ABSTRACT
The importance of incorporating ethics and legal compliance into machine-assisted decision-making is broadly recognized. Further, several lines of recent work have argued that critical opportunities for improving data quality and representativeness, controlling for bias, and allowing humans to oversee and impact computational processes are missed if we do not consider the lifecycle stages upstream from model training and deployment. Yet, very little has been done to date to provide system-level support to data scientists who wish to develop responsible machine learning methods. We aim to fill this gap and present FairPrep, a design and evaluation framework for fairness-enhancing interventions, which helps data scientists follow best practices in ML experimentation. We identify shortcomings in existing empirical studies for analyzing fairness-enhancing interventions and show how FairPrep can be used to measure their impact. Our results suggest that the high variability of the outcomes of fairness-enhancing interventions observed in previous studies is often an artifact of a lack of hyperparameter tuning, and that the choice of a data cleaning method can impact the effectiveness of fairness-enhancing interventions.

1 INTRODUCTION
While the importance of incorporating responsibility — ethics and legal compliance — into machine-assisted decision-making is broadly recognized, much of current research in fairness, accountability, and transparency focuses on the last mile of data analysis — on model training and deployment. Several lines of recent work argue that critical opportunities for improving data quality and representativeness, controlling for bias, and allowing humans to oversee and influence the process are missed if we do not consider earlier lifecycle stages [5, 9, 10, 15]. Yet, very little has been done to date to provide system-level support for data scientists who wish to develop and evaluate responsible machine learning methods. In this paper we aim to fill this gap.

We build on the efforts of Friedler et al. [4] and Bellamy et al. [1], and develop a generalizable framework for evaluating fairness-enhancing interventions called FairPrep. FairPrep implements a modular data lifecycle, enables the re-use of existing implementations of fairness metrics and interventions, and the integration of custom feature transformations and data cleaning operations from real world use cases. Our framework currently focuses on data cleaning (including different methods for data imputation), and model selection and validation (including hyperparameter tuning), and can be extended to accommodate earlier lifecycle stages, such as data integration and curation.
of existing machine learning pipelines, and (3) are not designed to enforce best practices.

This paper makes the following contributions:

- We discuss shortcomings and violations of sound experimentation practices in existing empirical studies and software for analyzing fairness-enhancing interventions (Section 2).
- We propose FairPrep, a design and evaluation framework that promotes data to a first-class citizen in fairness-related studies (Section 3).
- We demonstrate how FairPrep can be applied to illustrate the impact of violations of best practices of ML experimentation, and how it enables the inclusion of incomplete data into studies, which is not supported by existing frameworks (Section 4).

In what follows, we briefly detail these contributions. Additional information is available in our technical report [14].

2 SHORTCOMINGS OF PREVIOUS WORK

We inspect the code bases of existing studies [4], and evaluation frameworks [1] for fairness-enhancing interventions, and identify a set of shortcomings and violations of best practices that potentially invalidate some of these studies’ findings.

Insufficient isolation of held-out test data. A major requirement for the evaluation of ML algorithms is to simulate real world scenarios as closely as possible. In the real world, we train our model (and select its hyperparameters) on observed data from the past. This model is later used to make predictions for unseen target data for which the ground truth is unknown. To emulate this real-world deployment scenario, we evaluate the trained model on a test set that was randomly sampled from observed historical data. It is crucial that this test set be completely isolated from the process of model selection, which, consequently, is only allowed to use the training data (the remaining, disjunct observed historical data). Unfortunately, we encountered violations of the test set isolation requirement in the existing benchmarking framework by Friedler et al. [4], bringing into question the reliability of reported study results. Further, we found that the architecture of the IBM AlF360 toolkit [1] does not support data isolation best practices for feature transformation.

Hyperparameter selection on the test set. Grid search for hyperparameters of fairness-enhancing models and interventions in [4] computes metrics for all hyperparameter candidates on the test set, and returns the candidate that gave the best performance. This strongly violates the isolation requirement. Instead, an evaluation procedure should maintain an additional validation set to select hyperparameters, and only evaluate prediction quality of the resulting single best hyperparameter candidate on the test set, to measure how well the model generalizes on unseen data.

Lack of hyperparameter tuning for baseline algorithms. We additionally found that the study by Friedler et al. [4] did not tune the hyperparameters of the baseline algorithms for which pre-processing and post-processing interventions are applied, even though they tuned the hyperparameters of the fairness interventions. This is problematic because there is no guarantee that the baseline algorithm will converge to a good solution with the default parameters. Friedler et al. [4] found a high variability of the fairness and accuracy outcomes with respect to different train/test splits, which could be an artifact of the described lack of hyperparameter optimisation.

Lack of feature scaling. We observed that both existing frameworks [1, 4] do not normalise the numeric features of the input data, but keep them on their original scale. While some ML models such as decision trees are insensitive to feature scaling, many other algorithm components, such has the L1 and L2 regularizers of linear models, implicitly rely on standardized features.

Removal of records with missing values. Another point of critique is that the study of Friedler et al. [4] ignored records with missing values (by removing them before running experiments), which means that the study’s findings do not necessarily generalize to data with quality issues. Thereby, existing frameworks are unable to investigate the effects of fairness enhancing interventions on records with missing values, which could be especially important for cases where a protected group has a higher likelihood of encountering missing values in their data [8].

---

1. https://github.com/algofairness/fairness-comparison/blob/4e7341929ba9cc98f7843773169c3d8284f0b046e29e0/fairness/algorithms/ParamGridSearch.py#L41
2. https://github.com/algofairness/fairness-comparison/tree/35fcb37cc7954668ee22e8ac5bf26f9689a3d8/fairness/algorithms/baseline
We now demonstrate how FairPrep can be used to showcase one of the shortcomings from Section 2, and how it enables experimentation on incomplete data. For all experiments, data is randomly split into 70% training, 10% validation, and 20% test.

**Impact of hyperparameter tuning on the accuracy and fairness metrics of logistic regression models (in combination with various preprocessing and postprocessing interventions) on the germancredit dataset.** Hyperparameter tuning (red dots) results in higher accuracy and reduced variance of the fairness outcome compared to no tuning (gray dots) in many cases.

### 3 FRAMEWORK DESIGN

The identified shortcomings motivate us to propose FairPrep, an evaluation and experimentation framework.

**Design principles.** We implement FairPrep on top of scikit-learn [11] and AIF360 [1], and design it based on two principles: (i) **Data isolation** — to avoid target leakage, user code should only interact with the training set, and never access the held-out test set. User code can train models or fit feature transformers on the training data, which will be applied by the framework to the test set later on. The framework should furthermore especially take care of data with quality problems. For example, it should allow experimenters to quantify the effects of their code on records with missing values by computing metrics and statistics separately for these records. (ii) **Explicit modeling of the data lifecycle** — the evaluation framework defines an explicit, standardized data lifecycle that applies a sequence of data transformations and model training in a particular, predefined order. Users influence and define the lifecycle by configuring and implementing particular components. At the same time, the framework should support users in applying best practices from ML experimentation.

**Data lifecycle.** Figure 1 illustrates the data lifecycle during the execution of a run of FairPrep: (i) **Model selection on the training set and validation set:** we train different models on the training data, where we apply the following consecutive steps: (i) resampling of training data (e.g., bootstrapping or balancing, optional); (ii) treatment of records with missing values (either removal or imputation); (iii) feature transformation (e.g., scaling of numeric values, one-hot encoding of categorical values); (iv) potential application of a preprocessing intervention; (v) model training using grid search; (vi) computation of predictions on the train and validation set; (vii) potential application of postprocessing intervention to predictions from train and validation set. (ii) **User-defined choice of the best model:** Users can choose between the explored models based on accuracy-related and fairness-related metrics computed on the validation set, trading these off as appropriate in their context. (iii) **Application of the best model on the test set:** The user-selected best model (and its corresponding data transformations) are finally applied to the test set, and the resulting accuracy of fairness are reported by FairPrep.

### 4 EXPERIMENTAL EVALUATION

We now demonstrate how FairPrep can be used to showcase one of the shortcomings from Section 2, and how it enables
We obtained similar results for the decision tree model and omit the corresponding plots due to lack of space.

Our results indicate that the high variability of the fairness and accuracy outcomes with respect to different train/test splits observed by Friedler et al. [4] may be an artifact of the lack of hyperparameter tuning of the baseline models in these studies.

Enabling the inclusion of incomplete data. Next, we showcase how FairPrep can be used to quantify the effect of including records with missing values into an experimental study. These records are commonly filtered out in other studies and toolkits, as discussed in Section 2.

We use the adult dataset\footnote{https://archive.ics.uci.edu/ml/datasets/adult} for this experiment, with a total of 32,561 instances and 14 attributes, including the sensitive attributes race and sex, and 2,399 instances with missing values. The task is to predict whether an individual earns more or less than $50,000 per year. Fairness evaluation is conducted between the privileged group of white individuals (85% of records) and the underprivileged group of non-white individuals (15% of records).

Of the 14 attributes, three have missing values — workclass, occupation, and native-country. Based on our analysis, missing values do not occur at random, as the records with missing values exhibit very different statistics than the complete records. For example, the positive class label (high income) is associated with 24% of the complete records, but with only 14% of the records with missing values. Additionally, married individuals are in the vast majority in the complete records, while the most frequent marital-status among the incomplete records is never-married.

Furthermore, the records with missing values from the privileged group are very different from the records with missing values from the underprivileged group. For example, the attribute native-country is missing four times more frequently for non-white individuals than for white individuals. Among the incomplete privileged records, 15% are associated with a high income, the second largest age group consists of 60 to 70 year-olds, and the majority of the individuals is married. For the incomplete records from the underprivileged group, however, only 10.6% have a high income, there are very few individuals over 60, and the majority of the individuals is unmarried.

We use logistic regression as the baseline learner, with hyperparameter tuning analogous to previous experiments. As before, we apply two fairness enhancing interventions that preprocess the data: ‘disparate impact remover’ [3] and ‘reweighing’ [7]. We use three strategies to treat missing values: (i) complete case analysis, removing incomplete records; (ii) retain all records and impute missing values with ‘mode imputation’\footnote{https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer} (replace a missing value with the most frequent value for that feature); (iii) retain all records and apply model-based imputation with datwig [2]. We execute 530 runs and report metrics computed on the held-out test set.

Results. We investigate the classification accuracy for complete and incomplete records, under imputation with mode and datwig. First, we observe that records with imputed values achieve high accuracy. This is a significant result, since these records could not have been classified at all before imputation! Interestingly, we observe higher accuracy for records with missing values compared to the complete records. Based on our understanding of the data, we attribute this to the higher fraction of (easier to classify) negative examples among the incomplete records. Further, we do not observe a significant difference in accuracy between mode imputation and datwig imputation. We attribute this to the highly skewed distribution of the attributes to impute — a favorable setting for mode imputation. Because datwig does no worse than mode, and is expected to perform better in general [2], we only present results for datwig-based imputation in the next, and final, experiment.

We compute the accuracy and disparate impact of complete case analysis (e.g., the removal of incomplete records) versus the inclusion of complete records with datwig imputation. We observe a minimally higher accuracy in the case of including complete records, but in general find no significant positive or negative impact on disparate impact. Taken together, the results paint an encouraging picture: Imputation allows us to classify records with missing values, and do so accurately, and it does not degrade performance, either in terms of accuracy or in terms of fairness, for the complete records.

5 CONCLUSION

We identified shortcomings in existing empirical studies on fairness-enhancing interventions. Subsequently, we presented the design of our evaluation framework FairPrep. This framework empowers data scientists to conduct experiments on fairness-enhancing interventions with low effort, and at the same time enforces machine learning best practices. We demonstrated how FairPrep can be used to measure the impact of a lack of hyperparameter tuning, and how it enables the inclusion of incomplete data. We aim to extend FairPrep by integrating additional fairness-enhancing interventions [13], datasets, preprocessing techniques, and feature transformations. Additionally, we intend to extend its scope to scenarios beyond binary classification, and introduce human-in-the-loop elements by providing visualisations and allowing end-users to control experiments with low effort.

REFERENCES