}^{K} \phi(\mathbf{w}_k^\top \mathbf{x}_i - \mathbf{w}_j^\top \mathbf{x}_i + \theta_j)$$
$$+ \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in I_k^-} \sum_{\substack{j=1 \\ j \neq k}}^{K} \phi(\mathbf{w}_k^\top \mathbf{x}_i - \mathbf{w}_j^\top \mathbf{x}_i - \theta_k) \quad (3)$$

where function $\phi(z) : \mathbb{R} \to \mathbb{R}$ is a non-increasing convex closed loss function which can be one of those commonly used in linear classification such as the hinge loss $\phi(z) = (1-z)_+$, smooth hinge loss $\phi(z) = \frac{1}{2}(1-z)_+^2$ or the logistic loss $\phi(z) = \log(1 + \exp(-z))$. Note that $\{I_k = I_k^+ \cup I_k^-\}_{k=1}^K$ form a partition of $\{1, 2, \dots, n\}$. The intuition of this loss function is as follows. Take a data point $\mathbf{x}_i$ with $y_i = k$ and $y_i^b = k$ as an example. Our hybrid model (1) will classify $\mathbf{x}_i$ correctly as long as it does not fall into the region of a class other than k. To ensure $\mathbf{x}_i$ does not fall into another class's region, we need $\mathbf{w}_j^\top \mathbf{x}_i - \mathbf{w}_k^\top \mathbf{x}_i < \theta_j$ for every j other than k. Hence, with the non-increasing property of ϕ, the loss term $\phi(\mathbf{w}_k^\top \mathbf{x}_i - \mathbf{w}_j^\top \mathbf{x}_i + \theta_j)$ will encourage a positive value of $\mathbf{w}_k^\top \mathbf{x}_i - \mathbf{w}_j^\top \mathbf{x}_i + \theta_j$ which means we have $\mathbf{w}_j^\top \mathbf{x}_i - \mathbf{w}_k^\top \mathbf{x}_i < \theta_j$. On the other hand, for a data point $\mathbf{x}_i$ with $y_i = k$ and $y_i^b \neq k$, our hybrid model will classify $\mathbf{x}_i$ correctly only when $\mathbf{x}_i$ falls in the class k region, namely, $\mathbf{w}_k^\top \mathbf{x}_i - \mathbf{w}_j^\top \mathbf{x}_i \geq \theta_k$ for every j other than k. Hence, we use the loss term $\phi(\mathbf{w}_k^\top \mathbf{x}_i - \mathbf{w}_j^\top \mathbf{x}_i - \theta_k)$ to encourage a positive value of $\mathbf{w}_k^\top \mathbf{x}_i - \mathbf{w}_j^\top \mathbf{x}_i - \theta_k$.

3.2. Model Training

With the loss function defined in (3), the hybrid model can be trained by solving the convex minimization problem (2) for which many efficient optimization techniques are available in literature including subgradient methods (Nemirovski et al., 2009; Duchi et al., 2011), accelerated gradient methods (Nesterov, 2013; Beck & Teboulle, 2009), primal-dual methods (Nemirovski, 2004; Chambolle & Pock, 2011) and many stochastic first-order methods based on randomly sampling over coordinates or data (Johnson & Zhang, 2013; Duchi et al., 2011). The choice of algorithms for (2) depends on various characteristics of the problem, such as smoothness, strong convexity, and data size.

Since numerical optimization is not the focus of this paper, we will simply utilize the accelerated proximal gradient method (APG) by Nesterov (Nesterov, 2013) to solve (2) when ϕ is smooth. See the algorithm in the Appendix.

4. Experiments

We perform a detailed experimental evaluation of the proposed model on four public datasets. The goal here is to examine the predictive performance, transparency, and model complexity. We also analyze the medical dataset in detail to provide users a more intuitive understanding of the model. MALC is compared with some baseline models, and the human evaluation results are also presented in this section.

4.1. Experiments on Public Datasets

Datasets

We analyze four real-world datasets that are publicly available at (Chang & Lin, 2011; Ilangovan, 2017; Kaggle, 2018; Wang et al., 2017). 1) *Coupon* (Wang et al., 2017) (12079 $\times$ 113 and 3 classes) studies responses of consumers to the recommendation of coupons when users are driving in different contexts, using features such as the passenger, destination, weather, time, etc. The three classes are "decline", "accept and will use right away", and "accept and will use later" 2) *Covtype* (Chang & Lin, 2011) (581,012 $\times$ 54 and 7 classes) studies the forest cover type of wilderness areas which include Roosevelt National Forest of northern Colorado. The features in the covtype dataset are scaled to $[0, 1]$. 3) *Customer* (Kaggle, 2018) (7032 $\times$ 29 and 2 classes) was collected to study the behavior of customers and whether they return or not. 4) *Medical* (Ilangovan, 2017) (106,643 $\times$ 14 and 5 classes) provide information about Clinical, Anthropometric and Biochemical (CAB) survey done by Govt. of India. This survey was conducted in nine states of India with a high rate of maternal and infant death rates in the country. We focused on the subset of data for children under the age of five and predicted their illness type. We dropped some features not needed for classification, and the missing values in certain features were replaced by mean or mode values appropriately. For each dataset, we randomly sample 80% instances to form the training sets and use the remaining 20% as the testing sets. Since the *Medical* dataset is highly unbalanced among different classes, we downsample the majority class and upsample the minority class to make them balanced.

Training Black-box Models We first choose three state-of-the-art black-box classifiers, Random Forest (Liaw et al., 2002), XGBoost (Chen & Guestrin, 2016) and fully-connected neural network with two hidden layers. All of these models are implemented with R or python. The Random Forest model is built using the *ranger* package (Wright & Ziegler, 2015). The XGBoost model is built using the *xgboost* package (Chen et al., 2015). The neural network model is built using the *keras* package (Chollet & Allaire, 2017). For each model, we identify one or two hyperparameters, and, for each dataset, we apply an 80%-20% holdout method on the training set to select the values for these hyperparameters from a discrete set of candidates that give the best validation performance. For Random Forest, we use 500 trees and tune the minimum node size and maximal tree depth. For XGBoost, we tune maximal tree depth and the number of boosting iterations. For the neural network, we choose the sigmoid function as the activation function and tune the number of neurons and the dropout rates in the two hidden layers.

Training MALC We use the three black-box models' predictions on the training set as the input to build MALC models. In (2), we choose ϕ to be the smooth hinge loss and $r(\mathbf{w}) = \|\mathbf{w}\|_1$. We would like to obtain a list of models that span the entire spectrum of transparency, so we vary C_1 and C_2 to achieve that goal. Note that C_1 is directly related to transparency, and we use grid-search to find a suitable range to achieve transparency from zero to one. C_2 is related to the sparsity of the model. Overall, we choose C_1 from $[0.005, 0.95]$ and C_2 from $[0.03, 0.25]$. For each C_1 value, we use 80%-20% holdout on the training set to choose C_2 from a discrete set of candidates that give the best validation performance. After choosing the pairs of (C_1, C_2) values, the Algorithm APG is run up to $20,000$ iterations to make sure the change in objective value was less than 0.1%, in the last iterations, to ensure the convergence.

Efficient Frontier Analysis We characterize the trade-off between predictive accuracy and transparency using *efficient frontiers*. To create efficient frontiers, we vary the parameters C_1 and C_2 to generate a list of models producing an accuracy-transparency curve for each dataset. This provides the user with a range of models and the option to choose a desirable balance of transparency and predictive performance compared to the two discrete choices of a black-box or interpretable model. In Figure 3, each efficient frontier starts with a transparency value of zero, which corresponds to a pure black-box model. The general trend is as transparency increases, and accuracy tends to decrease. The rate of change of transparency w.r.t predictive performance is different for each dataset. For coupon and covtype datasets, accuracy decreases steadily as the transparency increases. However, The medical dataset provides an interesting scenario where the initial increase in transparency does not lead to a decrease in predictive performance. It only falls after a certain transparency threshold. Note that the transparency value of one corresponds to a pure linear (interpretable) model. But the interpretability comes at the cost of predictive performance, as evident by lower accuracy of linear models compared to the accuracy of the black-box models for all datasets. MALC provides the user with a unique framework of choosing a model from the whole spectrum of options available on an efficient frontier with their desired accuracy and transparency. We recommend the users to choose the models around the tipping point to ensure gain in transparency without a significant loss of accuracy.

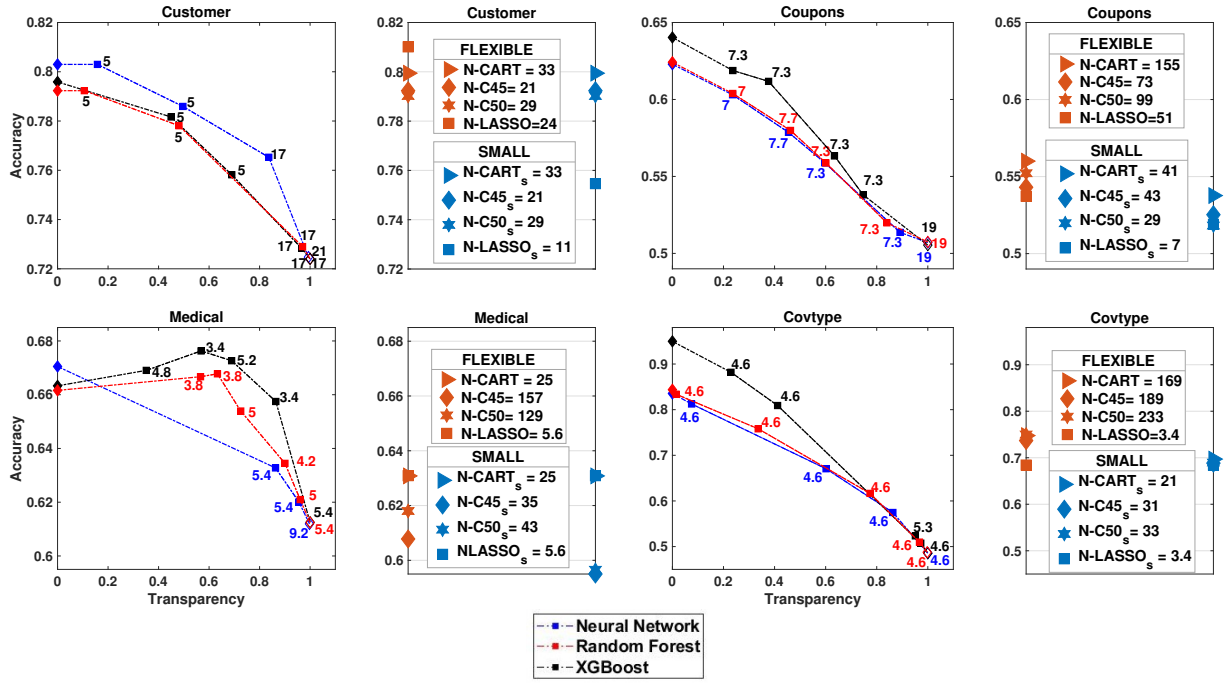


Figure 3. The efficient frontiers (EF) of MALC when collaborating with different black-box models. The numbers on EF represent the average number of features per class being used by that MALC model. The results of the pure interpretable baselines are shown on the right of each EF. N-model represents the no. of conditions for each decision tree model and no. of non-zero coefficients for the Lasso model. Flexible model results are presented on the left y-axis, and small model (number of nodes < 50) results on the right y-axis.

Number of Features Analysis We would also like to make sure the linear models are indeed interpretable, i.e., using a few non-zero terms in the model. We report in Figure 3 the average number of non-zero coefficients in MALC, which is calculated as a ratio of the number of non-zero coefficients in K linear models ($\mathbf{w}$) to the number of classes. Observe that MALC models require a relatively small number of features from the dataset to gain transparency, preserving linear models' interpretability.

The control over transparency-accuracy trade-off and the use of a small number of features to gain transparency make MALC a strong candidate for real-world applications, particularly when the user wants to avoid a pure black-box model or a pure interpretable but non-accurate model.

Comparison with Baselines MALC has a unique model form that is different from the current work in interpretable machine learning. Considering the situation, it is appropriate to compare MALC with stand-alone interpretable models (like decision trees, LASSO), to evaluate transparency by trading off accuracy. We compare with three decision trees CART (Breiman, 2017), C4.5 (Quinlan, 2014), C5.0 (Pandya & Pandya, 2015) and LASSO (Tibshirani, 1996) as stand-alone interpretable models.

For CART, we tune the complexity parameter, which imposes a penalty to the tree for having two many splits. For C4.5 and C5.0, we tune the minimum number of samples at splits. The regularization parameter was tuned for LASSO. Further details of the parameter setup for the baseline experiments are in the Appendix. We choose two models for each decision tree method, a most accurate model without any size constraint labeled as "flexible", and a most accurate model with less than 50 conditions (nodes), labeled as 'small.' See Figure 3.

Except for the customer dataset, interpretable baselines lose about 10% of accuracy compared to black-box models. This makes them less favorable by users, as shown later in our human evaluation. On the other hand, MALC provides more model choices for users while being consistently smaller than baseline models. Users can, therefore, choose their desired accuracy and transparency, based on efficient frontiers.

4.2. Case Study on the Medical Dataset

We show an example of MALC on the medical dataset. There are five classes in this dataset, "no illness", diarrhea/dysentery", "acute respiratory infection", "fever of any type", and "other illness". MALC was built in collaboration with a pre-trained random forest whose accuracy is 66.0%. After building five linear competitors, the accuracy of MALC reaches 66.4% while gaining transparency of 77.7%. The coefficients of the five linear models are shown

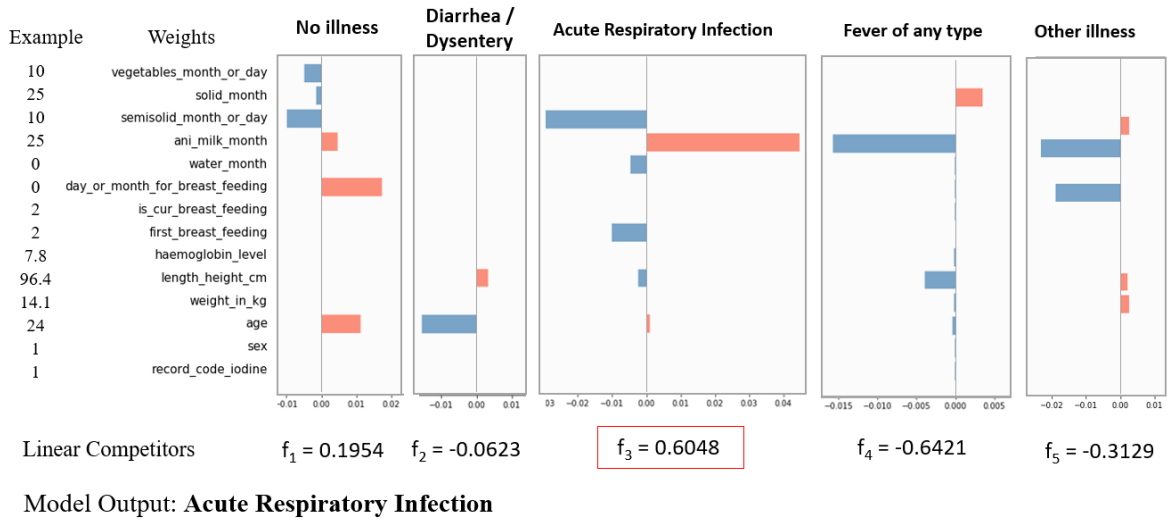


Figure 4. An example of MALC in collaboration with a pre-trained random forest. The linear model coefficients for each of the five classes are displayed.

in Figure 4. From the linear models, one can easily extract some of the key characteristics for each class. For example, the later children start to receive semisolid food, and the longer they are exclusively breastfed, the more likely they will be free of any of the illness (Class 1). Children who start receiving semisolid mashed food at a very young age, start receiving water at an early age, and are too late to start receiving animal milk/formula milk are more likely to have an acute respiratory infection (Class 3).

We chose an example instance and show the input features and the output of the linear models in Figure 4. This child started receiving animal milk/formula milk at the age of 25 months, almost six times the average age of receiving animal/formula milk (4.3 months). This child started receiving semisolid food at 10 months old, later than the average age of children (5.8 months) who start receiving semisolid food. This is helpful for the child's overall health conditions, as suggested by classifier 1. However, this effect is completely overtaken by the late usage of formula milk.

Also, the child was breastfed later than 64% of the children in the dataset. Combining these important features, classifier 3 outputs the highest score, with a large enough margin over the other four linear models. Thus this child is predicted to have an acute respiratory infection consistent with the true label.

An interesting observation for this model is that it performs slightly better than the black-box alone, which means the 77.7% transparency is obtained for free. This is the desired situation for hybrid models like MALC to be adopted.

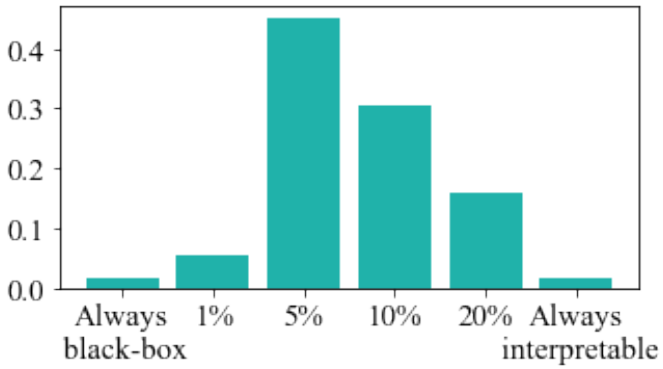
4.3. Human Evaluation

We study humans' preferences for black-box models, interpretable models, and MALC when they have different predictive accuracies and how the preferences vary for technical and non-technical users. See the supplementary material for the survey questions.

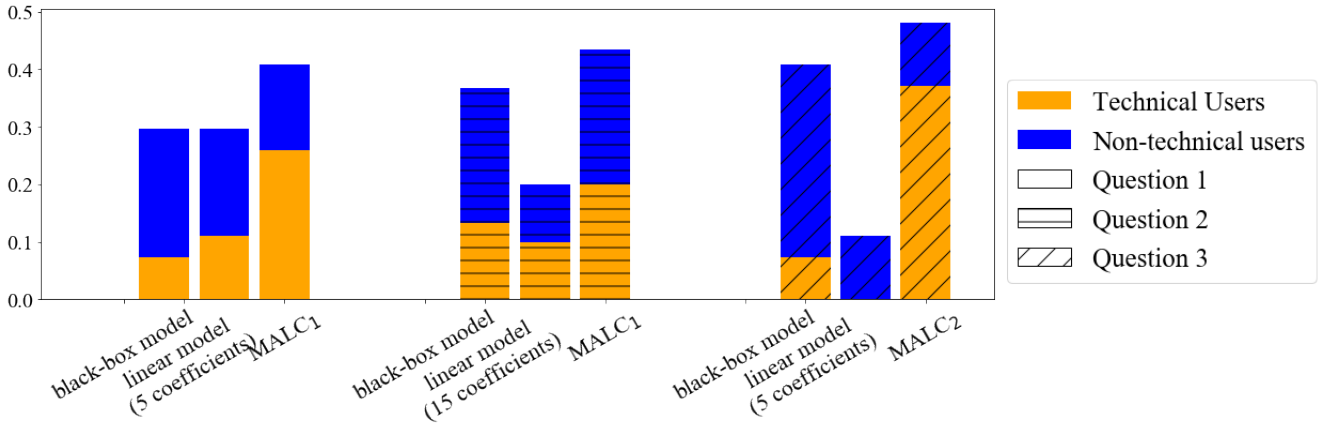
For this purpose, we designed a survey and collected responses from 71 subjects in total. The subjects were mainly recruited from two channels, graduate students from authors' home university and Amazon Mechanical Turk. Thus our study covers both technical and non-technical users: 25 subjects have at least a Master's degree, and the rest have lower or no degrees. The average age of the participants was 33.8, from the youngest 23 to the oldest 61. 77% of the subjects were male.

In this survey, we presented to subjects the context of customer retention prediction. We first teach subjects how to understand a linear model and MALC, respectively, to give them some idea what interpretability means, in contrast to a black-box model. To make sure users understand the model and screen out those who were not paying attention, we then provide users with a test question, asking them to use a simple linear model to classify an instance. Those who failed the question were excluded from the following reports.

For users who passed the testing question (61 out of 71), we ask them for the largest accuracy loss they can tolerate if required to use an interpretable model instead of a black-box model. Results are shown in Figure 5(a). Notice that there is a small group of people who would never choose interpretable models if there is an accuracy loss, and also a small group of people who will always choose interpretable



(a) The largest tolerable accuracy loss by different subjects.



(b) Distribution of human choices of different model types. MALC_2 has the same transparency as MALC_1 but higher accuracy.

Figure 5. Human Evaluation Results

models, no matter how much accuracy is lost. The majority of people can accept a certain level of accuracy loss, with the largest group willing to sacrifice at most 5% accuracy. The result also validates why in the previous experiments in Figure 3, interpretable baselines are insufficient since they lose more than 5% accuracy than black-box models.

We use δ to represent a subject's tolerable loss of accuracy. Then we ask users to choose from the following models: 1) a highly accurate black-box model, 2) we randomly show one from two linear models, one with 15 non-zero coefficients and one with five non-zero coefficients, both with an accuracy loss of δ compared to the black-box model and 3) a hybrid model MALC_1 with five non-zero coefficients, accuracy loss 0.5δ and transparency 50%. Subjects' choices are shown in Question 1 and Question 2 in Figure 5(b). We observe when the linear model becomes less interpretable (the number of non-zero coefficients increased from 5 to 15), some users abandoned linear models and switch to the black-box or the hybrid model. Note that the majority of the switch happened to non-technical users, as they are more sensitive to the increase in the cognitive load in understanding a model. Next in question 3, we replace MALC_1 with an improved hybrid model MALC_2, with 0.25δ accuracy loss and the same transparency as MALC_1 and ask users to choose again. As MALC becomes more accurate, more users choose MALC, mainly technical users: those who chose the linear model in the previous question all switched to MALC_2.

Results show that almost half of the users are willing to trade-off accuracy for interpretability, and the preference becomes stronger if MALC has a better predictive performance. , even when there is no extra gain in the trade-off: when MALC_1 lies on a straight line connecting the black-box model and an interpretable model. Interestingly, we notice that "accuracy oriented" users remain loyal to their choice of the black-box models even if we provide better MALC and "interpretability oriented" users also remain faithful to their choice of the linear model. Our survey demonstrates the diversity in users' choice of models.

5. Conclusion

We proposed a Model-Agnostic Linear Competitors Model for multi-class classification. MALC promotes transparency by building K linear models to collaborate with a pre-trained black-box model. We formulated the training of a MALC model as convex optimization, where predictive accuracy and transparency balance through objective function. The optimization problem is solved with the accelerated proximal gradient method.

MALC provides more flexible model choices for users. Experiments show that MALC was able to yield models with different transparency and accuracy values by varying the parameters, thus providing more model options to users. In real applications, users can decide the operating point based on the efficient frontier. The decision depends on the application use-case and desired balance of transparency and accuracy, which varies by different users, as shown in the human evaluation.

Discussion of Applications MALC can be used in situations where the requirement for predictive performance is stringent, while interpretability is highly appreciated. MALC helps "gain" some interpretability at the trade-off of a possible user-tolerable loss of accuracy. MALC is model-agnostic, making it flexible to collaborate with state-of-the-art black-box models to utilize their high predictive power while preserving some proprietary information of the black-box model.

In addition to collaborating with black-box models, MALC can collaborate with human decision-makers. Consider a domain expert as f_b, MALC can be trained needing only his decisions on previous data. Such situations apply to, for example, medical diagnosis, like the one in the case study. Our linear model can replace a human doctor in some easy cases, saving medical resources for the hospital and costs for patients.

The proposed work offers a new perspective in building handshakes between interpretable and black-box models, to build collaboration between them to exploit the strength of both. Hybrid models like MALC can serve as a comfortable stepping stone for users accustomed to black-box models to move towards more interpretable machine learning.

References

Aïvodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., and Tapp, A. Fairwashing: the risk of rationalization. *International Conference on Machine Learning*, 2019.

Alvarez-Melis, D. and Jaakkola, T. S. On the robustness of interpretability methods. *ICML Workshop on HumanInterpretability in Machine Learning*, 2018.

Aubert, R. E., Yao, J., Xia, F., and Garavaglia, S. B. Is there a relationship between early statin compliance and a reduction in healthcare utilization? *The American journal of managed care*, 16(6):459–466, 2010.

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Breiman, L. *Classification and regression trees*. Routledge, 2017.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. 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*, pp. 1721–1730. ACM, 2015.

Chambolle, A. and Pock, T. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.

Chang, C.-C. and Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chen, C., Lin, K., Rudin, C., Shaposhnik, Y., Wang, S., and Wang, T. An interpretable model with globally consistent explanations for credit risk. *arXiv preprint arXiv:1811.12615*, 2018.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794. ACM, 2016.

Chen, T., Benesty, M., Khotilovich, V., and Tang, Y. Xgboost: extreme gradient boosting. In *R package version 0.4*, 2015.

Chollet, F. and Allaire, J. R interface to keras., 2017. Retrieved from https://github.com/rstudio/keras.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

Garcez, A. d., Besold, T. R., De Raedt, L., Földiak, P., Hitzler, P., Icard, T., Kühnberger, K.-U., Lamb, L. C., Miikkulainen, R., and Silver, D. L. Neural-symbolic learning and reasoning: contributions and challenges. In

Proceedings of the AAAI Spring Symposium on Knowledge Representation and Reasoning: Integrating Symbolic and Neural Approaches, Stanford, 2015.

Garcez, A. S. d., Broda, K. B., and Gabbay, D. M. *Neural-symbolic learning systems: foundations and applications*. Springer Science & Business Media, 2012.

Hu, Z., Ma, X., Liu, Z., Hovy, E., and Xing, E. Harnessing deep neural networks with logic rules. *arXiv preprint arXiv:1603.06318*, 2016.

Hua, Z. and Zhang, B. A hybrid support vector machines and logistic regression approach for forecasting intermittent demand of spare parts. *Applied Mathematics and Computation*, 181(2):1035–1048, 2006.

Ilangovan, R. Clinical, anthropometric & bio-chemical survey, 2017. Retrieved from https://www.kaggle.com/rajanand/cab-survey.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pp. 315–323, 2013.

Kaggle. Telco customer churn, 2018. Retrieved from https://www.kaggle.com/blastchar/telco-customer-churn.

Kohavi, R. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *KDD*, volume 96, pp. 202–207, 1996.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Letham, B., Rudin, C., McCormick, T. H., Madigan, D., et al. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3):1350–1371, 2015.

Liaw, A., Wiener, M., et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

Lipton, Z. C. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.

Lissack, M. Dealing with ambiguity–the 'black box'as a design choice. *SheJi (forthcoming)*, 2016.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pp. 4765–4774, 2017.

McGarry, K., Wermter, S., and MacIntyre, J. Hybrid neural systems: from simple coupling to fully integrated neural networks. *Neural Computing Surveys*, 2(1):62–93, 1999.

Nemirovski, A. Prox-method with rate of convergence $o(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Pandya, R. and Pandya, J. C5.0 algorithm to improved decision tree with feature selection and reduced error pruning. *International Journal of Computer Applications*, 16(3):18–21, 2015.

Quinlan, J. R. *C4. 5: programs for machine learning*. Elsevier, 2014.

Ribeiro, M. T., Singh, S., and Guestrin, C. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144. ACM, 2016.

Ribeiro, M. T., Singh, S., and Guestrin, C. Anchors: High-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Richter, M. M. and Weber, R. O. *Case-based reasoning*. Springer, 2016.

Ross, A. S., Hughes, M. C., and Doshi-Velez, F. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*, 2017.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretablemodels instead. *Nature Machine Intelligence*, 180:206–215, 2019.

Shin, C.-K., Yun, U. T., Kim, H. K., and Park, S. C. A hybrid approach of neural network and memory-based learning to data mining. *IEEE Transactions on Neural Networks*, 11(3):637–646, 2000.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. Interpretable decision sets: A joint framework for description and prediction. In *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*. ACM, 2020.

Taha, I. A. and Ghosh, J. Symbolic interpretation of artificial neural networks. *IEEE Transactions on knowledge and data engineering*, 11(3):448–463, 1999.

Thibault, L., Marie-Jeanne, L., Christophe, M., Xavier, R., and Detyniecki, M. The dangers of post-hoc interpretability:unjustified counterfactual explanations. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological),*, 58(1):267–288, 1996.

Towell, G. G. and Shavlik, J. W. Knowledge-based artificial neural networks. *Artificial intelligence*, 70(1-2):119–165, 1994.

Wang, J., Fujimaki, R., and Motohashi, Y. Trading interpretability for accuracy: Oblique treed sparse additive models. In *SIGKDD*, pp. 1245–1254. ACM, 2015.

Wang, T. Gaining no or low-cost transparency with interpretable partial substitute. *International Conference on Machine Learning*, 2019.

Wang, T., Rudin, C., Doshi, F., Liu, Y., Klampfl, E., and MacNeille, P. A bayesian framework for learning rule set for interpretable classification. *Journal of Machine Learning Research*, 2017.

Wright, M. N. and Ziegler, A. ranger: A fast implementation of random forests for high dimensional data in c++ and r. In *arXiv preprint arXiv:1508.04409*, 2015.

Yang, C., Shi, X., Jie, L., and Han, J. I know you'll be back: Interpretable new user clustering and churn prediction on a mobile social application. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 914–922. ACM, 2018.

Zeng, J., Ustun, B., and Rudin, C. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(3):689–722, 2017.