FROTE: FEEDBACK RULE-DRIVEN OVERSAMPLING FOR EDITING MODELS

Öznur Alkan [1] **Dennis Wei** [1] **Massimiliano Matteti** [2] **Rahul Nair** [1] **Elizabeth M. Daly** [1] **Diptikalyan Saha** [1]

ABSTRACT

Machine learning (ML) models may involve decision boundaries that change over time due to updates to rules and regulations, such as in loan approvals or claims management. However, in such scenarios, it may take time for sufficient training data to accumulate in order to retrain the model to reflect the new decision boundaries. While work has been done to reinforce existing decision boundaries, very little has been done to cover these scenarios where decision boundaries of the ML models should change in order to reflect new rules. In this paper, we focus on user-provided feedback *rules* as a way to expedite the ML models' update process, and we formally introduce the problem of pre-processing training data to edit an ML model in response to feedback rules such that once the model is retrained on the pre-processed data, its decision boundaries align more closely with the rules. To solve this problem, we propose a novel data augmentation method, the *F*eedback *R*ule-Based *O*versampling *Te*chnique (FROTE). Extensive experiments using different ML models and real world datasets demonstrate the effectiveness of the method, in particular the benefit of augmentation and the ability to handle many feedback rules.

1 INTRODUCTION

Machine learning (ML) classifiers are increasingly employed in critical decision-making processes such as loan approvals, credit score assignment (Khandani et al., 2010), and claims management (Singh & Urolagin, 2020). Much focus in the research community has been on improving accuracy of such ML models, evaluated on test data with a similar distribution as the training data. However, to deploy such ML models in the real world, one must address problems that arise from the model being inherently governed and limited by the training data. In many applications, domain expert knowledge could be used to improve performance either where data coverage is sparse, or where decision boundaries may have changed over time. Loan approval policies are an example where training data may reflect historical policies but not new policies with shifted decision boundaries.

Naive options for incorporating expert feedback include manually relabelling historical data and labelling new data. Both are costly in terms of human intervention, and doing the latter alone compromises the accuracy of the deployed model until enough new data is collected. While active learning can reduce the amount of new data needed, the burden may still be too high (Cakmak et al., 2010; Guillory & Bilmes, 2011), and moreover during deployment, it may not be possible to select which instances to label. Recent work (Daly et al., 2021) has proposed a more efficient feedback mechanism using rules. This approach uses algorithms for learning decision rules (Lakkaraju et al., 2016; Ribeiro et al., 2018; Dash et al., 2018) to provide explanations for arbitrary ML classifiers. The expert's task is then limited to reviewing and modifying a set of classifier predictions and rule-based explanations, resulting in a *feedback rule set* (FRS). Daly et al. (2021) propose a post-processing layer to account for the feedback rules; however, the feedback is not incorporated into the underlying model.

In this paper, we propose an algorithm called FROTE (*F*eedback *R*ule-Based *O*versampling *Te*chnique) to edit an ML model for tabular data in response to user feedback rules. FROTE thus complements the input transformation method of (Daly et al., 2021). Given an input dataset, the algorithm first modifies the training data if allowed, and then augments it so that re-training the model on the augmented data results in better alignment with the feedback rules. FROTE can thus be used with any classification algorithm that takes training data as input and produces a classifier as output; the algorithm (which could be proprietary) is treated as a black box. Unlike Daly et al. (2021), the user feedback is directly encoded in the model.

[1]IBM Research [2]Microsoft. Correspondence to: Öznur Alkan <oznurkirmemis@gmail.com>, Dennis Wei <dwei@us.ibm.com>, Massimiliano Matteti <mmattetti@microsoft.com>, Rahul Nair <rahul.nair@ie.ibm.com>, Elizabeth M. Daly <elizabeth.daly@ie.ibm.com>, Diptikalyan Saha <diptsaha@in.ibm.com>.

Proceedings of the 5^{th} MLSys Conference, Santa Clara, CA, USA, 2022.

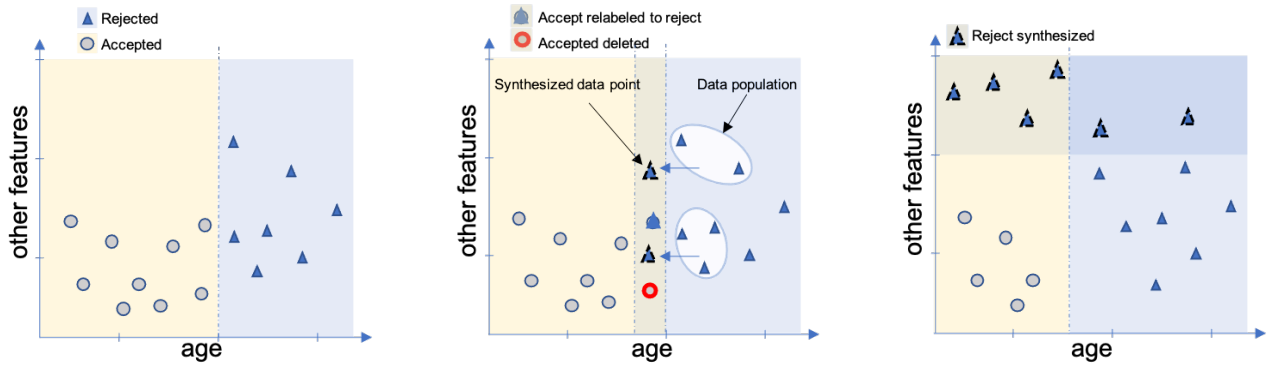

Figure 1. Left: Original classification boundary. *Middle:* FROTE generates synthetic instances to move decision boundary (after relabelling and removing if permitted). *Right:* Generating synthetic instances where existing data is limited.

We use Figure 1 to be suggestive of a loan approval scenario and to illustrate our solution. Suppose there is a new policy to lower the ages of applicants for whom loans are approved. Rather than crafting rules from scratch, the user relies on the existing ML model and accompanying rule-based explanations to capture relevant dependencies among a potentially large number of features, and only modifies rules that involve age. Given the resulting feedback rule set, the user may wish to relabel and remove existing instances as shown in Figure 1(b). FROTE then generates synthetic instances that reflect both the feedback rules as well as the existing data. Synthetic data generation can address the challenge of insufficient training data in the region to be adjusted, as seen in Figure 1(c). For data generation, we build upon the SMOTE method (Chawla et al., 2002) in several ways; other methods could also be adapted.

Our contributions can be summarized as follows: 1) We formulate the problem of editing an ML model by pre-processing a dataset based on user feedback rules. 2) A novel data augmentation-based solution, FROTE, is presented. 3) FROTE is extensively evaluated using different ML models, real-world datasets, and feedback rule set parameters to demonstrate its effectiveness, in particular the benefit of augmentation, improved performance over the state-of-the-art, and the ability to handle many feedback rules.

2 Related Work

To the best of our knowledge, the problem studied in this paper is novel in that it differs in at least one of the following aspects from the existing literature: 1) general editing of ML models 2) based on user-specified feedback rules 3) via model-agnostic data augmentation/pre-processing, where the rules can enforce existing boundaries, or introduce new boundaries through changing the dataset.

Data augmentation/pre-processing has been explored in different problem settings. The *class imbalance* problem, which deals with the unequal distribution of classes in training data, was tackled in the seminal work of Chawla et al. (2002). Their Synthetic Minority Oversampling Technique (SMOTE) randomly selects minority data points as base instances and generates new data points that are convex combinations of the base instances and their k nearest neighbours. Han et al. (2005) extend SMOTE by synthesizing data points that reinforce existing decision boundaries. Due to its simplicity in the design of the procedure, as well as its robustness, SMOTE has been applied to different type of problems and has proven successful in a variety of applications from several different domains (Fernández et al., 2018). While we build on SMOTE for data generation, our model editing use case differs in going beyond reinforcing existing boundaries to adjusting and introducing new ones. While our contributions build upon these prior works in terms of generating synthetic data instances, our use case is not only to reinforce existing decision boundaries, but also to enable a user both to adjust those decision boundaries and introduce new ones.

More recently, a more specific use case has gained attention, where data is processed in order to understand and mitigate underlying biases through focusing on fairness. Within the *fairness* and *bias mitigation* literature, pre-processing methods such as relabelling and reweighing (Calders et al., 2009), data synthesis (Sharma et al., 2020), and data transformation (Calmon et al., 2017) have been proposed.

We argue that the problem we tackle is a more general form of user feedback that can support user concerns through feedback rules, rather than the ones based on only the specified protected features.

Within the *transfer learning* literature, Dai et al. (2007); Eaton & desJardins (2011) address a similar problem where

test data does not follow the same distribution as training data. They propose an iterative mechanism that re-weights the old data to minimize error observed on the new data. In (Eaton & desJardins, 2011), desJardins and Eaton pursue a similar strategy. Neither approach however generates synthetic instances.

In the *generative models* domain, synthetic data generation is used for several tasks. For example, generative adversarial networks (GANs) aim to improve the realism of generated samples until the adversary cannot distinguish real from synthetic data (Goodfellow et al., 2014). In Tanaka & Aranha (2019); Douzas & Bacao (2018); Xu et al. (2019), GANs and conditional GANs with different network architectures are used to generate synthetic data to overcome class imbalance as well as privacy issues. In (Douzas & Bacao, 2018), a conditional version of GAN (cGAN) is used to generate data for the minority class of various imbalanced datasets. Overall, when comparing the performance of the classifiers on imbalanced data sets that were augmented by the GAN and SMOTE, the former provides better results but with the cost of an higher complexity correlated to the training of the networks. Xu et al. (Xu et al., 2019) generate tabular synthetic data using conditional tabular GANs. Again these do not support model editing based on rules.

Incorporating prior knowledge into support vector machines (SVM) was reviewed by Lauer & Bloch (2008). Two forms of prior knowledge were considered: 1. invariances to transformations, to permutations and in domains of input space, 2. knowledge on the unlabelled data, the imbalance of the training set or the quality of the data. Maclin et al. (2006) make use of knowledge bases of rules and virtual support vectors to add constraints to the optimization. Different from our solution, these works target only SVM models. Another work from Kapoor et al. Kapoor et al. (2010) support user influence over ML algorithms by manipulating confusion matrices, where the user is allowed to manipulate the initial confusion matrix over the different classes.

Leveraging expert rules has been explored in the *assisted labelling* literature. Snorkel (Ratner et al., 2017) takes a weak supervision approach to labelling training data by bringing together label predictions from different sources, including labelling functions that can be expert-provided patterns. The labelling sources include labelling functions which can be expert provided patterns and heuristics to predict labels. A generative model is built to estimate the accuracy and correlations of the different labelling sources and produces probabilistic training data where each data point has a probabilities distribution over all the labels and then can be used to train a model. Awasthi et al. (2020) consider hybrid supervision from labelled instances as well as rules that generalize them. The assisted labelling problem is different from ours in that they seek to label unlabelled data whereas we already have a model trained on a labelled dataset and wish to edit the model, without negatively impacting accuracy for data unaffected by the rules. In addition, in assisted labelling, experts have to devise rules from scratch whereas in model editing, they may only have to modify rules that capture what the model has already learned. They provide a solution where labels are unavailable or noisy and seek to label unlabelled data. Our goal is somewhat different, where we assume the presence of a dataset and a model that may be considered trusted and validated but the user wants to make some adjustments or edits without negatively impacting the model accuracy for unaffected data which should remain unchanged.

The most closely related work by Daly et al. (2021) addresses user feedback rules, but not by editing the ML model. Instead, transformations that map between the original and feedback rules are obtained to yield a post-processing layer called Overlay. When a new data point arrives for prediction, Overlay checks to see if a feedback rule corresponds to the data point and if so, applies the transformation, returning the prediction of the transformed data point. While Overlay enables immediate changes to an ML system by applying the above transformations to the input, without retraining the model, Daly et al. (2021) note that it is a "patch". As more feedback rules and their corresponding patches are produced, the overall system consisting of the ML model and these patches may become overly complex and difficult to maintain. It is not difficult to imagine that even a single expert could generate a large number of feedback rules. Additionally, experiments by Daly et al. (2021) suggest and our experiments confirm (Table 2) that one limitation of Overlay occurs when a feedback rule differs too significantly from the underlying model, a limitation that FROTE overcomes. Moreover, in applications such as finance or spam detection, Overlay's transformations may incur additional undesirable latency. For the reasons above, once short-term patches have been applied, it may be preferable to directly incorporate user feedback into the model, which is the problem that FROTE solves.

3 PRELIMINARIES

As discussed in the Introduction, the premise of this work is that 1) the distribution of future data (i.e. test data) is different from that of training data, due for example to a policy change or to the training data not being representative, and 2) a domain expert understands the nature of the change and communicates that through a set $\mathcal{F}$ of *feedback rules*, i.e. a *feedback rule set* (FRS). To establish notation, let $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$ denote a set of attributes for decision-making, and $y \in \mathcal{Y} = \{c_1, c_2, \ldots, c_l\}$ denote a class label. The existing training data is a set of n instances $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, assumed to be drawn i.i.d. from a joint distribution $p_{X,Y}$.

3.1 Feedback Rules

We consider a generalization of decision rules beyond recent works (Lakkaraju et al., 2016; Molnar, 2019) to allow feedback rules that are *probabilistic*. A feedback rule $R = (s, \pi)$ is thus a statement of the form IF *the clause s is true* THEN *the class label Y is distributed according to π*. These are discussed in turn below.

Clauses and coverage. A clause is a conjunction of one or more predicates (also referred to as conditions) of the form (attribute, operator, value). In our solution, the operators allowed for categorical attributes are $\{=, \neq\}$, and for numeric attributes are $\{=, >, \geq, <, \leq\}$. An example of a clause with three predicates is *age $<$ 29 AND marital-status = 'single' AND income $>$ 150K*. We say that $\mathbf{x} \in \mathcal{X}$ *satisfies* a clause s, and reciprocally, a rule (s, π) *covers* $\mathbf{x}$, if all the predicates in s are *true* when evaluated on $\mathbf{x}$. Given a dataset D, *coverage* of a rule (s, π) and an FRS $\mathcal{F} = \{(s_r, \pi_r)\}_{r=1}^{m}$ of m feedback rules are defined as follows:

$$\mathrm{cov}(s, D) = \{(\mathbf{x}, y) \in D : \mathbf{x} \text{ satisfies } s\}, \tag{1}$$

$$\mathrm{cov}(\mathcal{F}, D) = \bigcup_{r=1}^{m} \mathrm{cov}(s_r, D). \tag{2}$$

Note that coverage involves only clauses s and attributes $\mathbf{x}$. If D is omitted as in $\mathrm{cov}(s)$, then it is understood to be the entire domain $\mathcal{X}$.

The reason for using logical clauses as above is that they semantically resemble natural language and the way humans think (Zhang & Deng, 2015; Letham et al., 2015; Molnar, 2019). Therefore it can be more natural for users to provide feedback in the form of a rule, either of their own creation or by modifying an algorithm-provided rule-based explanation. This does require the rule's conditions to be built from intelligible features and favours smaller numbers of conditions and rules (Lakkaraju et al., 2016).

Label distribution. Given a feedback rule (s, π) and $\mathbf{x} \in \mathrm{cov}(s)$, we assume that the class label is distributed as $Y \sim \pi$. We will mostly work with the *deterministic* case where π is the Kronecker delta distribution for a class c, i.e., $Y = c$ with probability 1. This is the easiest case for a human expert, who only has to specify the class c. However, allowing probabilistic rules is useful for at least two reasons: 1) accommodating conflicts between rules (discussed next), and 2) allowing uncertainty in rules and providing robustness against over-confident rules.

Rule conflicts. When feedback from multiple experts is to be considered, the possibility of conflicts should be taken into account due to contradictory opinions. Two rules (s_1, π_1), (s_2, π_2) are conflicting if their coverages intersect, $\mathrm{cov}(s_1) \cap \mathrm{cov}(s_2) \neq \emptyset$, and $\pi_1 \neq \pi_2$. We assume that all such conflicts are resolved, for example through one of the following options:

1. Removal of the intersection, i.e., clause s_1 is changed to s_1 *AND NOT* s_2, and s_2 to s_2 *AND NOT* s_1.

2. Creation of a new rule for the intersection with a mixture of the distributions, e.g. $(\pi_1 + \pi_2)/2$ or a more general weighting. The intersection is then excluded from the two original rules as in option 1.

3. If the two rules are provided by different experts, asking them to come to a consensus.

We assume that the final FRS is *conflict-free* through repeated application of the above operations for conflict resolution.

When we consider the above alternatives, expert consensus may be preferred as it is the most informed strategy. In the absence of expert consensus, the mixture distribution of the second option represents a mathematical consensus. Second option could also be preferred when users are hesitant to delete real data instances. On the other hand, first option could be preferred if users are conservative and would rather have no rule when rules conflict.

In addition to the above alternatives, using probabilistic generative models can also be used in principle to handle rule conflicts (Ratner et al., 2017). For example, if there are multiple experts, rules from experts could be treated as labelling functions, and inferred probabilistic labels can be treated as assigned probabilities to rules in our setting. This can be an alternative for conflict resolution and can aid in assigning probabilities automatically instead of requiring experts to provide them.

3.2 Problem Formalization

We are given 1) a conflict-free feedback rule set $\mathcal{F}$, 2) an initial training dataset D, and 3) a classification *algorithm* A that, given a dataset D, trains a classification model M_D. The task is to create a dataset $\hat{D}$ by augmenting D such that when the model is retrained on $\hat{D}$ using A to yield $M_{\hat{D}}$, the objective function in (3) is minimized. To define the objective function, let $L_1, L_2 : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$ be two loss functions that compare two labels. We also assume for ease of exposition that the rule coverage sets are disjoint, which can be achieved by 1) resolving conflicts as described above, and 2) merging rules that overlap but do not conflict. Then

the objective function can be written as;

$$J\big(M_{\hat{D}}, \mathcal{F}\big) = \sum_{(s_r,\pi_r)\in\mathcal{F}} \Pr(X \in \mathrm{cov}(s_r)) \times \mathbb{E}_{X\sim p_X, Y\sim\pi_r}\left[L_1\big(M_{\hat{D}}(X), Y\big) \mid X \in \mathrm{cov}(s_r)\right] + \Pr(X \notin \mathrm{cov}(\mathcal{F})) \times \mathbb{E}_{X,Y\sim p_{X,Y}}\left[L_2\big(M_{\hat{D}}(X), Y\big) \mid X \notin \mathrm{cov}(\mathcal{F})\right]. \quad (3)$$

The summation in (3) applies to instances in the coverage of the FRS and evaluates the retrained model's predictions $M_{\hat{D}}(X)$ against labels Y distributed according to each feedback rule's π_r. We refer to the complement of this term (i.e. 1 minus it) as *model-rule agreement* (MRA). The motivation for the name MRA comes from the case where L_1 is the 0-1 loss. Then the expectation of $1 - L_1(M_{\hat{D}}(X), Y)$ is the probability of agreement between $M_{\hat{D}}(X)$ and Y.

The last term in (3) applies to instances outside $\mathrm{cov}(\mathcal{F})$ and evaluates the predictions against labels following the original distribution $p_{X,Y}$. We refer to this term as outside-coverage performance.

4 Proposed Solution

Given an input dataset D, the goal of our proposed solution FROTE is to produce an augmented dataset $\hat{D}$ so that retraining the model on $\hat{D}$ minimizes the loss function defined in equation (3). The initial dataset D could be the one used to train the original model, or it could be a modified version of this dataset. We show in the Experiments section and supplement that FROTE works with different types of initial datasets. The steps of FROTE are given in Algorithm 1.

Base instance selection. The adaptation of SMOTE used by FROTE requires a set of *base instances* chosen from the original dataset. These provide the basis for augmentation to ensure that generated instances are similar to original instances. Base instance selection occurs in two steps: preselection of a *base population* (BP), denoted $\mathcal{P}$, before the main augmentation loop (line 4), and selection of subsets of the BP, denoted $\mathcal{B}$, within the loop (line 7). These are described in the Base Instance Selection subsection.

Augmentation loop. In each iteration of FROTE, base instances are selected from the BP (line 7) and corresponding synthetic instances are generated (line 8) as described in the Synthetic Instance Generation subsection. A temporary dataset D' is created (line 9) by combining these synthetic instances with $\hat{D}$, the current active dataset. The model is retrained on D' (line 10) and the loss function J is evaluated (line 11). If the loss decreases (lines 12-15), D' becomes the current active dataset $\hat{D}$. Otherwise, the generated instances are discarded and $\hat{D}$ is unchanged. This augmentation loop proceeds until one of the termination criteria is met: 1. the *oversampling quota* (controlled by oversampling fraction q) is used up, or 2. the *iteration limit* τ is exceeded.

User Constraints. We regard τ and q as constraints determined by user preferences: τ is the number of times the user is willing to run training algorithm A, and q is the allowed amount of augmentation relative to the initial dataset. Given τ and q, the number of generated instances per iteration is set to $q|D|/\tau$ (line 1) to uniformly distribute the quota.

Algorithm 1: FROTE

Input: input dataset D, ML algorithm A, feedback rule set $\mathcal{F}$
User Constraints: iteration limit τ, oversampling fraction q
Output: output dataset $\hat{D}$

1 $\eta \leftarrow q|D|/\tau$, $\hat{D} \leftarrow D$
2 $M_{\hat{D}} \leftarrow$ apply training algorithm A to $\hat{D}$
3 $\hat{j} \leftarrow \hat{J}_{\hat{D}}(M_{\hat{D}}, \mathcal{F})$
4 $\mathcal{P} \leftarrow \mathrm{PreSelectBP}(\hat{D}, \mathcal{F})$
5 $i, N \leftarrow 0$
6 **while** $i < \tau$ *and* $N \leq q \times |D|$ **do**
7 $\quad \mathcal{B} \leftarrow \mathrm{SelectBaseInstances}(\mathcal{P}, \eta)$
8 $\quad \mathcal{S} \leftarrow \mathrm{Generate}(\mathcal{B})$
9 $\quad D' \leftarrow \hat{D} \cup \mathcal{S}$
10 $\quad M_{D'} \leftarrow$ apply training algorithm A to D'
11 $\quad j' \leftarrow \hat{J}_{\hat{D}}(M_{D'}, \mathcal{F})$
12 $\quad$ **if** $j' < \hat{j}$ **then**
13 $\quad\quad \hat{D} \leftarrow D'$, $N \leftarrow N + |S|$
14 $\quad\quad \hat{j} \leftarrow j'$
15 $\quad\quad \mathcal{P} \leftarrow \mathrm{PreSelectBP}(\hat{D}, \mathcal{F})$,
16 $\quad$ **end**
17 $\quad i \leftarrow i + 1$
18 **end**

4.1 Base Instance Selection

Whereas SMOTE randomly selects data points from the minority class as the base population, our problem is more challenging as it is driven by the loss J in (3) and the ideal selection of base instances would maximally decrease this loss. Referring to Algorithm 1, we denote by $\mathcal{B}$ the set of selected base instances, $\mathcal{S} = \mathrm{Generate}(\mathcal{B})$ the synthetic instances generated from $\mathcal{B}$, and $A(D')$ the model obtained from the temporary dataset $D' = \hat{D} \cup \mathcal{S}$. Then the goal is to choose $\mathcal{B}$ to minimize

$$J(M_{D'}, \mathcal{F}) = J\big(A\big(\hat{D} \cup \mathrm{Generate}(\mathcal{B})\big), \mathcal{F}\big). \quad (4)$$

There are multiple challenges in minimizing (4): 1) Choosing $\mathcal{B}$ is a combinatorial subset selection problem. The size of the subset $|\mathcal{B}| = \eta$ may be large (e.g. 100), and the size of the BP $\mathcal{P}$ is larger still. 2) The training algorithm A is a black box. Furthermore, it may be expensive to run to evaluate (4). 3) The expectations in J must be approximated with empirical averages. We address this by using the current active dataset $\hat{D}$, replacing J with the empirical approximation $\hat{J}_{\hat{D}}$ over $\hat{D}$ (lines 3, 11). As a consequence however, even eval-

uating (4) for all singleton $\mathcal{B}$, e.g. all instances in $\mathcal{P}$, would incur complexity of at least $O(|\mathcal{P}||\hat{D}|)$. This implies that even a greedy selection algorithm, which would evaluate $O(\eta|\mathcal{P}|)$ subsets, would have cubic complexity $O(q|D|^3/\tau)$ assuming $\eta = q|D|/\tau$ and $|\mathcal{P}| \propto |D|$.

Herein we take a simple approach to base instance selection, consisting of 1) pre-selecting a BP to focus only on the coverage set $\text{cov}(\mathcal{F}, D)$, 2) selecting subsets *randomly*, and 3) exploring more informed strategies that maintain low computational complexity.

Base population pre-selection (line 4). Motivated by the MRA term in equation (3), we restrict the BP to the coverage $\text{cov}(\mathcal{F}, D)$. In our implementation, we maintain *per-rule* BPs, i.e., $\mathcal{P}[r]$ for $R_r \in \mathcal{F}$, and accordingly initialize $\mathcal{P}[r] = \text{cov}(s_r, D)$. However, rules may have little or no coverage in the original dataset D, and the method described in the Synthetic Instance Generation subsection requires coverage of at least $k+1$. To handle this scenario, FROTE uses *rule relaxation* to obtain a *maximal* partial rule set, denoted as $\tilde{\mathcal{F}}$. During augmentation, an instance is selected to be part of the base population if it is *strongly covered*, i.e. the instance matches a rule within $\mathcal{F}$ exactly, or if it is *weakly covered*, i.e. the instance only matches a rule partially. The latter case is designed to handle a relaxed case when a rule in $\mathcal{F}$ has zero support. In this case, we determine the maximal partial rule, a version of the rule with the minimal condition deletion that gives the maximum support. In other words, we tried to find out the minimum change we can make to the rule to give us the largest non-zero support. Since the number of conditions within each rule set is low, such a maximal partial rule can be determined by a breath-first search exhaustively by first removing one condition and then two and so on.

Base population pre-selection. Base population pre-selection procedure *PreSelectBP* is outlined in Algorithm 2. For each rule in the feedback rule set $\mathcal{F}$, FROTE requires coverage of at least $k+1$ to generate synthetic instances, where k represents the number of nearest neighbors. Therefore, conditions of a feedback rule R are relaxed if the coverage of R is less than $k+1$ (lines 7-18). During rule relaxation, the goal is to remove minimum number of conditions from R^s that will result in a maximum rule coverage. To achieve this, *PreSelectBP* performs a breadth first search on a tree of $|R^s|$ levels, where at each level the nodes are the remaining conditions in R^s. At each level, *PreSelectBP* chooses a condition whose removal results in maximum coverage in comparison with other conditions that exist at that level (lines 8-18). The procedure returns the union of the instances within the coverage of the relaxed feedback rules.

Random subset selection (line 7). The simplest choice for selecting base instances is to randomly select η instances from the BP, motivated in part by Chawla et al. (2002). We refer to this strategy as *random* in the paper. More specifically, base instances are selected on a per-rule basis as detailed in the supplement. Despite its simplicity, we find during the experiments that *random* appears to work well empirically.

Subset selection via integer programming (line 7). We also consider an integer programming (IP) approach, referred to as *IP*. Unlike *random*, *IP* takes into account the current ML model $M_{\hat{D}}$ in seeking to generate synthetic instances that have a greater effect on the objective J. The model is accounted for using *borderline* instances, which are data points that lie close to the decision boundaries of the model and thus have more impact (Han et al., 2005).

To quantify the value of different base instances, we associate a weight w_i with each base instance i in the BP $\mathcal{P}$. Weights are pre-computed using a similar strategy followed in Han et al. (2005), where instances are classified as *noisy*, *safe*, or *borderline* based on the number of nearest neighbours with the same and different class labels, and the highest weight is assigned to borderline instances (see supplement for details).

Algorithm 2: PreSelectBP

Input: input dataset D,
feedback rule set $\mathcal{F}$,
number of nearest neighbours k
Output: initial base population BP

```
1  L ← k + 1
2  BP ← ∅
3  for each rule R in F do
4      if cov(R, D) < L then
5          max_sup ← 0
6          max_cond_R ← nil
7          while max_sup < L do
8              for each condition c in R^s do
9                  R' ← R
10                 remove condition c from R'^s
11                 if R'^s is empty then
12                     max_sup ← |D|
13                     max_cond_R ← R'
14                 end
15                 else
16                     if cov(R', D) > max_sup then
17                         max_sup ← cov(R', D)
18                         max_cond_R ← R'
19                     end
20                 end
21                 R ← max_cond_R
22             end
23         end
24     end
25     BP ← BP ∪ cov(R, D)
26 end
```

Let z_i be a binary variable such that $z_i = 1$ if the i-th instance in the BP $\mathcal{P}$ is selected, and $z_i = 0$ otherwise. Given $\mathcal{P}$, we define a matrix $\mathbf{A}$ with entries a_{ji} and dimensions $m \times p$, where m represents the number of rules and $p = |\mathcal{P}|$, such that $a_{ji} = 1$ if instance i is covered by feedback rule j and $a_{ji} = 0$ otherwise. Then the problem of selecting base instances can be stated as the following IP:

$$\max_{z \in \{0,1\}^p} \sum_{i \in \mathcal{P}} w_i z_i, \text{s.t.}, k{+}1 \leq \sum_{i \in \mathcal{P}} a_{ji} z_i \leq \frac{\eta}{m}, j = 1, \ldots, m. \tag{5}$$

The objective is to maximize the weighted selection of base instances subject to lower and upper bounds on the number of instances selected for each rule. Since the data augmentation step described in the next section seeks k neighbours, the lower bound is set to $k + 1$. This also preserves the per-rule diversity in the BP. The upper bound is the number of instances to generate divided by the number of rules. Non-uniform allocations of instances to rules are also possible.

Despite (5) being an IP, in practice it can be solved quickly as linear relaxations directly provide integral optimal solutions in most cases. Furthermore, *IP* avoids any evaluation of the objective function (4) in selecting base instances. In the supplement, we also discuss an approach that simplifies the evaluation of (4) by using online learning in place of the more expensive black-box algorithm A.

4.2 Synthetic Instance Generation

Motivated from SMOTE and its extension to categorical attributes, SMOTE-NC (Chawla et al., 2002), we design a methodology to generate synthetic instances (line 8 of Algorithm 1) for each selected base instance in line 7. SMOTE generates synthetic instances that lie between a base instance and one of its k nearest neighbours, selected at random. For numerical attributes, the generated value is distributed uniformly on the line segment between the base instance and the neighbour. sing Equation 6:

$$f_v = x_i^f + (x_j^f - x_i^f) \times \omega(0, 1) \tag{6}$$

where $\omega(0, 1)$ denotes a random number between (0,1). For categorical attributes, the value is the majority value among the neighbours. Following the recommendation of Chawla et al. (2002); Han et al. (2005), we set the number of neighbours $k = 5$.

FROTE's generation method differs from SMOTE in the following ways: First, nearest neighbours are found without the constraint that they have the same class label as the base instance, but with the constraint that they satisfy the same feedback rule (possibly relaxed). Second, we require that the generated instance satisfies the conditions of the original, *unrelaxed* rule. This happens automatically if the rule was not relaxed, but if it was, then special logic is needed as described in the supplement. Third, the class label for the generated instance is sampled from the distribution π of the rule (or simply assigned if the rule is deterministic) rather than being equal to the label of the base instance.

Table 1. Properties of the datasets used during the experiments. #Ins, #Labels, and #Feat. stands for the number of instances, number of class labels and number of features (numeric/nominal) of the datasets, respectively.

Dataset	#Ins.	#Feat.	#Labels
Adult	45222	12(4/8)	2
Breast Cancer	569	32(32/-)	2
Nursery	12958	8(-/8)	4
Wine Quality (white)	4898	11(11/-)	7
Mushroom	8124	21(-/21)	2
Contraceptive	1473	9(2/7)	3
Car	1728	6(-/6)	4
Splice	3190	60(-/60)	3

5 Experimental Evaluation

5.1 Experimental Setup

Datasets, ML Models, Feedback Rules

To evaluate the effectiveness of FROTE, we experimented with eight real-world benchmark datasets from UCI[1], properties of which are provided in Table 1. To generate realistic feedback rules, we follow the process mentioned in the Introduction by leveraging *Boolean Rules via Column Generation (BRCG)* algorithm Dash et al. (2018) to obtain a rule set explanation for an initial ML model, and then artificially perturbing these rules to simulate users providing feedback that deviates from the model's predictions. For each rule extracted from Dash et al. (2018), we performed the following three perturbations until we generate 100 rules for each dataset with coverage satisfying $0.05 \leq |\text{cov}(s, D)|/|D| < 0.25$: For each rule extracted, 1. A predicate is randomly selected from the rule's clause and the operator is *reversed*. For instance, if the operator is $\neq$, it is changed to $=$, and similarly if the operator is $\leq$, it is changed to $\geq$, respectively. 2. Value of the selected predicate is updated based on its values in the training dataset. For instance, for categorical attributes, any randomly selected value other than the value of the current predicate is picked and assigned. Similarly for the numerical attributes, a value within the range of the minimum and the maximum values of that attribute observed in the training dataset is assigned. 3. An existing condition from any other rule is randomly picked and added to the rule's conditions. We generated 100 feedback rules in this manner for each dataset, where each generated rule has coverage satisfying $0.05 \leq |\text{cov}(s, D)|/|D| < 0.25$. Rules are de-

[1] https://archive.ics.uci.edu/ml/datasets.php

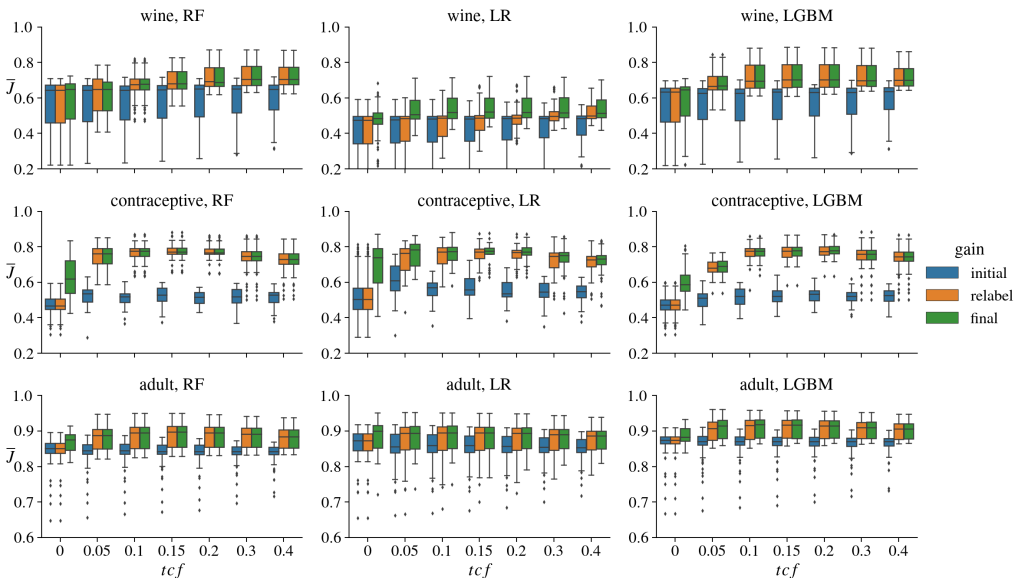

Figure 2. Experiments with models trained on initial training dataset (*initial*), after relabelling (*relabel*), and after FROTE completes augmentation (*final*). *random* selection strategy is used. Standard box plot shows interquartile range (IQR) and whiskers show $1.5 \times IQR$ based on 30 draws for each of $|\mathcal{F}| \in \{1, 3, 5\}$. Results with other datasets included in Section B.

terministic except for the probabilistic rules experiment in Section B.

Classification models. We used three classification algorithms: scikit-learn's Random Forest (RF) and Logistic Regression (LR), and LightGBM (LGBM) (Ke et al., 2017). Default parameter settings are used except for `max_iter` $=$ 500 for LR and `max_depth` $=$ 3 for RF. For finding nearest neighbours in FROTE, scikit-learn's Nearest Neighbors (Pedregosa et al., 2011) algorithm with `algorithm=ball_tree` is utilized.

FRS selection and train-test splitting. We experimented with FRS sizes $|\mathcal{F}| \in \{1, 3, 5, 8, 10, 15, 20\}$, and for each run, we randomly draw this many rules from the pools of 100 generated as described above. We used the following mechanism to vary the level of support of the FRS in the initial training data. For each dataset D and FRS $\mathcal{F}$, D is partitioned into *coverage* ($\text{cov}(\mathcal{F}, D)$) and *outside-coverage* ($D - \text{cov}(\mathcal{F}, D)$) sets. $D - \text{cov}(\mathcal{F}, D)$ is randomly partitioned on a $(80\% - 20\%)$ basis into training and test. For the coverage set $\text{cov}(\mathcal{F}, D)$, we vary the *training coverage fraction* (tcf), i.e. the fraction of the coverage set included in the training set. That is, $tcf \times |\text{cov}(\mathcal{F}, D)|$ randomly selected instances are added to the training partition of $D - \text{cov}(\mathcal{F}, D)$, and the remainder to the test partition of $D - \text{cov}(\mathcal{F}, D)$. We experimented with $tcf \in \{0, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4\}$. $tcf = 0$ tests the scenario where the FRS has no coverage in the initial training set, for example when a new rule emerges.

We perform 30 to 50 runs as described in the previous paragraph for each experimental setting, depending on the size of the dataset. This method of randomly drawing a new rule set and train-test split for each run increases the variability of rule sets tested (and their impact on the results) compared to fixing a rule set and performing cross-validation with it. All algorithm variations are compared using the same rule sets and splits.

Metrics. FROTE uses only the training dataset for augmentation and all evaluation results are reported on the held-out test set. We report values of the complement of J, $\overline{J} = 1 - J$, where $\overline{J}$ is a weighted average as in (3), weighted by rule coverage probabilities $\Pr(\text{cov}(s_r, D))$ in the test set, the first term is the MRA discussed previously (with L_1 as 0-1 loss), and the last term is F_1 score to evaluate model performance on the outside-coverage population. In running FROTE however, we simply use a 0.5-0.5 weighting between MRA and F_1 score in evaluating $\hat{J}_{\hat{D}}$. This is because the test set coverage probabilities are not known to FROTE and may not be equal to the training set probabilities.

Input dataset choices. We experiment with three choices of input dataset D to FROTE. In addition to 1) taking the training dataset as it is (denoted *none* for no modification), instances in $\text{cov}(\mathcal{F}, D)$ that do not have the same class label as the feedback rules covering those instances may be 2) relabelled to agree with the covering rules (*relabel*) or 3) dropped (*drop*). *relabel* is used in all experiments except for the one that evaluates input dataset choices. It is important to note that *relabel* and *drop* may not be possible if the user is reluctant to make changes to the existing dataset for various

data integrity reasons.

Configuration. The number of instances generated per iteration (η) is set to 200 for Adult dataset, 50 for Nursery, Mushroom, Splice, and Wine datasets, and 20 for Car, Contraceptive and Breast Cancer datasets. $\tau = 200$ is used as the iteration limit for all the experiments. We used $k = 5$ and $q = 0.5$ for all the experiments except the ones we evaluated the effect of these two parameters. All experiments were limited to 24 hours and runs that exceed this time limit were terminated. We ran all experiments on a 2.6GHz CPU with 20GB of RAM and they were run deterministically with consistent random number generator seed (42).

5.2 Results and Discussion

Benefit of augmentation. In Figure 2, we compare the test set $\overline{J}$ values obtained from models trained on 1) the initial training dataset, 2) after relabelling based on the FRS (*relabel*), and 3) after FROTE completes augmentation. The comparison is shown for the three ML models, a range of training coverages, and three of the datasets with the remainder in Section B. Even after relabelling, FROTE's augmentation improves $\overline{J}$ for all models and datasets compared to relabelling alone (*final* vs. *relabel*). This finding is further supported by similar plots in Section B of *differences* in $\overline{J}$ between *final* and *relabel*, the vast majority of which are positive. Not surprisingly, the same conclusion holds more strongly for the *drop* and *none* options (see Section B).

Two trends are evident from Figure 2. First, the improvement over *relabel* is larger for smaller tcf, and notably for the difficult case of $tcf = 0$ in which the initial training dataset has no coverage of the FRS. This shows that *relabel* is not sufficient and there is a greater need for augmentation when tcf is low. Second, the improvement is larger for LR, which indicates that linear models may require more data to push decision boundaries.

Comparison with the existing work. To the best of our knowledge, the closest work to ours is Overlay (Daly et al., 2021), which includes two approaches, *Soft Constraints* and *Hard Constraints*. The former treats the user feedback as a soft constraint and uses the prediction on the transformed instance, and the latter considers the feedback as a hard constraint and uses the feedback rules' prediction for all applicable instances. A similar setting as in the previous experiments is used for this comparison. For each run with a dataset, 3 rules are randomly selected and provided as the Full Knowledge Rule Set (FKRS) (Daly et al., 2021) for Overlay, and as the FRS for FROTE. For each rule set, 50% of the coverage population is included in the training data and rest in the test data. Similarly, for the outside-coverage population, a $50\% - 50\%$ split is performed. The model is trained on the training dataset, and FROTE, *Soft Constraints* and *Hard Constraints* are evaluated on the held-out test set. Overlay is presented for binary classification problem and the experiments reported in (Daly et al., 2021) are performed using binary datasets. Therefore we experimented with only the 3 binary datasets (out of 8), and results are displayed in Table 2 (Results with the adult dataset together with separate MRA and F-Scores are in Section B.) We observe that FROTE performs significantly better than both approaches of Overlay for all datasets. The performance of *Soft Constraints* and *Hard Constraints* differs greatly, which suggests the user feedback rules are too divergent from the decision boundaries of the initial ML model for Overlay to perform well, in line with the findings of Daly et al. (2021). This demonstrates that our solution for integrating user feedback into models through pre-processing achieves a better performance in comparison with a state-of-art post-processing approach.

Table 2. Comparison with Overlay-Soft (soft constraints) and Overlay-Hard (hard constraints) of Daly et al. (2021) on Breast-Cancer and Mushroom datasets. Means and standard deviations computed from 50 runs.

Dataset	Model	$\Delta\overline{J}$		
		Overlay-Soft	Overlay-Hard	FROTE
B.Cancer	LR	-0.008 ± 0.045	-0.237 ± 0.212	0.030 ± 0.008
	RF	0.001 ± 0.003	-0.215 ± 0.204	0.041 ± 0.018
	LGBM	0.006 ± 0.011	-0.207 ± 0.180	0.033 ± 0.015
Mushr.	LR	0.001 ± 0.004	-0.158 ± 0.213	0.014 ± 0.015
	RF	0.001 ± 0.004	-0.153 ± 0.208	0.009 ± 0.008
	LGBM	-0.017 ± 0.091	-0.150 ± 0.206	0.009 ± 0.009

Number of feedback rules. One advantage of FROTE is its capability to work with rule sets containing any number of rules. Figure 3 displays $\overline{J}$ values in the same manner as Figure 2 for feedback rule sets having $8, 10, 15$ and 20 rules. The improvement in $\overline{J}$ is maintained up to 20 rules. Results with other datasets are provided in Section B. Overall, they demonstrate the efficacy of FROTE with larger rule sets.

Base instance selection strategy. We now compare the performance of the two base instance selection strategies, *random* and *IP*. Table 3 shows the $\overline{J}$ improvements for models trained on the final augmented dataset relative to the initial dataset. The amount of augmentation required (as a fraction of the input dataset size) for these improvements for both strategies is included in Section B. There is not a clear winner between *random* and *IP* in terms of $\overline{J}$ (the "win-loss-tie" record based on 3 decimal places is 11-8-5), although *IP* generally adds fewer instances to the dataset. One possible reason behind relatively good performance of *random* is although *IP* appears more informed, *random* may avoid "overfitting", in the sense of selecting base instances that improve the objective function evaluated on the augmented training dataset but not on the held-out test set. Looking at the MRA and F-Score separately (provided in

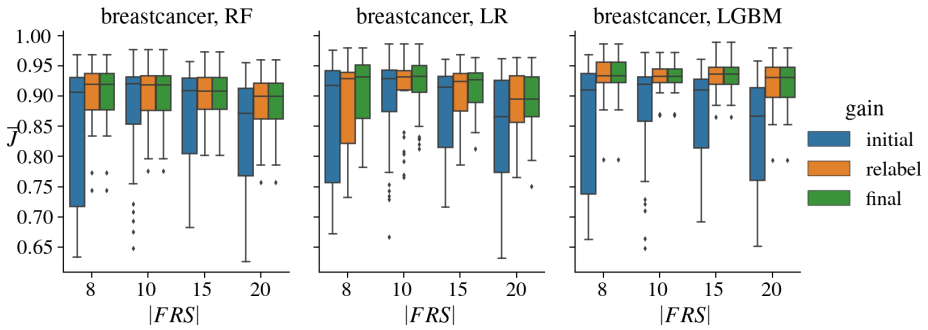

Figure 3. Effect of feedback rule set size for the Breast Cancer dataset and *random* selection strategy. The same comparison as in Figure 2 is shown between *initial*, after *relabel*, and *final* (after FROTE). Each box and whiskers is computed from 30 runs with $tcf = 0.2$.

Table 3. Comparison of *random* and *IP* base instance selection strategies. Means and standard deviations computed from all runs for a given dataset and model.

Dataset	Model	$\Delta\overline{J}$	
		random	*IP*
B. Cancer	RF	0.000 ± 0.003	0.001 ± 0.006
	LR	0.006 ± 0.022	0.006 ± 0.026
	LGBM	0.001 ± 0.008	0.002 ± 0.010
Car	RF	0.005 ± 0.020	0.006 ± 0.020
	LR	0.022 ± 0.034	0.020 ± 0.029
	LGBM	0.008 ± 0.033	0.008 ± 0.027
Mushroom	RF	0.001 ± 0.017	0.004 ± 0.034
	LR	0.005 ± 0.023	0.011 ± 0.049
	LGBM	0.004 ± 0.037	0.006 ± 0.041
Adult	RF	0.003 ± 0.014	0.003 ± 0.011
	LR	0.008 ± 0.023	0.004 ± 0.012
	LGBM	0.004 ± 0.015	0.003 ± 0.011
Wine	RF	0.001 ± 0.007	0.001 ± 0.007
	LR	0.056 ± 0.096	0.055 ± 0.094
	LGBM	0.003 ± 0.015	0.003 ± 0.010
Contracep.	RF	0.032 ± 0.081	0.038 ± 0.085
	LR	0.041 ± 0.099	0.051 ± 0.102
	LGBM	0.027 ± 0.066	0.026 ± 0.057
Nursery	RF	0.031 ± 0.099	0.023 ± 0.076
	LR	0.043 ± 0.088	0.029 ± 0.069
	LGBM	0.035 ± 0.108	0.030 ± 0.096
Splice	RF	0.003 ± 0.017	0.002 ± 0.012
	LR	0.011 ± 0.031	0.007 ± 0.018
	LGBM	0.014 ± 0.049	0.009 ± 0.037

Section B) for the results in Table 3, we see an improvement in MRA without significant decrease (in some cases an increase) in F-Score for both techniques, for all results. However, the degree of improvement is dependent on the dataset and model.

6 Broader Impact and Discussion

One important point to note is that there is generally an inflection point in terms of the number of data points added where the cost to overall model performance starts to outweigh the improvement in MRA. This inflection point also depends on the model used and the dataset. It can be explained by the *data difficulty factors* described in (Stefanowski, 2016), namely *an effect of too strong overlap between classes*, and *a presence of too many examples of one class inside the other class's region*. Another point we want to highlight is that, we define *good candidates* as original data points that satisfy a rule's conditions completely. If there are no such points, then FROTE uses rule relaxation as discussed in the *Base population pre-selection* subsection and in Algorithm 2. While rule relaxation tries to select instances that are more similar to the population targeted by the rule, it may indeed select instances that are far from satisfying the original rule. However, this is mitigated by requiring generated instances to still satisfy the conditions of the original unrelaxed rule.

One limitation of the work is that it may be restricted to tabular data, however, we believe similar mechanisms could be used when considering images where Boolean rules could show relevant images segments or features.

Our work supports model editing where the final ML model will encode the decision processes of not just the underlying data but also external knowledge. This ability can be leveraged to correct incorrect assumptions in the original data or encoded updated policies. On the other hand this introduces the ability for the model builder to influence the model outcomes which could intentionally or unintentionally introduce bias. The user feedback however is interpretable and transparent and user influence is in the form of a Boolean feedback rule. This supports easy integrating into a governance framework such as proposed in (Arnold et al., 2019) where clear auditing of the original data, the feedback rules and the newly created dataset can be stored to transparently log the updates to the model and capture the lineage of the data. Post processing analysis to compare the original and the resulting model could also be leveraged to ensure unintended biases have not been introduces (Bellamy et al.,

2019) along with generating an interpretable model comparison of the two models as proposed by Nair et. al. (Nair et al., 2021). Additionally, FROTE achieves this while trying to minimise the model accuracy for other segments of the dataset. This is in contrast to human labelling or relabelling tasks where the downstream impact of the newly labeled data points may be unclear. Additionally, the source of the newly labeled data, their level or expertise, familiarity with the data are all opaque. One could argue peer reviewing a feedback rule set to obtain consensus among stake holders is relatively easy compared to ensuring a consistent view is being used among data labellers.

7 Conclusion

We presented the problem of pre-processing training dataset to edit an ML model based on feedback rules. We proposed FROTE, a novel technique based on data augmentation, to solve this problem. Empirical studies on real datasets with different ML models demonstrate its effectiveness. Our work supports model editing where the final model encodes decision processes of not just the underlying data but also external knowledge. This ability can be leveraged to correct deficiencies in the original data or adapt to updated policies. User feedback is interpretable and transparent as it is in the form of Boolean rules, supporting clear auditing and governance. A promising future direction is to experiment on different base population selection strategies and optimization techniques to select the base instances and their neighbors together, in order to improve the performance.

References

Arnold, M., Bellamy, R. K., Hind, M., Houde, S., Mehta, S., Mojsilović, A., Nair, R., Ramamurthy, K. N., Olteanu, A., Piorkowski, D., et al. Factsheets: Increasing trust in ai services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5):6–1, 2019.

Awasthi, A., Ghosh, S., Goyal, R., and Sarawagi, S. Learning from rules generalizing labeled exemplars. *arXiv preprint arXiv:2004.06025*, 2020.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

Cakmak, M., Chao, C., and Thomaz, A. L. Designing interactions for robot active learners. *IEEE Transactions on Autonomous Mental Development*, 2(2):108–118, 2010.

Calders, T., Kamiran, F., and Pechenizkiy, M. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pp. 13–18. IEEE, 2009.

Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. Optimized pre-processing for discrimination prevention. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 3995–4004, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, June 2002. ISSN 1076-9757.

Dai, W., Yang, Q., Xue, G.-R., and Yu, Y. Boosting for transfer learning. In *Proceedings of the 24th international conference on Machine learning*, pp. 193–200, 2007.

Daly, E. M., Mattetti, M., Alkan, Ö., and Nair, R. User driven model adjustment via boolean rule explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 5896–5904, 2021.

Dash, S., Günlük, O., and Wei, D. Boolean decision rules via column generation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 4660–4670, Red Hook, NY, USA, 2018. Curran Associates Inc.

Douzas, G. and Bacao, F. Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with applications*, 91: 464–471, 2018.

Eaton, E. and desJardins, M. Selective transfer between learning tasks using task-based boosting. In *AAAI*, 2011.

Fernández, A., García, S., Herrera, F., and Chawla, N. V. Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *J. Artif. Int. Res.*, 61(1):863–905, January 2018. ISSN 1076-9757.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

Guillory, A. and Bilmes, J. A. Simultaneous learning and covering with adversarial noise. In *ICML*, 2011.

Han, H., Wang, W.-Y., and Mao, B.-H. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International Conference on Intelligent Computing*, pp. 878–887. Springer, 2005.

Kapoor, A., Lee, B., Tan, D., and Horvitz, E. Interactive optimization for steering machine classification. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1343–1352, 2010.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154, 2017.

Khandani, A. E., Kim, A. J., and Lo, A. W. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010.

Lakkaraju, H., Bach, S. H., and Leskovec, J. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 1675–1684, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322.

Lauer, F. and Bloch, G. Incorporating prior knowledge in support vector machines for classification: A review. *Neurocomputing*, 71(7):1578–1594, 2008. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2007.04.010. URL https://www.sciencedirect.com/science/article/pii/S0925231207001439. Progress in Modeling, Theory, and Application of Computational Intelligenc.

Letham, B., Rudin, C., McCormick, T., and Madigan, D. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9:1350–1371, 09 2015. doi: 10.1214/15-AOAS848.

Maclin, R., Shavlik, J., Walker, T., and Torrey, L. A simple and effective method for incorporating advice into kernel methods. In *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference*, 01 2006.

Molnar, C. *Interpretable Machine Learning*. 2019. https://christophm.github.io/interpretable-ml-book/.

Nair, R., Mattetti, M., Daly, E., Wei, D., Alkan, O., and Zhang, Y. What changed? interpretable model comparison. In *IJCAI 2021*, 2021.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., and Ré, C. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, pp. 269. NIH Public Access, 2017.

Ribeiro, M. T., Singh, S., and Guestrin, C. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Sharma, S., Zhang, Y., Ríos Aliaga, J. M., Bouneffouf, D., Muthusamy, V., and Varshney, K. R. Data augmentation for discrimination prevention and bias disambiguation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pp. 358–364, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375865. URL https://doi.org/10.1145/3375627.3375865.

Singh, J. and Urolagin, S. Use of artificial intelligence for health insurance claims automation. In *Advances in Machine Learning and Computational Intelligence*, pp. 381–392. Springer, 2020.

Stefanowski, J. *Dealing with Data Difficulty Factors While Learning from Imbalanced Data*, pp. 333–363. Springer International Publishing, Cham, 2016. ISBN 978-3-319-18781-5. doi: 10.1007/978-3-319-18781-5_17. URL https://doi.org/10.1007/978-3-319-18781-5_17.

Tanaka, F. H. K. d. S. and Aranha, C. Data augmentation using gans. *arXiv preprint arXiv:1904.09135*, 2019.

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. Modeling tabular data using conditional gan. In *Advances in Neural Information Processing Systems*, 2019.

Zhang, Y. and Deng, A. Redundancy rules reduction in rule-based knowledge bases. pp. 639–643, 08 2015. doi: 10.1109/FSKD.2015.7382017.

A SOLUTION DETAILS

Subset selection via integer programming. We elaborate on the integer programming formulation for the subset selection problem presented in the main paper. For a given $\mathcal{F}$ we would like to determine a subset of the training data within $\mathcal{F}$ that has the greatest influence on the model decision boundaries.

The weight w_i reflects the value of a data point for the final selection. Instances near the decision boundary are more valuable, as it has a greater potentially to influence the model. This weight is pre-computed as follows:

For each j, $j \in D$ compute, p as the number of k neighbours who have the same label, and q as the number of k neighbours who have a different label. Here, the label refers to the predicted label from a model we seek to edit. If $q >> p$, the observation can be considered *noisy*, $p >> q$, then the observation can be considered as *safe*, and if $p \approx q$ the observation can be considered as *borderline* (Han et al., 2005). Correspondingly, the weights w_i can be assigned based on these three cases such that, the borderline points is assigned the largest weight. In our experiments, we set $w_i = 3$ for borderline and $w_i = 1$ for noisy and safe data points computed within $k = 10$ nearest neighbors.

Subset selection with online learning. As mentioned in the main text, we also considered the use of online learning to simplify the evaluation of objective function (3), and specifically to avoid running training algorithm A. We instead take a proxy approach in which 1) the current model $M_{\hat{D}} = A(\hat{D})$ is approximated by a (parametric) model $\hat{M}$ to which online learning can be applied, and 2) the retrained model $M_{D'}$ is approximated by the result of online learning, starting from $\hat{M}$ and updating based on the generated instances $\text{Generate}(\mathcal{B})$. Recalling that J is also replaced by its empirical approximation $\hat{J}_{\hat{D}}$ over $\hat{D}$, the online learning approximation can thus be written as

$$J\left(A\left(\hat{D} \cup \text{Generate}(\mathcal{B})\right), \mathcal{F}\right) \approx \hat{J}_{\hat{D}}\left(\text{OL}\left(\hat{M}, \text{Generate}(\mathcal{B})\right), \mathcal{F}\right). \tag{7}$$

We investigated the use of 7 to approximate objective function (3) for singleton sets $\mathcal{B} = \{i\}$, $i \in \mathcal{P}$. Such evaluations on singletons might be summed to provide a crude approximation to (3) for non-singleton $\mathcal{B}$; the *IP* objective function (4) is also a sum approximation in this sense. They could also constitute the first iteration in a greedy algorithm for selecting $\mathcal{B}$.

Our experience thus far however is that even the evaluation of (7) is still too computationally intensive to be practical (at least in terms of facilitating experimentation). To be more specific, we used the Vowpal Wabbit library[2] for online learning with a plain logistic regression model $\hat{M}$. Step 1) of approximating $M_{\hat{D}}$ with $\hat{M}$ is done by training $\hat{M}$ on dataset $\hat{D}$ and the outputs of $M_{\hat{D}}$ on $\hat{D}$. This has computational complexity $O(|\hat{D}|)$. Likewise, step 2), i.e. approximating $M_{D'}$ by updating $\hat{M}$ for each generated instance $\text{Generate}(\{i\})$, $i \in \mathcal{P}$, also has complexity $O(|\mathcal{P}|) = O(|\hat{D}|)$. However, evaluating $\hat{J}_{\hat{D}}$ for each of these updated models results in complexity $O(|\hat{D}|^2)$, and we have found this to be the slow step in our limited experiments. Future work could consider further approximations to the objective function that avoid higher than first-order complexity in $|\hat{D}|$.

Synthetic instance generation. Synthetic instance generation is used by the *Generate()* procedure within FROTE, as outlined in Algorithm 1 of main paper (line 9). It is called for each *base instance* and a randomly selected neighbor of it in order to generate synthetic instances. Synthetic instance generation uses two subroutines for populating categorical and numerical attributes.

For populating categorical attributes, algorithm iterates through each categorical attribute to assign a value. For each categorical attribute, initially, all possible values for that attribute are calculated and stored. These attribute values are sorted in the decreasing order of the number of times they occur in the neighbors. Therefore, the first element in the list is the value that occurs in the majority of the k nearest neighbor instances of the base instance. If the corresponding attribute is part of one of the conditions of the rule, then a special check is needed to make sure that the assigned value satisfies the corresponding condition(s). For instance, the algorithm ensures that for a condition with "$\neq$" operator, the value assigned to the corresponding attribute is different than the *value* of that corresponding condition.

The procedure iterates over each of the numerical attributes, and for each attribute, if it is not part of any of the conditions of the rule, the value to the corresponding numerical attribute is assigned using a similar approach to SMOTE (Chawla et al., 2002). If the attribute exists in a condition where the operator is '=', then the value of the corresponding condition is assigned. However, if the attribute exists in a condition where the operator is one of {'>','≥','<','≤'}, extra checks are performed to ensure that the generated value satisfies the corresponding conditions. Specifically, a window is defined with a minimum and maximum value (lines 21-29) based on the specific operators. These bounds keep track of the minimum and/or the maximum values that can be assigned to the corresponding feature of the new instance. They are further adjusted based on the base and neighbor instance values to make sure that

[2] https://vowpalwabbit.org

the new value that will be assigned will stay within the value limits defined by the comparison operators. Finally a *diff* value is assigned based on a tightest window determined by these minimum and maximum values together with the base and the neighbor instances' corresponding attribute values, and *diff* is then used to generate a value for the corresponding attribute.

B EXPERIMENTAL EVALUATION

B.1 Further Experimental Results

Benefit of augmentation. We compare the test set $\overline{J}$ values obtained from the models that are trained on 1) the initial dataset before FROTE, 2) after applying the modification strategy, and 3) after FROTE completes augmentation. In Figure 4, additional plots for Figure 3 of the main paper are given, where the results with Splice, Nursery, Breast Cancer, Mushroom and Car datasets are included. In Figure 4, the improvements of the $\overline{J}$ values observed after 1. modification strategy is applied, and 2. between the augmentation process and mod strategy, is displayed. Both Figure 3 of the main paper and Figure 4 show the results with the *relabel* strategy. Figure 5 and Figure 6 show the results with the *none* strategy, and Figure 7 and Figure 8 show the results with the *drop* strategy. As can be observed from the figures, the variance appears to be higher for both *none* and *drop* strategies, since for the former, existing contradictory instances are remained in the dataset, and for the latter, the base instances are selected through rule relaxation which increases the variety in the base instances. However, for all mod strategies, we can conclude that augmentation can improve MRA without much compromise-in some cases increase- in Δ**F-Score**.

Comparison with the existing work. Additional results for the comparison experiments with (Daly et al., 2021) are included in Tables 9 and 10. We observe from the tables that our solution performs better than a state-of-art post-processing approach, which confirms with the findings presented in the main paper. When we examine the results in Table 10, we see that even *Hard Constraints* has a significantly higher MRA than the *Soft Constraints* for all datasets, it performs very poorly on the outside coverage population, as can be seen from the ΔF-Score values. This demonstrates that a pure post-processing approach can suffer if the rules are deviated from the underlying model. Similar findings are observed for the *Soft Constraints*, however *Soft Constraints* suffers less from the deviation in the rules, since it considers models decisions after applying changes to the data instance based on the rules learnt so far.

Relation between the level of deviation of the rule labels from the ground truth labels. For scenarios when rules' labels deviate from the labels of data instances in test subset, FROTE performs better than both Overlay techniques. Experiments in (Daly et al., 2021) used rules that are learnt from the whole dataset, therefore they ensure that ground truth labels are well aligned with rule labels. In our experiments, we perturb the rules so that the rule labels do not align with ground truth any more. We performed additional experiments to demonstrate this effect, in other words, the relation between the level of deviation of the rule labels from the ground truth labels on the performance of different techniques.

The following steps are followed to perform this experiment:

1. Extract rules from the **whole dataset** using BRCG technique (Dash et al., 2018) through using a logistic regression surrogate.

2. Perturb the rules. The rule perturbation procedure takes a rule and a *decrease in accuracy* threshold, τ. It tries to change one of the values in one of the conditions in if part of the rule such that, the ratio of the difference between the support of the original rule (R) and the perturbed rule (R') is between [τ-0.05, τ+0.05]. For the numerical attributes, we randomly change the numerical value in the condition, whereas for the categorical attributes, we randomly assign one of the values in the permitted value list for that corresponding attribute. One thing to note is that, with this perturbation procedure, we cannot always achieve an accuracy decrease as required by τ. In such a scenario, that rule in the set remains as it is, therefore not perturbed.

3. After steps 1 and 2, we store each ruleset, which is the FRS for FROTE, and $FKRS$ for the Overlay together with its support, corresponding perturbed rule and the perturbed rule's support.

4. Divide the dataset into ($\%80$-$\%20$) train-test split in an informed manner such that, we randomly select ($\%80$) of the population that are accurately covered by the rules in train split. Same performed for the population that are *not* accurately covered by the rules.

5. Both FROTE and Overlay takes the *train* dataset, same model, ruleset created in Step 3 (same ruleset for both FRS and FKRS).

6. After the algorithms complete, the solutions are evaluated on the held out test dataset and the MRA and F-Score's are recorded.

The above procedure is run for 50 times randomly for each dataset, and results are averaged over these 50 runs. Both *means* and *standard deviations* are shown as part of the results. During these experiments, we use Breast Cancer and Banknote datasets from UCI [3], which were both used during evaluations in (Daly et al., 2021). In Table 4, examples of the perturbed rules for Banknote dataset are given.

Results are given in Table 5. For the Banknote dataset, F-Score columns are empty since the rules (FRS/FKRS) cover the whole population of the dataset. Therefore, there does not exist any instance within both the train and the test dataset that are not covered by the rules.

[3] https://archive.ics.uci.edu/ml/datasets/

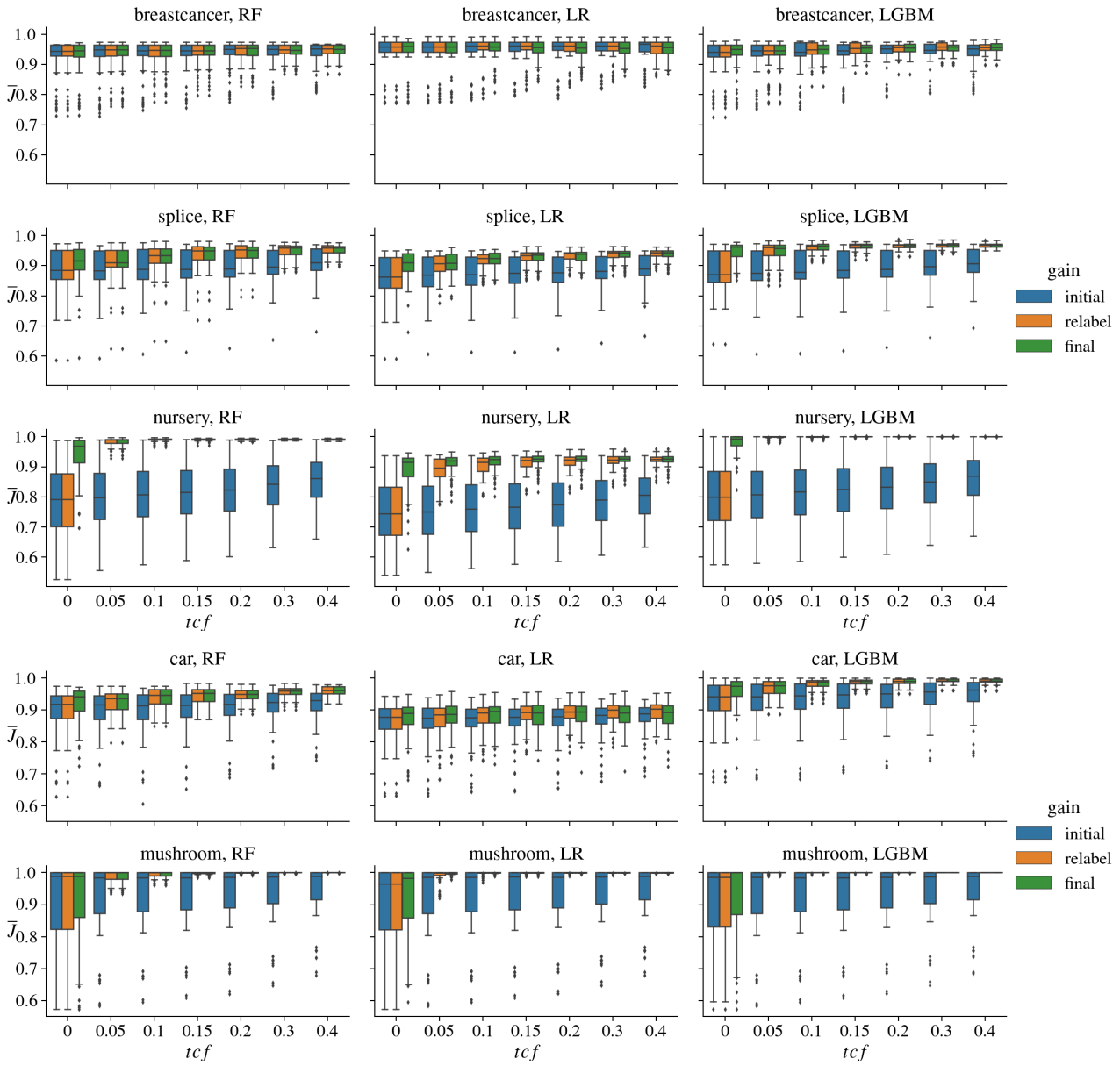

Figure 4. Additional plots for Figure 2 in the main paper. Experiments with models trained on the initial dataset before FROTE (*initial*), after applying the *relabel* mod strategy, and after FROTE completes augmentation (*final*). The comparison is shown as a function of the training coverage fraction of the feedback rule sets and for different ML models and the Splice, Nursery and Breast Cancer, Mushroom and Car datasets. The *random* selection strategy is used. Standard box plot showing interquartile range (IQR) and whiskers showing 1.5 times IQR based on 30 random draws for each of $|\mathcal{F}| \in \{1, 3, 5\}$.

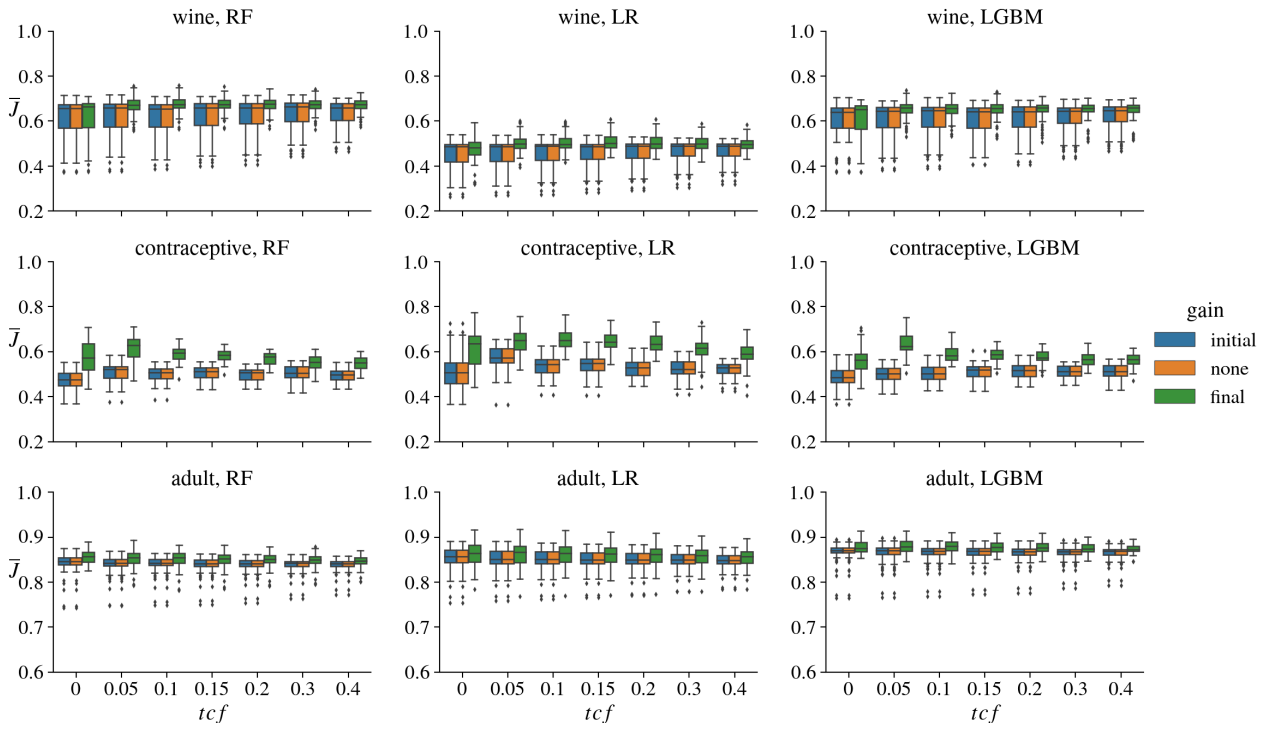

Figure 5. Additional plots for Figure 2 in the main paper. Experiments with models trained on the initial dataset before FROTE (*initial*), after applying the *none* strategy, and after FROTE completes augmentation (*final*). $mod - imp$ and $final - imp$ represent the differences in $\overline{J}$ between *mod* and *initial* and *final* and *mod*, respectively. The comparison is shown as a function of the training coverage fraction of the feedback rule sets and for different ML models and all the datasets. The *random* selection strategy is used. Standard box plot showing interquartile range (IQR) and whiskers showing 1.5 times IQR based on 30 random draws for each of $|\mathcal{F}| \in \{1, 3, 5\}$.

As we can see from the results, the more the ruleset accuracy decreases, the more the MRA of Overlay-Soft tends to decrease for both datasets. However, FROTE's performance increases as we have less instances that are correctly covered by the rules. Therefore, for less accurate rules with respect to the current dataset, FROTE performs better. In other words, if we have less accurately supporting instances in the dataset, FROTE performs better than Overlay considering MRA. For F-Score, performances are similar, however Overlay-Hard's performance is very low when we check the F-Score. This is due to the fact that, Overlay-Hard completely depends on the rule logic, therefore if rules are accurate, Overlay-Hard performs very well for instances that are covered by the rules. That also explains why its performance is highest for Banknote dataset, since the whole dataset is covered by the rules.

Augmentation progress. In Figure 9, we evaluate $\overline{J}$ on the held-out test set for intermediate models trained on D' (i.e. augmented training dataset at the end of each iteration) as a function of the number of synthetic instances added, to illustrate how these change for different models and tcf values. For all models, $\overline{J}$ improves more quickly for lower training coverage. RF needs fewer instances to reach $\overline{J} = 1$ in comparison with LR and LGBM. This again suggests that non-linear models like RF may require less data to edit than linear models.

Number of feedback rules. Additional plots for displaying the effect of number of rules on the performance of the solution are given in Figure 10. For all datasets, we experimented with $|\mathcal{F}|$={8,10,15,20}, however for some datasets, for $|\mathcal{F}|$=15 and $|\mathcal{F}|$=20, no such conflict-free $\mathcal{F}$ can be found out of 500 rules. Therefore, we included the results for the experiments for which a conflict-free rule set with the experimented size can be formed.

As it is observed from the results, FROTE improves J both after the relabel modification strategy and after the data augmentation. Overall, results demonstrate the efficacy of the approach even with larger rule sets.

Base instance selection strategy. Performance of the two base instance selection strategies, *IP* and *random* were compared in the main paper using the improvement in $\overline{J}$. In Table 6, we have included the number of instances added to achieve those improvements. In Table 7, improvements in MRA and F-Score are reported separately. We observe that the improvement in $\overline{J}$ is highly dominated by the improve-

Table 4. Examples for rules and corresponding perturbed rules for Banknote dataset for different perturbation degrees (decrease in accuracy for the rule sets).

Dataset	Decrease in Accuracy	Original Rule/Perturbed Rule

Banknote	[%5-%15]	IF (curtosis $\leq$ 4.5641) THEN target_class = 0 (661)
		IF (curtosis $\leq$ 2.5875) THEN target_class = 0 (563)
		
		IF (variance $\leq$ -3.3125) THEN target_class = 1 (137)
		IF (variance $\leq$ -3.4982) THEN target_class = 1 (123)
		
		IF (curtosis $\leq$ 1.4238 and skewness $\leq$ 5.82499 and variance $\leq$ 2.2892) THEN target_class = 1 (325)
		IF (curtosis $\leq$ 1.4238 and skewness $\leq$ 5.82499 and variance $\leq$ 0.8765) THEN target_class = 1 (286)
		
		IF (skewness $\leq$ 0.7201 and variance $\leq$ -0.40804) THEN target_class = 1 (292)
		IF (skewness $\leq$ 0.7201 and variance $\leq$ -0.93395 THEN target_class = 1 (256)
		
		IF (curtosis $\leq$ 7.8929 and skewness $\leq$ -3.3895 and variance $\leq$ 0.49571) THEN target_class = 1 (80)
		IF (curtosis $\leq$ 7.8929 and skewness $\leq$ -3.3895 and variance $\leq$ -0.6987) THEN target_class = 1 (70)

Banknote	[%65-%75]	IF (curtosis $\leq$ 4.5641) THEN target_class = 0 (661)
		IF (curtosis $\leq$ -1.8100) THEN target_class = 0 (186)
		
		IF (variance $\leq$ -3.3125) THEN target_class = 1 (137)
		IF (variance $\leq$ -4.9341) THEN target_class = 1 (39)
		
		IF (curtosis $\leq$ 1.4238 and skewness $\leq$ 5.82499 and variance $\leq$ 2.2892) THEN target_class = 1 (325)
		IF (curtosis $\leq$ 1.4238 and skewness $\leq$ 5.82499 and variance $\leq$ -2.1109) THEN target_class = 1 (87)
		
		IF (skewness $\leq$ 0.7201 and variance $\leq$ -0.40804) THEN target_class = 1 (292)
		IF (skewness $\leq$ 0.7201 and variance $\leq$ -3.04252 THEN target_class = 1 (74)
		
		IF (curtosis $\leq$ 7.8929 and skewness $\leq$ -3.3895 and variance $\leq$ 0.49571) THEN target_class = 1 (80)
		IF (curtosis $\leq$ 7.8929 and skewness $\leq$ -3.3895 and variance $\leq$ -1.92584) THEN target_class = 1 (22)

Table 5. Comparison with Overlay-Soft (soft constraints) and Overlay-Hard (hard constraints) of (Daly et al., 2021) on BreastCancer and Banknote datasets. Logistic Regression model is used. Means and standard deviations computed from 50 runs.

Dataset	Decrease in Accuracy(%)	$F-Score$			MRA		
		Overlay-Hard	Overlay-Soft	FROTE	Overlay-Hard	Overlay-Soft	FROTE
B. Cancer	$[\%5-\%15]$	0.188 ± 0.163	0.877 ± 0.204	0.936 ± 0.035	0.876 ± 0.111	0.871 ± 0.027	0.879 ± 0.023
	$[\%25-\%35]$	0.313 ± 0.295	0.811 ± 0.210	0.939 ± 0.033	0.868 ± 0.156	0.886 ± 0.027	0.920 ± 0.020
	$[\%45-\%55]$	0.250 ± 0.267	1.000 ± 0.000	0.955 ± 0.031	0.885 ± 0.111	0.786 ± 0.032	0.920 ± 0.020
	$[\%65-\%75]$	0.303 ± 0.281	0.866 ± 0.151	0.938 ± 0.037	0.906 ± 0.127	0.771 ± 0.028	0.967 ± 0.014
Banknote	$[\%5-\%15]$	–	–	–	0.908 ± 0.069	0.776 ± 0.136	0.586 ± 0.007
	$[\%25-\%35]$	–	–	–	0.888 ± 0.069	0.760 ± 0.112	0.625 ± 0.006
	$[\%45-\%55]$	–	–	–	0.898 ± 0.083	0.783 ± 0.102	0.707 ± 0.009
	$[\%65-\%75]$	–	–	–	0.910 ± 0.072	0.768 ± 0.115	0.813 ± 0.011

ment in MRA.

Probabilistic rules. In this experiment, we consider probabilistic rules, where the label distribution π is not just a Kronecker delta for one of the classes. The experiment provides a brief demonstration of the ability of probabilistic rules to represent uncertainty and mitigate the effect of an over-confident expert rule. We consider an extreme case of this where the expert provides a single feedback rule, but the test distribution remains the same as the training distribution, i.e., the expert is wrong and the rule does not take effect. (We use only a single feedback rule to try to isolate the effect of having a probabilistic rule and avoid interactions among rules.) We also set $tcf = 0$ (so *relabel* and *drop* initializations are not applicable).

We run FROTE with the following label distribution π for instances generated under the rule: With probability p, the label is equal to the class c specified by the feedback rule. With probability $1-p$, it is equal to the label of the corresponding base instance, except when that label is c, in which case the label of the generated instance is chosen uniformly at random from classes other than c. Thus overall, the labels of generated instances are equal to c with probability p, and otherwise they approximately follow the distribution of the training data (as represented by the base instances) restricted to classes other than c. The case $p = 1$ is the deterministic case used in the other experiments. With $p < 1$, the user of FROTE can express less than full confidence in the expert rule and rely more on the existing training data.

Table 8 shows the MRA and $\overline{J}$ improvements for different probabilities p. In this case, since the feedback rule is not in effect for test data, MRA just measures agreement with respect to labels following the original distribution $p_{X,Y}$, within the coverage of the rule. The MRA column shows

Table 6. Experiments with *IP* and *random* selection strategies for all datasets and models. Number of instances are included as an additional column to Table 2 of the main paper. $\Delta\#Ins/|D|$ is the number of instances added (as a fraction of the dataset size) that leads to the reported improvements. Means and standard deviations are computed from all runs performed with a given dataset and model.

Dataset	Model	$\Delta\overline{J}$ (*random*)	$\Delta\overline{J}$ (*IP*)	Δ#Ins$/|D|$ (*random*)	Δ#Ins$/|D|$ (*IP*)
B.Cancer	RF	0.000 ± 0.003	0.001 ± 0.006	0.011 ± 0.016	0.015 ± 0.042
	LR	0.006 ± 0.022	0.006 ± 0.026	0.298 ± 0.326	0.199 ± 0.266
	LGBM	0.001 ± 0.008	0.002 ± 0.010	0.011 ± 0.016	0.014 ± 0.036
Car	RF	0.005 ± 0.020	0.006 ± 0.020	0.001 ± 0.003	0.001 ± 0.004
	LR	0.022 ± 0.034	0.020 ± 0.029	0.227 ± 0.225	0.113 ± 0.183
	LGBM	0.008 ± 0.033	0.008 ± 0.027	0.001 ± 0.003	0.001 ± 0.004
Mushroom	RF	0.001 ± 0.017	0.004 ± 0.034	0.001 ± 0.002	0.001 ± 0.002
	LR	0.005 ± 0.023	0.011 ± 0.049	0.036 ± 0.136	0.016 ± 0.064
	LGBM	0.004 ± 0.037	0.006 ± 0.041	0.001 ± 0.002	0.001 ± 0.002
Adult	RF	0.003 ± 0.014	0.003 ± 0.011	0.005 ± 0.047	0.004 ± 0.009
	LR	0.008 ± 0.023	0.004 ± 0.012	0.356 ± 0.507	0.185 ± 0.352
	LGBM	0.004 ± 0.015	0.003 ± 0.011	0.059 ± 0.096	0.046 ± 0.083
Wine	RF	0.001 ± 0.007	0.001 ± 0.007	0.004 ± 0.033	0.003 ± 0.016
	LR	0.056 ± 0.096	0.055 ± 0.094	0.136 ± 0.135	0.096 ± 0.098
	LGBM	0.003 ± 0.015	0.003 ± 0.01	0.002 ± 0.009	0.003 ± 0.009
Contracep.	RF	0.032 ± 0.081	0.038 ± 0.085	0.000 ± 0.001	0.001 ± 0.001
	LR	0.041 ± 0.099	0.051 ± 0.102	0.011 ± 0.019	0.008 ± 0.013
	LGBM	0.027 ± 0.066	0.026 ± 0.057	0.001 ± 0.003	0.001 ± 0.003
Nursery	RF	0.031 ± 0.099	0.023 ± 0.076	0.001 ± 0.003	0.001 ± 0.002
	LR	0.043 ± 0.088	0.029 ± 0.069	0.144 ± 0.162	0.031 ± 0.044
	LGBM	0.035 ± 0.108	0.030 ± 0.096	0.001 ± 0.003	0.001 ± 0.002
Splice	RF	0.003 ± 0.017	0.002 ± 0.012	0.009 ± 0.047	0.008 ± 0.044
	LR	0.011 ± 0.031	0.007 ± 0.018	0.091 ± 0.116	0.046 ± 0.079
	LGBM	0.014 ± 0.049	0.009 ± 0.037	0.009 ± 0.047	0.007 ± 0.040

that setting $p = 1.0$, i.e., completely following the expert rule, does not give as good a performance as setting p to a lower, less confident value. This pattern however is not as clear looking at the $\overline{J}$ column. In reality, the best value of p is not known a priori as it depends on the exact extent to which the test data (in this case, the distribution $p_{X,Y}$) conforms to the expert rule. Nevertheless, Table 8 suggests that there is a benefit to using a probabilistic rule with $p < 1$ if there is reason to be less confident in the validity of the feedback rules.

Experiments with k parameter. In Table 11, experiments with different k values are reported. For each k parameter being evaluated, the improvement in $\overline{J}$ is reported. As can be seen from the results, larger k values decreases the $\Delta\overline{J}$ for Wine and Car datasets for all models, whereas the $\Delta\overline{J}$ is higher for higher values of k for the Random Forest model both for the Contraceptive and Breast Cancer datasets. Therefore, the effect of k on the performance is not conclusive, and the effect of it depends on both the dataset and the model under consideration.

Table 7. MRA and F-Score reported separately for the results in Table 1 of the main paper. Same with Table 1, results are reported for ***IP*** and *random* selection strategies.ΔMRA and ΔF-Score represent the improvement in the corresponding metrics ($mean \pm std$). Means and standard deviations are computed from all runs performed with a given dataset and model.

Dataset	Model	Δ MRA (***IP***)	ΔMRA (*random*)	ΔF-Score (***IP***)	ΔF-Score (*random*)
Breastcancer	RF	0.003 ± 0.042	0.002 ± 0.038	0.000 ± 0.005	0.000 ± 0.003
	LR	0.047 ± 0.116	0.039 ± 0.102	-0.006 ± 0.014	-0.006 ± 0.015
	LGBM	0.013 ± 0.092	0.014 ± 0.098	0.000 ± 0.006	0.000 ± 0.005
Car	RF	0.018 ± 0.063	0.015 ± 0.069	0.000 ± 0.003	0.000 ± 0.003
	LR	0.096 ± 0.112	0.109 ± 0.135	-0.020 ± 0.028	-0.026 ± 0.031
	LGBM	0.024 ± 0.083	0.024 ± 0.099	0.000 ± 0.002	0.000 ± 0.002
Mushroom	RF	0.009 ± 0.081	0.002 ± 0.027	-0.000 ± 0.000	-0.000 ± 0.000
	LR	0.045 ± 0.158	0.024 ± 0.111	-0.000 ± 0.001	-0.000 ± 0.001
	LGBM	0.024 ± 0.141	0.018 ± 0.128	-0.000 ± 0.000	-0.000 ± 0.000
Adult	RF	0.011 ± 0.053	0.012 ± 0.073	-0.000 ± 0.001	-0.0 ± 0.001
	LR	0.072 ± 0.170	0.075 ± 0.192	-0.003 ± 0.005	-0.003 ± 0.007
	LGBM	0.026 ± 0.108	0.026 ± 0.117	0.000 ± 0.001	0.000 ± 0.001
Wine	RF	0.018 ± 0.096	0.020 ± 0.107	-0.001 ± 0.005	-0.000 ± 0.005
	LR	0.360 ± 0.306	0.354 ± 0.309	-0.020 ± 0.023	-0.023 ± 0.026
	LGBM	0.043 ± 0.169	0.037 ± 0.155	0.001 ± 0.005	0.001 ± 0.008
Contraceptive	RF	0.070 ± 0.151	0.059 ± 0.144	-0.000 ± 0.009	-0.000 ± 0.007
	LR	0.115 ± 0.214	0.095 ± 0.203	-0.007 ± 0.019	-0.010 ± 0.025
	LGBM	0.048 ± 0.104	0.049 ± 0.119	-0.001 ± 0.010	-0.000 ± 0.009
Nursery	RF	0.059 ± 0.192	0.074 ± 0.226	-0.000 ± 0.001	-0.000 ± 0.001
	LR	0.097 ± 0.217	0.131 ± 0.240	-0.002 ± 0.004	-0.008 ± 0.013
	LGBM	0.073 ± 0.227	0.082 ± 0.242	-0.000 ± 0.000	-0.000 ± 0.000
Splice	RF	0.006 ± 0.026	0.009 ± 0.036	-0.001 ± 0.004	-0.001 ± 0.004
	LR	0.025 ± 0.046	0.035 ± 0.071	-0.002 ± 0.006	-0.004 ± 0.010
	LGBM	0.022 ± 0.098	0.032 ± 0.116	0.000 ± 0.002	0.000 ± 0.002

Table 8. Experiments with probabilistic rules. Means and standard deviations computed from 50 runs for LR model and for the given datasets. For each run, $|FRS| = 1$, and $tcf = 0$. *random* selection strategy is utilized during the experiments.

Dataset	Probability	ΔMRA	$\Delta\overline{J}$
Mushroom	$p = 0.4$	0.206 ± 0344	0.007 ± 0.012
	$p = 0.6$	0.242 ± 0.386	0.009 ± 0.014
	$p = 0.8$	0.249 ± 0.390	0.009 ± 0.014
	$p = 1.0$	0.173 ± 0.296	0.006 ± 0.011
Wine	$p = 0.4$	0.416 ± 0.305	-0.011 ± 0.024
	$p = 0.6$	0.448 ± 0.317	-0.010 ± 0.021
	$p = 0.8$	0.423 ± 0.348	-0.011 ± 0.020
	$p = 1.0$	0.338 ± 0.327	-0.008 ± 0.016
B. Cancer	$p = 0.4$	0.005 ± 0.015	0.003 ± 0.007
	$p = 0.6$	0.005 ± 0.015	0.002 ± 0.006
	$p = 0.8$	0.007 ± 0.015	0.002 ± 0.007
	$p = 1.0$	0.005 ± 0.015	0.003 ± 0.006

Table 9. Comparison with Overlay-Soft (soft constraints) and Overlay-Hard (hard constraints) of Daly et al. (2021) on Adult dataset. Means and standard deviations computed from 50 runs.

Dataset	Model	$\Delta\overline{J}$		
		Overlay-Soft	Overlay-Hard	FROTE
Adult	LR	-0.015 ± 0.034	-0.107 ± 0.111	0.025 ± 0.039
	RF	0.114 ± 0.013	-0.121 ± 0.019	0.036 ± 0.039
	LGBM	0.102 ± 0.021	-0.018 ± 0.180	0.240 ± 0.043

Table 10. Experiments with *Overlay* (Daly et al., 2021). Overlay-Hard and Overlay-Soft refers to the Hard Constraints and Soft Constraints approaches of Overlay. The comparison is shown for different ML models and the Breast Cancer, Mushroom and Adult datasets. *random* selection strategy is used for FROTE. Means and standard deviations are computed from 50 runs, where for each run a different set of 3 rules are used.

Model	**ΔMRA**			**ΔF-Score**		
			B. Cancer			
	Overlay-Soft	Overlay-Hard	FROTE	Overlay-Soft	Overlay-Hard	FROTE
LR	0.008 ± 0.021	0.021 ± 0.232	0.080 ± 0.168	-0.012 ± 0.058	-0.313 ± 0.248	-0.014 ± 0.022
RF	0.005 ± 0.015	0.071 ± 0.224	0.097 ± 0.194	0.000 ± 0.000	-0.299 ± 0.238	-0.002 ± 0.007
LGBM	0.016 ± 0.032	0.072 ± 0.218	0.880 ± 0.238	-0.000 ± 0.001	-0.272 ± 0.198	-0.001 ± 0.009
			Mushroom			
	Overlay-Soft	Overlay-Hard	FROTE	Overlay-Soft	Overlay-Hard	FROTE
LR	0.046 ± 0.091	0.202 ± 0.34	0.049 ± 0.033	-0.001 ± 0.004	-0.168 ± 0.223	-0.000 ± 0.001
RF	0.021 ± 0.114	0.205 ± 0.34	0.040 ± 0.032	0.000 ± 0.000	-0.166 ± 0.220	0.000 ± 0.000
LGBM	0.155 ± 0.302	0.208 ± 0.34	0.049 ± 0.033	-0.023 ± 0.093	-0.163 ± 0.218	0.000 ± 0.000

Table 11. Experiments with different values of k parameter. Results are reported using *random*. $\Delta\overline{J}$ represent the improvement in the corresponding metric ($mean \pm std$). Means and standard deviations are computed from 20 runs for each row in the table.

Dataset	**Model**	$k=3, \Delta\overline{J}$	$k=5, \Delta\overline{J}$	$k=8, \Delta\overline{J}$	$k=10, \Delta\overline{J}$
Contraceptive	RF	0.268 ± 0.156	0.204 ± 0.150	0.199 ± 0.116	0.159 ± 0.080
	LR	0.470 ± 0.280	0.484 ± 0.289	0.464 ± 0.286	0.447 ± 0.264
	LGBM	0.305 ± 0.132	0.288 ± 0.142	0.222 ± 0.096	0.256 ± 0.130
Wine	RF	0.525 ± 0.193	0.518 ± 0.185	0.513 ± 0.174	0.503 ± 0.189
	LR	0.750 ± 0.257	0.750 ± 0.256	0.748 ± 0.255	0.747 ± 0.256
	LGBM	0.506 ± 0.15	0.455 ± 0.158	0.437 ± 0.125	0.429 ± 0.132
Breastcancer	RF	0.178 ± 0.147	0.200 ± 0.145	0.226 ± 0.193	0.230 ± 0.218
	LR	0.315 ± 0.265	0.308 ± 0.258	0.208 ± 0.225	0.168 ± 0.204
	LGBM	0.320 ± 0.252	0.323 ± 0.223	0.276 ± 0.236	0.269 ± 0.195
Car	RF	0.305 ± 0.104	0.250 ± 0.090	0.195 ± 0.057	0.185 ± 0.093
	LR	0.660 ± 0.276	0.563 ± 0.286	0.473 ± 0.190	0.414 ± 0.163
	LGBM	0.340 ± 0.137	0.233 ± 0.089	0.209 ± 0.078	0.183 ± 0.068

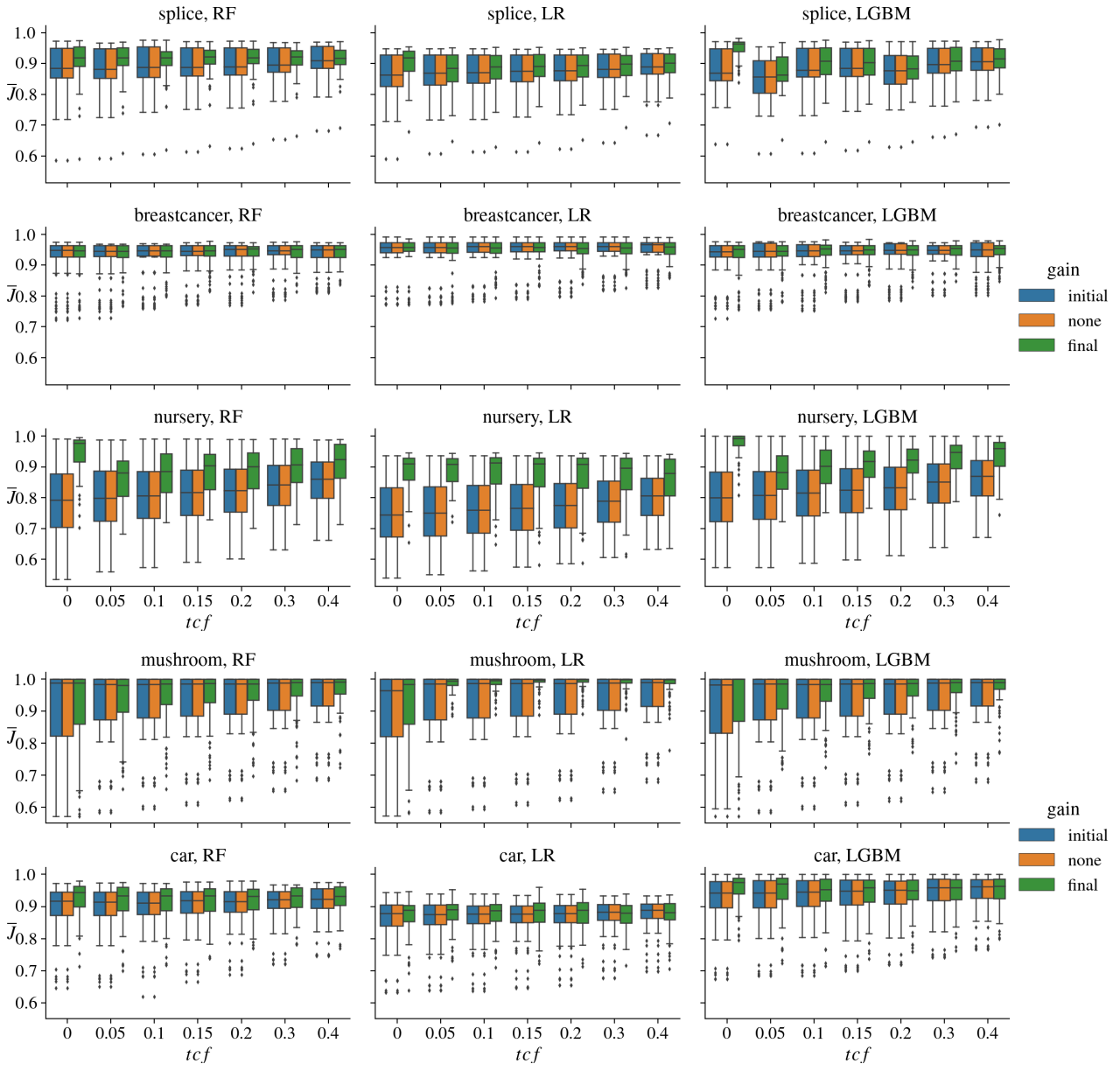

Figure 6. Additional plots for Figure 2 in the main paper. Experiments with models trained on the initial dataset before FROTE (*initial*), after applying the *none* strategy, and after FROTE completes augmentation (*final*). $mod - imp$ and $final - imp$ represent the differences in $\overline{J}$ between *mod* and *initial* and *final* and *mod*, respectively. The comparison is shown as a function of the training coverage fraction of the feedback rule sets and for different ML models and all the datasets. The *random* selection strategy is used. Standard box plot showing interquartile range (IQR) and whiskers showing 1.5 times IQR based on 30 random draws for each of $|\mathcal{F}| \in \{1, 3, 5\}$.

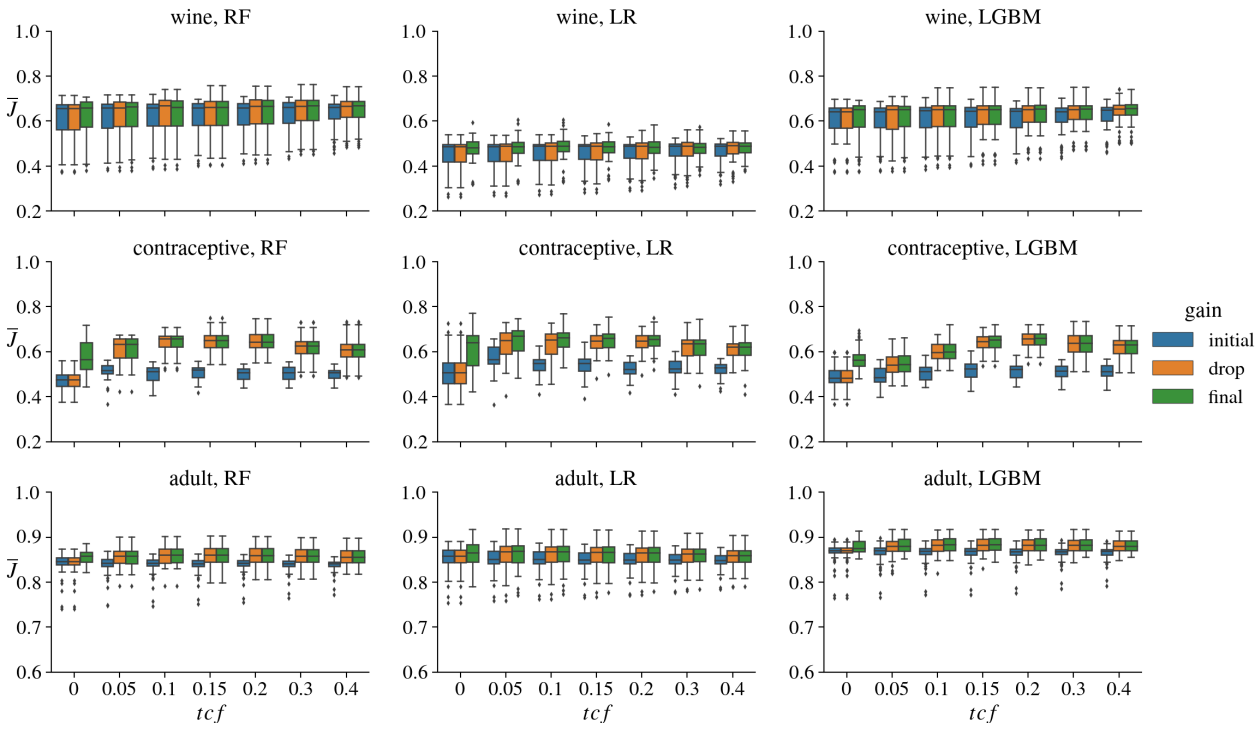

Figure 7. Similar setting with Figure 1 except results are presented for *drop* modification strategy.

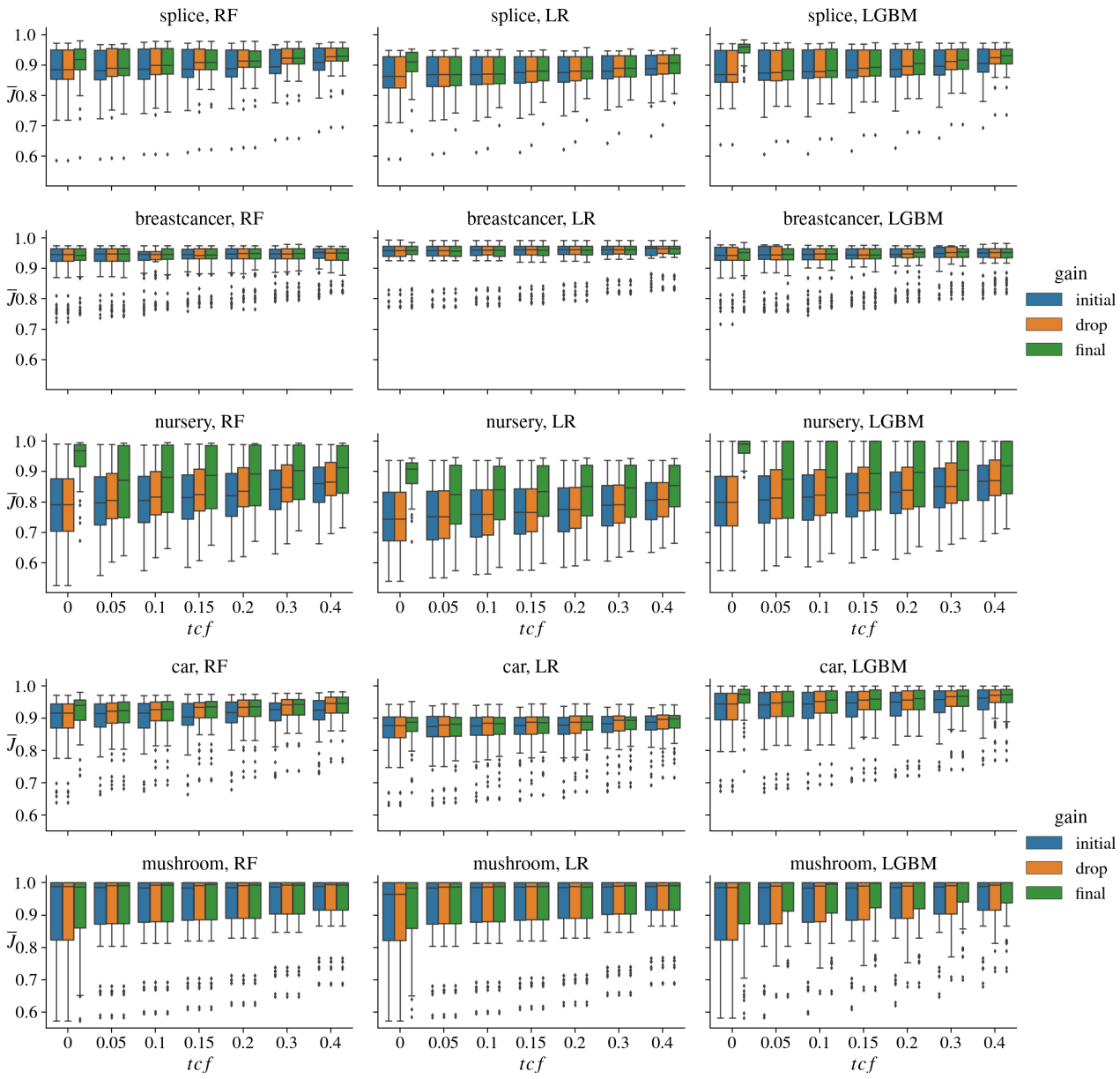

Figure 8. Similar setting with Figure 1 except results are presented for *drop* modification strategy.

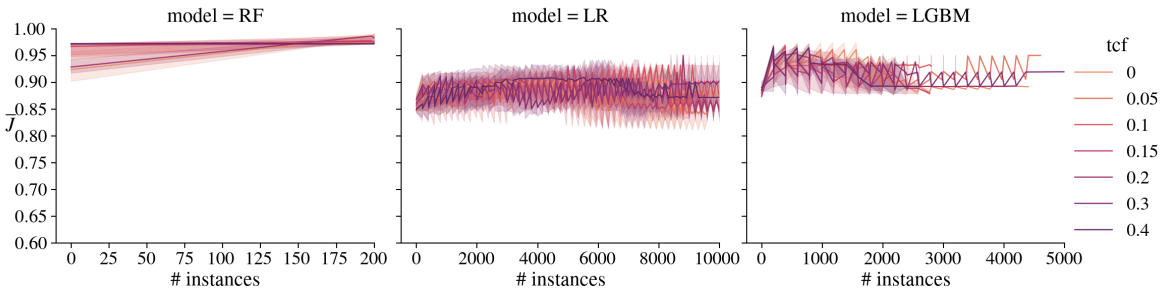

Figure 9. Augmentation progress evaluated on the held-out test set for different models and tcf values on the Adult dataset. The objective function $\overline{J}$ (median and 5-95 percentiles) is shown as a function of the number of instances added to the dataset during augmentation. Results are averaged over 90 runs, and for all runs, $|\mathcal{F}| = 3$, the mod-strategy is *relabel*, and *random* selection is used.

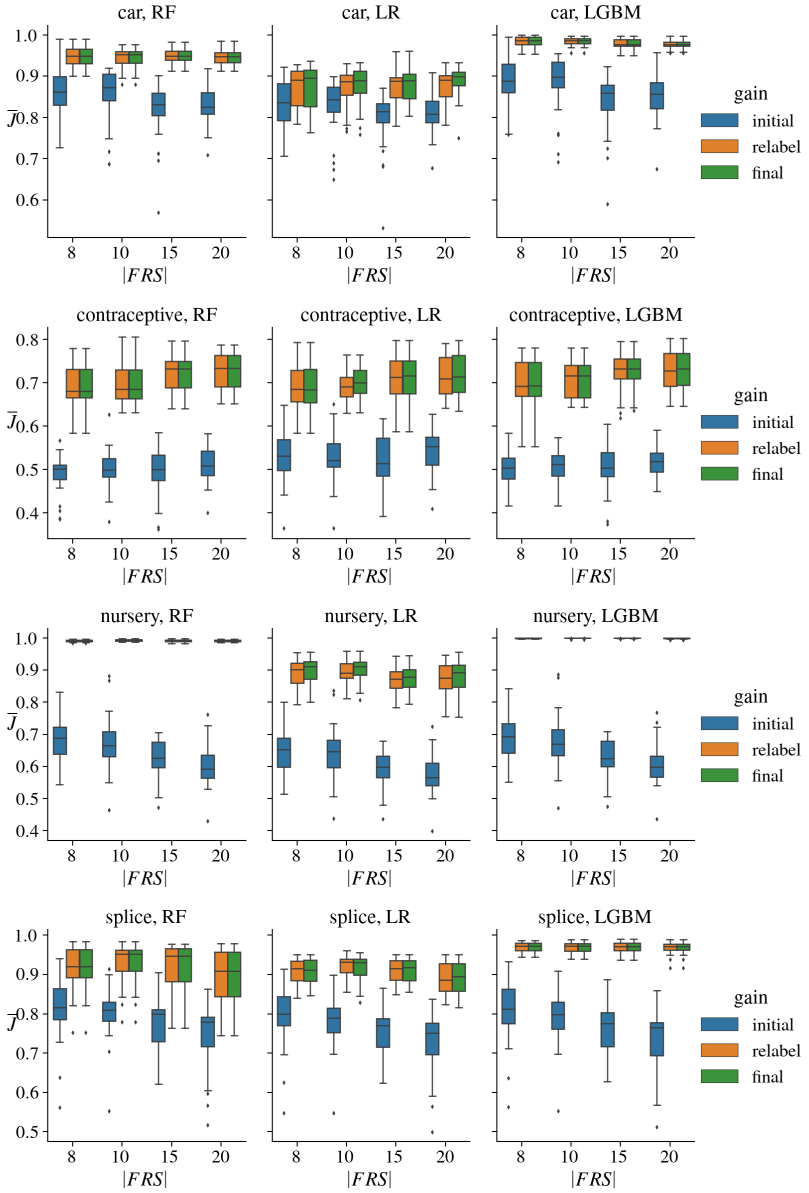

Figure 10. Additional plots for Figure 3 in the main paper. Effect of feedback rule set size for the Car, Contraceptive, Nursery and Splice datasets are given using the *random* selection strategy. The same comparison as in Figure 1 is shown between initial (before FROTE), after relabel, and final (after augmentation). Each box and whiskers is computed from 20 runs with $tcf = 0.2$, $\alpha = 0.8$, $k = 5$.