A Bio-Inspired Framework for Machine Bias Interpretation

First Author
first.author@afilliation.com
First University
City, Country

Second Author
second.author@afilliation.com
Second University
City, Country

Third Author
third.author@afilliation.com
Second University
City, Country

ABSTRACT

Machine learning algorithms use the past and the present to predict the future. But when given biased historical data, these algorithms can quickly become discriminatory. The area of machine learning fairness has emerged to detect and de-bias these algorithms, but has received widespread criticism for its one-size-fits-all approach, which allows certain cases of bias to slip through the cracks. In this study, we take a deeper look at the mechanisms by which machine learning algorithms develop harmful bias. We introduce a new method to interpret discriminatory systems, an Evolutionary algorithm for Feature Interaction (EFI), which we apply to several commonly used machine learning algorithms in two real-world problem instances: violent crime and median house price prediction. In the results, we discover several complex forms of bias, including the encoding of race through seemingly unrelated attributes and the direct use of race information as a proxy for neighborhood quality and conditions. Ultimately we suggest that more informative interpretation tools such as EFI can be used to not only explain machine learning outcomes, but supplement and improve existing machine bias detection approaches to provide a more robust and in-depth evaluation of machine learning algorithms.

CCS CONCEPTS

• **Computing methodologies** → Genetic algorithms.

KEYWORDS

Interpretability, fairness, machine bias, feature importance, feature interaction

ACM Reference Format:
First Author, Second Author, and Third Author. 2018. A Bio-Inspired Framework for Machine Bias Interpretation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Artificial intelligence (AI) and machine learning (ML) have revolutionized the modern world, but also raised a host of difficult questions. Whose interests do AI systems serve? How can we detect bias and unfairness in these systems? What can we do about it? This paper focuses on two such questions: 1) how can we understand the decisions made by AI systems, and 2) how can we ensure that AI systems are treating people fairly? Understanding AI systems is important both in its own right, and as a step toward evaluating fairness.

The predictive power of ML algorithms comes from their ability to internalize and represent features and their complex relationships. However, this same complexity poses a challenge when it comes to deciding questions of trust and accountability. Interpretable Machine Learning (iML) has the goal of making black-box ML systems understandable to human interpreters [13]. The challenge of fairness has received significant attention due to discoveries of models making skewed decisions with respect to attributes like race and gender [6]. ML algorithms can internalize the complex systems of inequality reflected in our data, and reinforce cycles of systemic oppression. There have been many advances in ML fairness, including statistical measures of model bias and automatic bias removal [16]. Although these approaches may be theoretically sound, many practitioners are unsatisfied with their black-box and one-size-fits-all nature [2], and suggest a crucial interplay between fairness and other machine learning objectives like interpretability, transparency, and explainability.

In this study, we develop a method to improve interpretability in complex ML models. We introduce an Evolutionary algorithm for Feature Interaction (EFI)[1], whose contributions to interpretability are twofold: 1) a bio-inspired approach to efficiently search for synergistic feature interactions and 2) a novel, permutation based method to approximate feature interaction strength with improved efficiency. In the following sections, we present EFI's methodology and apply our algorithm to both synthetic and real world example problems. We show that descriptive interpretability tools can supplement fairness metrics by providing an in depth understanding of complex ML biases. EFI could also be used as a robust method of selecting combinations of attributes to examine for intersectional fairness.

2 BACKGROUND

In this section, we introduce and motivate the methodology of our approach with an overview of the current state and challenges in ML interpretability and intersectional fairness.

2.1 Interpretability

According to the Predictive Descriptive Relevant (PDR) framework for iML development [13], interpretable models and interpretability tools should maintain a high degree of accuracy (predictive), comprehensively describe the model's predictive process (descriptive), and offer insights that are comprehensible to the human interpreter

[1] https://anonymous.4open.science/r/EFI

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

(relevant). Post-hoc "after the fact" tools offer a convenient solution, as they are applied in the post-processing stage of the machine learning design cycle so do not affect predictive accuracy. However, post-hoc tools have historically lacked descriptive accuracy, offering only rudimentary information regarding the relevance or *importance* of features and variables [3] [11]. While the relevance of features can be useful in noise removal and even narrowing the feature space relevant for interpretation, it tells us little about the model's predictive process.

2.2 Feature Interaction

In order to better understand the inner workings of a ML model, modern post-hoc interpretability tools focus on the interpretation of the predictive process itself. These tools make a distinction between the *main effect* of features and their *interaction effect* with other features. A feature interaction, defined as a learned relationship between two or more features is *strong* if the model gains more information from a set of features together than it gains from each individual feature apart. This interaction is *synergistic* if exactly all features are required to produce this effect [19]. Discovering the global structure of synergistic feature interactions is crucial in understanding the model's underlying predictive process.

Although modern interaction-based interpretability tools make the jump from single feature importance to multi-feature interaction [8] [14], they often fall short in terms of their scalability. Given an n-dimensional data set, there are approximately 2^n higher-order (greater than pairwise) feature interactions in a fitted model. Furthermore, a higher-order feature interaction is composed of exponentially many lower-order interactions, whose signal must filtered out of the result. Therefore, the estimation of a feature combination's interaction strength, weakness, synergy, or redundancy, let alone multiple, carries a large computational cost.

2.3 Fairness

The same problem of scalability also arises in considerations of ML fairness. In response to criticism that ML fairness metrics can give the false sense of having solved a socio-technical problem algorithmically, by patching just one axis of unfairness while ignoring others [10, 18], there has been a push to recognize Crenshaw's notion of 'intersectionality' [4] in treatments of ML fairness.[2]

However, problems arise in the implementation of intersectional approaches to fairness. As Kearns et al. (2018) write, "There are exponentially many ways of carving up a population into subgroups, and we cannot necessarily identify a small number of these a priori as the only ones we need to be concerned about." They define "fairness gerrymandering" as only looking for unfairness in a small number of pre-defined groups, and show that avoiding gerrymandering is computationally hard in the worst case [12].

In addition to the scalability problem is the problem of data scarcity at many of the intersections of minority groups [20]. While we do not directly study fairness here, our interpretability tools offer a promising method of efficiently choosing the most relevant feature interactions in a classifier, which could be adopted as a solution to fairness gerrymandering in intersectional fairness work.

[2]This response is at best partial, addressing only one of Hoffmann's three criticisms, and skirting all 5 of the traps identified by Selbst et al.

3 METHODOLOGY

In this section, we introduce EFI, a novel, bio-inspired approach to synergistic feature interaction discovery. EFI applies a powerful search algorithm designed to efficiently explore the space of higher-order feature interactions (Fig. 1). When the search is complete, we consolidate the results and provide a thorough interpretation of the model's predictive process.

In the following sections, we discuss the technical details of EFI, beginning with a formal definition of the inputs, the bio-inspired search for synergistic feature interactions, and proposed measures to interpret the search results.

3.1 Data Preprocessing and Model Training

The input to our algorithm is a data set whose rows correspond to instances or examples and columns correspond to features or variables. One outcome is designated the target and is the value that we would like to predict. In a house price prediction data set, the instances might be houses that have sold in the past, the target is the price of the house, and the features are variables like location, number of bedrooms, etc. The goal in house price prediction is to learn the relationships between these features and the target variable such that given a house for sale, we can accurately predict its sale price.

We split the data set into training and testing sets using *scikit-learn*'s stratified sampling, which balances the target distribution across splits [15]. Next, we initialize three commonly used ML models – Random Forest (RF), Multi-Layer Perceptron (MLP), and Gradient Boosting (GBDT) – and fit them to the training set. We select these models due to their high capacity to learn complex, higher-order, non-linear feature relationships. Due to their high capacity, their default configurations provide sufficient predictive accuracy.

3.2 Multi-Objective Genetic Algorithm

Given the fitted model and an unseen testing set, EFI applies a Multi Objective Genetic Algorithm (MOGA) to search for synergistic feature interactions. A MOGA is a population-based search and optimization technique that draws from the natural process of biological evolution [7].

Given an optimization problem, a set of objectives to direct the search, and a fitness function to measure the quality of solutions, a MOGA initializes a random population of candidate solutions, which are evolved over many generations. A single generation begins with parent selection, where the current fittest candidate solutions are chosen to recombine. In the recombination stage, parents share certain aspects of their solutions, in order to create new, possibly novel offspring. In the next step, the offspring are mutated and added to the population, and the fittest overall solutions are selected to survive into the next generation. When the average population fitness begins to converge, the algorithm is terminated and the population of high-fitness candidate solutions is returned.

In our MOGA (Fig. 1b), the objectives are to search for the smallest feature combinations with the strongest feature interaction. Under the objective of minimizing the size of features combination we incentivize the population to prioritize synergistic feature interactions containing exactly the features required to produce a

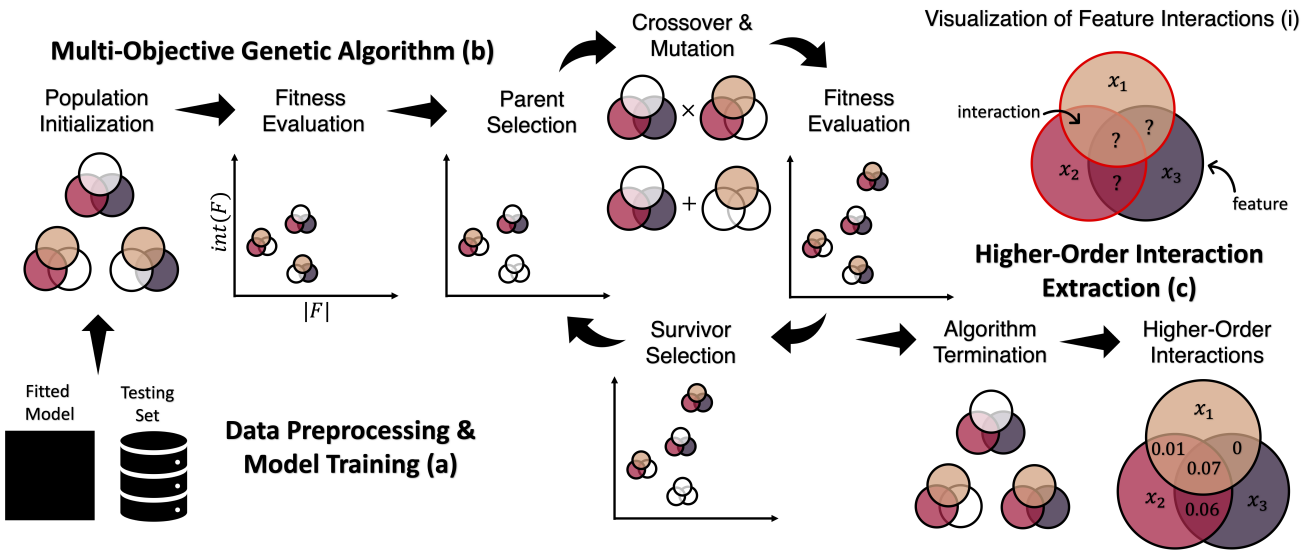

Figure 1: Given a fitted model and unseen testing testing set (a), EFI employs a Multi-Objective Genetic Algorithm (b) to search for feature combinations containing synergistic higher-order feature interactions (i). After multiple rounds, search results are consolidated and analyzed to describe the most synergistic higher-order feature interactions present in the model (c).

strong interaction effect. At the beginning of evolution, a population of 200 randomly generated feature combinations is initialized. In order to select feature combinations for crossover and survival that optimize these objectives, we apply the *fast-nondominated-sort* selection algorithm, proposed in the Non-dominated Sorting Genetic Algorithm II (NSGA-II) [5]. *Dominance*, allows for the sorting of candidate solutions in terms of multiple objectives.

Definition 3.1 (Dominance)**.** Given a pair of feature combinations (F_i, F_j) with attributes *int* (interaction) and $|F|$ (size), F_i *dominates* F_j iff.

$$|F_i| < |F_j| \ \ \& \ \ int(F_i) > int(F_j) \tag{1}$$

The attribute *int* estimates the strength of the interaction between the features. Fast-nondominated-sort divides the population into fronts, where the first (Pareto) front is the set of feature combinations that are not dominated in terms of either search objective (Fig. 2). The second front is the set of feature combinations that are not dominated by any other feature combination not in the first front, and so on. Once the population has been sorted, entire fronts are selected until selecting the next front would yield too many feature combinations. At this stage, fast-nondominated-sort selects the feature combinations from the current front with the largest crowding-distance, defined as the Euclidean distance to its nearest neighbor on the front (Fig. 2). This step prioritizes the selection of diverse feature combinations from the final selection front, as feature combinations with a higher crowding-distance are more likely to be unique [5].

The maximum crowding-distance on the Pareto front is also used to detect population stagnation, which determines when to terminate evolution. According to [17], the stagnation of crowding distance in the Pareto front is a stable measure of population convergence in MOGAs, as it indicates that the Pareto front is maintaining both a fixed size and structure.

3.2.1 Fitness Function. Given a candidate feature combination, we estimate its interaction strength using a novel permutation-based method inspired by an approach proposed by Oh [14]. Permutation-based methods such as permutation feature importance [8] typically measure the importance of a feature as the error produced by shuffling its values in the testing set. If the feature is relevant to the model, then corrupting its values should produce a positive change in error. Proposed as a simple alternative to partial dependence-based feature interaction measures such as Friedman's H-statistic and partial dependence plots [8], Oh's approach is defined as follows:

Definition 3.2 (Oh's Interaction Measure)**.** Given a predictive model and a feature combination $\{x_1, x_2\}$, Oh's approach estimates the interaction as the difference in error from shuffling the features together and the summed effect of shuffling individual features apart.

$$Oh(F) = (err(x_1) + err(x_2)) - err(\{x_1, x_2\}) \tag{2}$$

If there exists an interaction between the features, the associated error will be reflected multiple times in the sum and only once in the negation. While Oh's measure provides a simple and efficient alternative to partial-dependence based approaches, it presents a potential theoretical problem: permuting a feature combination has a potentially propagating effect. Consider the scenario posed in Fig. 1i, which depicts a 3-feature interaction where the colored circles represent individual features and their shared and labeled

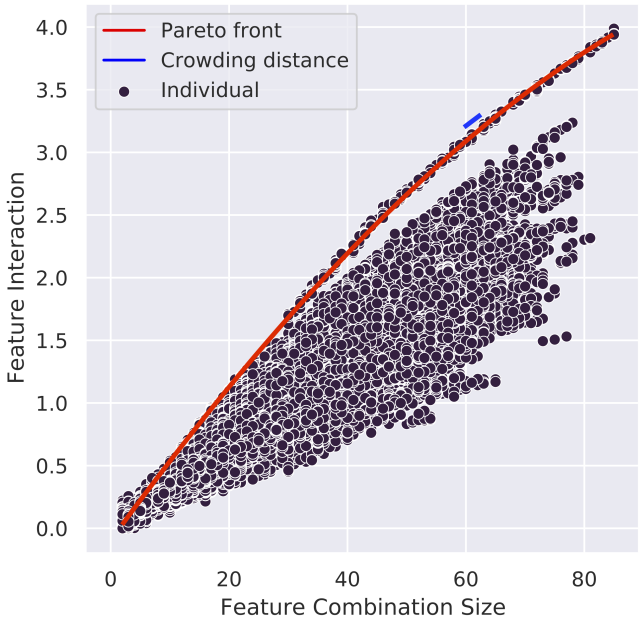

Figure 2: Visualization of the combined population from 1000 runs of the MOGA on the MLP, the Pareto front of individuals, and an example of the crowding distance between two individuals on the same front.

regions the interaction between them. If we permute the features x_2 and x_3, depicted as outlining these feature regions with a red border, it will have the effect of breaking all interactions that these features are involved in. In this case, x_2 and x_3 are involved in all interactions, including the pairwise interactions between x_1 and x_2 as well between x_2 and x_3. Thus, the interaction effects of feature combinations other than x_2 and x_3 will contribute to the measured result.

In order to account for this problem, we propose an alternative feature interaction measure. Instead of permuting a set of features and measuring the error that results, we *don't* permute a set of features and measure the amount of accuracy that remains. Due to its inverse relationship to Oh's permutation based method, we coin this method *Not-Perm*.

Definition 3.3 (Not-Perm Interaction)**.** Given a predictive model and feature combination F, the interaction strength is estimated as the difference in remaining accuracy from not shuffling the features compared to the summed effect of not shuffling individual features apart.

$$int(F) = \frac{1}{K}\sum_{k=1}^{K} acc(F) - \left(\sum_{x_i \in F} acc(x_i)\right) \tag{3}$$

Note that this process is repeated $K = 10$ times, as some permutations will be more or less effective in removing the signal from outside of the feature combination.

It's important to note that the feature interaction strength measured by our evaluation function *int* alone does not necessarily quantify pure synergy in the mathematical sense, where the strength of the embedded lower order interactions are removed. However, in the context of our MOGA, redundant interactions are reduced by the search objective of minimizing feature subset size.

3.3 Higher-Order Interaction Extraction

Because the MOGA is an inherently stochastic process, the population of a single run is not guaranteed to converge to the set of optimal feature combinations. In order to increase the likelihood that synergistic feature combinations are discovered, the MOGA is run 1000 times in parallel, resulting in a set P of approximately 200,000 high-performing feature combinations.

We propose several equations to estimate the interaction effect of an individual feature x_i, the interaction contained within feature combinations F, and the dependency between feature pairs (x_i, x_j).

Definition 3.4 (Feature-Wise Interaction Effect)**.** A feature x_i's contribution to feature interactions is estimated as the weighted occurrence of x_i in high-quality feature combinations F_i of P

$$occ(x_i) = \frac{\sum_{F_i \in P} \frac{int(F_i)}{|F_i|}}{|P|} \tag{4}$$

This metric operates under the assumption that given a feature combination with an associated feature interaction score, each feature contributes equally in the interaction. Thus, a feature receives a high interaction score if it appears many times in high-quality combinations.

Definition 3.5 (Higher-Order Feature Interaction Strength)**.** A feature combination F_i's overall interaction strength is estimated as the weighted occurrence of F_i as a subset of high-quality feature combinations in P.

$$occ(F_i) = \frac{\sum_{F_i \in P} \frac{int(F_i)}{2^{|F_i|}-|F_i|-1}}{|P|} \tag{5}$$

Note the difference between Definition 3.4 and 3.5. Because we now consider feature combinations as opposed to individual features, we make the assumption that each lower-order interaction contributes equally to the resulting interaction score, and divide the interaction score by the number of possible lower-order interactions. Similarly, a feature combination receives a high interaction score for appearing many times in high quality feature combinations.

Definition 3.6 (Feature Dependency)**.** The dependency of x_i on x_j is estimated as the weighted occurrence of x_i in feature combinations F_i relative to the weighted occurrence of x_i in feature combinations F_{ij} where x_j also occurs.

$$dep(x_i, x_j) = \frac{\sum_{F_i \in P} \frac{int(F)}{2^{|F_i|}-|F_i|-1}}{\sum_{F_{ij} \in P} \frac{int(F_{ij})}{2^{|F_{ij}|}-|F_{ij}|-1}} \tag{6}$$

Despite its one way nature, this measure does not imply, or attempt to imply causality. A large dependency from one feature to another suggests that the depended on feature occurs *whenever* the depending on feature occurs. From this we can infer that the depended on feature plays a supporting role in the interaction between these features.

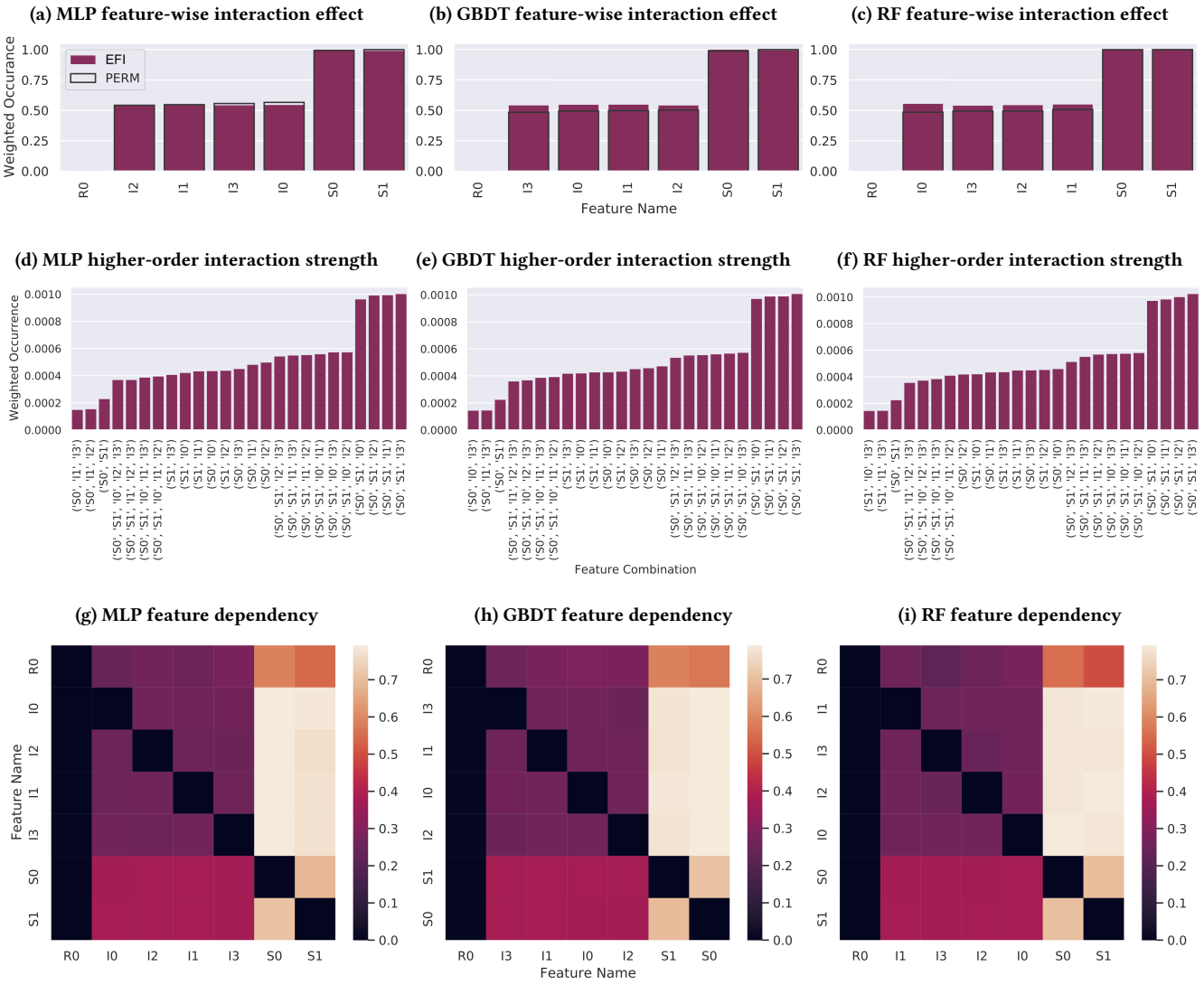

Figure 3: EFI's analysis of the MLP, GBDT and RF fitted to the synthetic data describing the interactivity (a-c) of individual features, the interaction strength (d-f) of specific feature combinations and the pairwise dependency between features (g-i).

4 RESULTS AND DISCUSSION

In this section, we introduce one synthetic problem instance, and re-examine two well-studied real-world problem instances to demonstrate the efficacy and applicability of our approach.

4.1 6-bit Multiplexor

In order to provide a concrete example of the concept of synergistic higher-order feature interaction and how EFI discovers this phenomenon, we begin with a simple synthetic problem instance where the ground truth higher-order interactions are known: the 6-bit multiplexer function. Given a 4-bit input string, 2 selector bits and 1 random bit, the multiplexer function maps the selector bit input to an integer, and returns the value from the input string at that index.

In this function, there exist 4 separate 3-way interactions between the two selector variables and each input variable. These interactions are *pure* in the mathematical sense, as exactly all 3 features are required to produce an accurate prediction. If a predictive model $\hat{f}$ correctly learns the multiplexor function, these ground-truth interactions should be present in the model.

In order to estimate the descriptive accuracy of EFI in discovering these higher-order feature interactions, we fit Gradient Boosting (GBDT), Random Forest (RF), and Multi-Layer Perceptron (MLP) classification models to a set comprised of all possible binary strings of length 7, each labeled with the correct output of the function. Note that we fit the models to the entire data distribution to allow

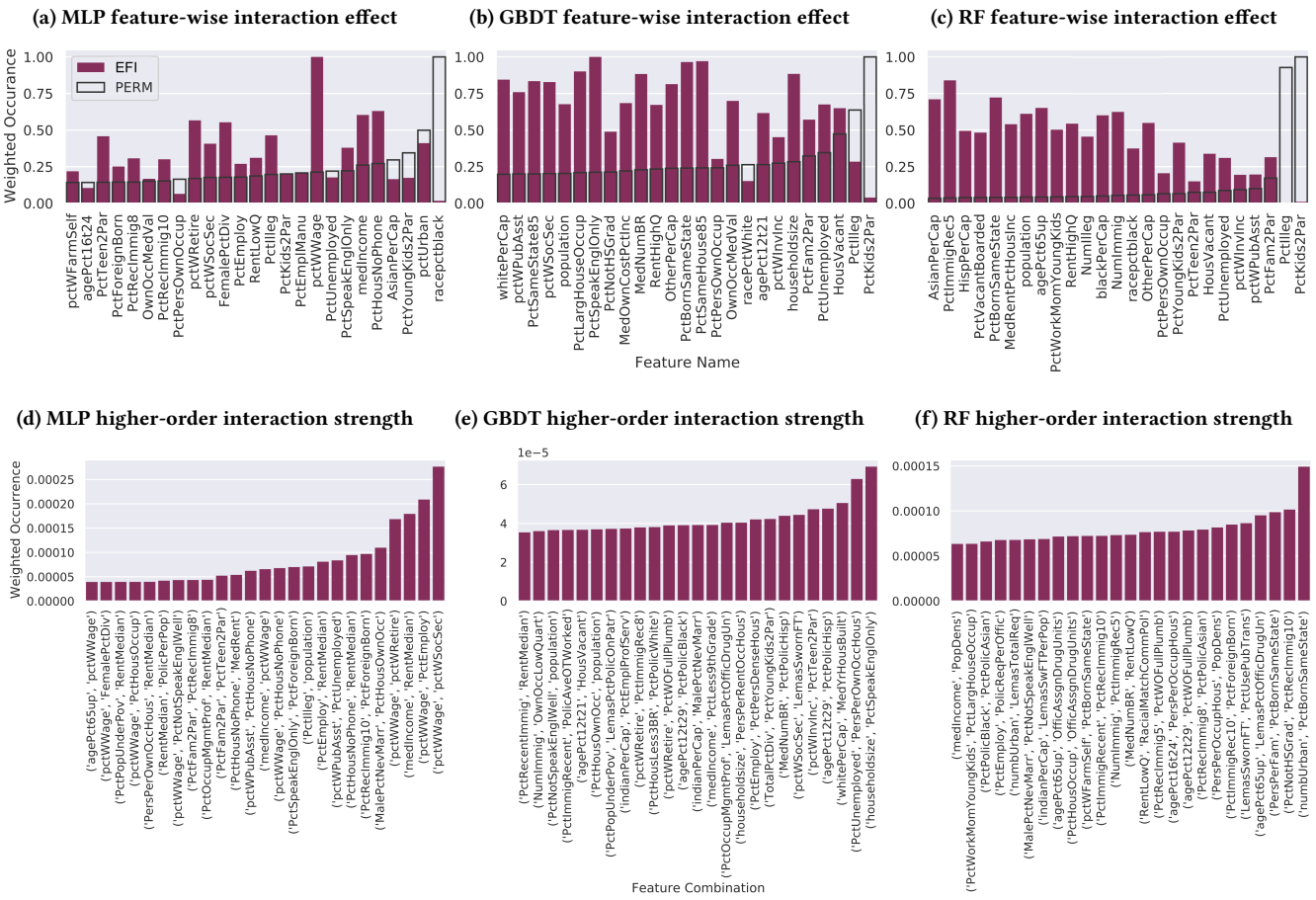

Figure 4: EFI's analysis of the MLP, GBDT and RF fitted to the LEMAS data set, describing the interactivity (a-c) of individual features and the interaction strength (d-f) of specific feature combinations.

them to accurately learn the multiplexor function. For each of these models, EFI is run 1000 times in parallel and the resulting populations are aggregated and analyzed.

The first relevant outcome is that EFI's feature-wise interaction effect closely matches classic permutation-based importance (PERM), which measures both the main and interaction effect of features (Fig. 3a-c). This match is a correct result, as features have no *direct* relationship to the output of the multiplexor function only an indirect relationship through interactions. This result shows that Not-Perm is successful in removing the main effect of features.

When examining the feature interactions with the highest strength (Fig. 3d-f), the 4 ground-truth interactions between the two selector variables and each individual input variable emerge with highest estimated interaction effect. The next most frequently occurring interactions are not pure in the mathematical sense, as they actually contain two separate interactions: one between the selector variables and one input variable and another between the selectors and another input variable. This is another encouraging result, as it shows that the multi-objective search mechanism, which prioritizes smaller, more synergistic feature combinations, was effective in prioritizing pure synergistic interactions.

The dependencies discovered by EFI (Fig. 3g-i) also provide useful insight. In these models, EFI discovered a strong dependency of input variables on selector variables, as shown by the beige regions of the heat map. This dependencies tell us that the input variables typically occur in feature combinations where selector variables occur. This is a correct result, as the values of selector variables are *always* required to make accurate predictions. This dependency is *asymmetric*, as the selector variables do not depend as heavily on input variables. In contrast, the selector variables have a strong co-dependency (symmetric), as they work together to index the correct input variable.

EFI's interpretation of these models closely mirrors the inner dynamics of the multiplexor function. This result provides evidence for EFI's descriptive accuracy in discovering higher-order synergistic interactions in synthetic problem instances. In the next two sections, we explore two real-world problem instances in order to show EFI's efficacy in machine bias detection and explanation problems.

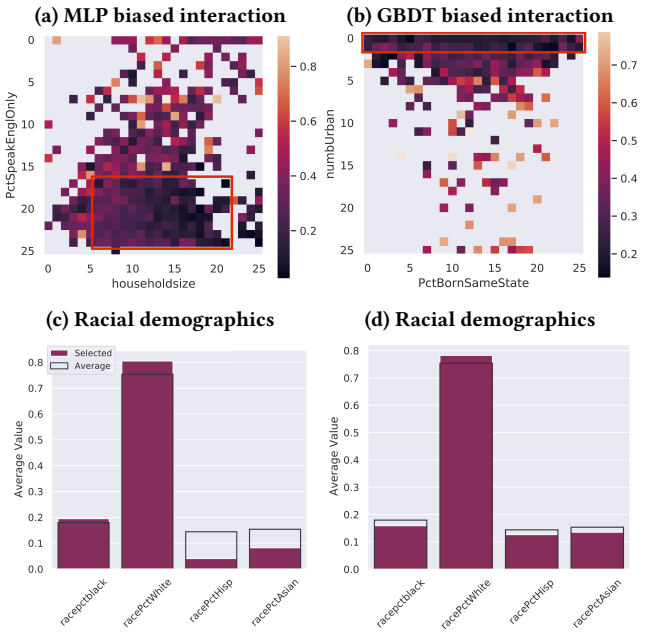

Figure 5: Further analysis of the potentially biased feature interactions learned by the MLP (a) and GBDT (b) showing that neighborhoods in the bounded regions with low predicted violent crime are disproportionally white (c-d).

4.2 Violent Crime Prediction

Machine crime prediction is one of the most studied problems in ML where algorithmic bias has been found to reinforce and exacerbate existing systemic inequalities [1, 6]. The goals of this section and the next are to demonstrate how EFI picks out and explains the mechanisms through which ML models internalize bias. We focus on popular real-world data sets in order to show how EFI performs on well understood data. The first of these data sets is the 1980 Law Enforcement Management Administrative Services (LEMAS) data set[3], which encodes per-capita violent crime rates[4] along with 122 demographic statistics from 1,994 US neighborhoods.

For this problem instance, we repeat the process of 1000 independent runs of the MOGA, population aggregation, and feature analysis with the same three models used previously, then examine the results. In the next sections, we focus on interactions that are relevant to the machine bias problem and discuss their possible implications.

In the MLP model, EFI discovers a large interaction effect for various income-related features, such as "pctWWage: the percentage of households with wage or salary income in 1989", "medIncome: the median household income", "pctWRetire: the percentage of households with retirement income", and "rentMedian: the rental housing median rent" (Fig. 4a). Of the possible interactions between these features, the feature combination with the highest score is the pairwise interaction between "pctWWage" and "pctWSocSec: the percentage percentage of households with social security income" (Fig. 4d). That these closely related features are related both to one another and to crime rates is unsurprising.

A comparison between EFI's feature-wise interaction effect and classic permutation-based feature importance (main and interaction effect) (Fig. 4a) shows that the feature "racepctblack: the percentage of population that is african american" and "pctUrban: the percentage of people living in areas classified as urban" have high importance in the model, but few interactions with other variables. This suggests that the model has learned direct relationships between rates of reported violent crimes, and the proportion of Black residents in a neighbourhood, or whether a neighbourhood is urban. Again, these are unsurprising relationships given what is known about this data set. This result demonstrates how EFI can identify cases of straightforwardly discriminatory decision-making, should a model learned on this data set be used to build a predictive policing tool.

In the GBDT model, fewer significant interactions appear. The most frequently occurring feature combination "householdsize: the mean people per household" and "PctSpeakEnglOnly: the percent of people who speak only English" (Fig. 4e) has a weighted occurrence of nearly an order of magnitude less than the largest weighted occurrence in the MLP model's results. This result can be attributed to the fact that the two most important features (Fig. 4b), "pctIlleg: the percentage of kids born to never married" and "PctKids2Par: the percentage of kids in family housing with two parents" have a very strong main effect. Although the interaction ("householdsize", "PctSpeakEnglOnly") has a relatively low score, a closer analysis provides valuable insight.

In order to better understand this feature interaction, we discretize the distribution between these two feature into a heat map, where the color of each cell represents the average predicted value of violent crime in neighborhoods of that bin (Fig. 5a). The transparency of each cell represents the number of neighborhoods in each bin, to indicate which bins are the most relevant for analysis. The simplified distribution shows that there is a highly populated region of roughly average "householdsize" and large "PctSpeakEnglOnly" with generally low predicted violent crime. We select the data points that fall within this region, and plot the average racial demographics of these neighborhoods (Fig. 5c). The selected region is nearly 80% white, and has lower racial diversity that other neighborhoods. The model has thus discovered an interaction that acts as a proxy for the whiteness of a neighbourhood. This demonstrates how EFI can identify cases where decision-making in ML models depends on proxies for protected classes like race.

In the RF model, a similar phenomenon occurs, where few significant interactions prevail due to the strong main effect of the two most important features: "PctIlleg" and "PctKids2Par" (Fig. 4). The importance of these features in the RF are presumably due to the same mechanisms which identified them in the GBDT model, which maintains a similar tree-based structure. Although the most frequently occurring feature combination ("numbUrban", "PctBornSameState") has relatively low weighted occurrence, it also has an interesting explanation. Another heat map is defined (Fig. 5b) which shows a highly populated region of low "numUrban" and

[3]LEMAS is Available on the UCI Machine Learning Repository http://archive.ics.uci.edu/ml/datasets/communities+and+crime.

[4]The data set description does not go into detail about what gets counted as a violent crime, beyond saying that it includes numbers for "murder, rape, robbery, and assault". Based on the information given at https://bjs.ojp.gov/topics/crime, we infer that this includes incidents "reported to and recorded by police."

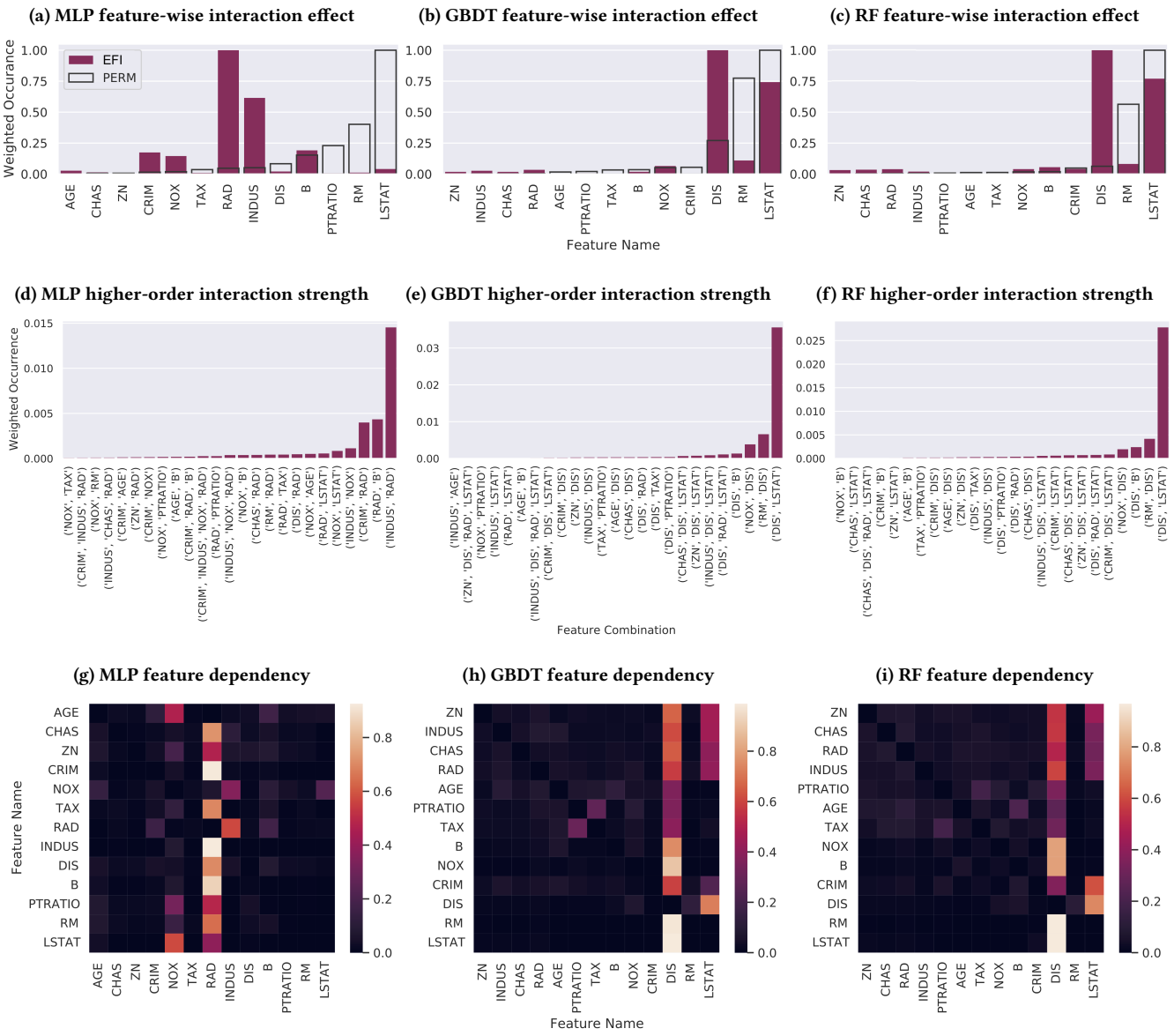

Figure 6: EFI's analysis of the MLP, GBDT and RF fitted to the Boston Housing data set, describing the interactivity (a-c) of individual features, the interaction strength (d-f) of feature combinations and the pairwise dependency between features (g-i).

average to large "PctBornSameState" where predicted violent crime is very low. A similar analysis of the racial demographic in this region reveals that these neighborhoods are again nearly 75% white (Fig. 5d).

This result shows a strong similarity between the GBDT and RF models. In addition to the direct learned relationships between "PctIlleg" and "PctKids2Par" with violent crime, the model has learned an association between their interaction and crime rate. In both interactions, ("householdsize", "PctSpeakEnglOnly") and ("numbUrban", "PctBornSameState"), the models has learned a region of neighborhoods in the corresponding feature spaces that is predominantly white. Once again, a proxy for race is being used by the model, which would correspond to discriminatory decision-making if such a model were used in predictive policing.

Looping back to the theory of feature interaction, these combinations provide an excellent example of feature synergy, where additional information is implicitly contained in a learned relationship. In this case, the models learned two pairwise relationships that incorporate complex demographic information.

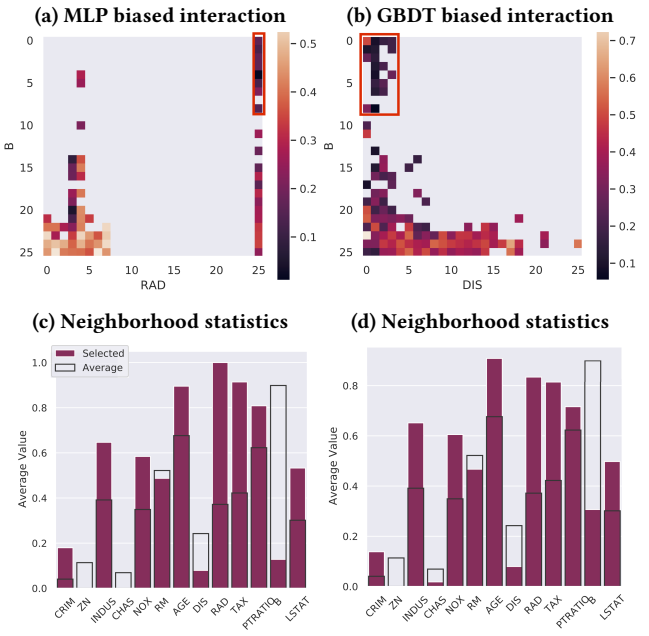

Figure 7: Further analysis of the potentially biased feature interactions learned by the MLP (a) and GBDT (b) showing that neighborhoods in the bounded regions that have a low distance (DIS) or high accessibility (RAD) to urban centers and a low value of "B" have a disproportionately low predicted median house price and generally worse neighborhood statistics (c-d)

4.3 Median House-price Prediction

Median house-price prediction is another widely studied machine bias problem. In order to examine the types of biases learned by ML models in this setting, we use the controversial Boston Housing data set.[5] Originally proposed in 1978 in an analysis of air quality in urban neighborhoods [9] the Boston Housing data set provides the median house price and 13 demographic statistics of 506 Boston neighborhoods drawn from the 1970 US Census.

Part of the controversy surrounding this data set concerns the feature "B". The Boston Housing data set originally contained a feature "Bk: the proportion of blacks by town". This feature was removed and engineered into the feature "B: $1000(Bk - 0.63)^2$", which corresponds to the variance of "Bk" from 63%. It is not entirely clear why the data was so manipulated, nor why 63% was chosen, but a possible answer is floated by Carlisle in a 2019 Medium post[6]. Carlisle notes that Harrison and Rubinfeld [9] hypothesize a parabolic relationship between Bk and house prices, where white flight pushes up prices in predominantly white neighbourhoods, and housing discrimination pushes up prices in predominantly Black neighbourhoods.[7] A low value of "B" represents a neighborhood with a proportion of Black residents close to 63%, while a high value corresponds to a neighborhood with either far more or far fewer than 64% Black residents. In other words, racially segregated neighbourhoods would have high "B" values, though whether a high "B" neighbourhood is predominantly Black or white is not apparent from the data.

In the MLP model, the feature pair "INDUS: the proportion of non-retail business acres per town" and "RAD: an index of accessibility to radial highways" emerges as the most synergistic feature interaction (Fig. 6d), having the first and second highest estimated interaction effect (Fig. 6a). In the dependency structure (Fig. 6), "RAD" is depended on significantly by almost every other feature, suggesting that this feature plays a supporting role in many interactions. "RAD" also depends heavily on "INDUS," suggesting that "RAD" and "INDUS" have a strong two-way interaction, and explains this combination's prevalence in the results. In addition, EFI discovers a relatively strong interaction between the feature "B" and the feature "RAD," occurring with the third most weighted frequency (Fig. 6d). In the distribution defined by this feature subspace (Fig. 7a), neighborhoods with high accessibility to radial highways but a low value of "B" receive a very low predicted median house prices. In the selected neighborhoods, the average feature values show worse than average neighborhood conditions (Fig. 7c).

In the RF and GBDT models, EFI discovers a high feature-wise interaction effect for "LSTAT: the percent lower status of the population," "RM: the average number of rooms per dwelling", and "DIS: the weighted distances to five Boston employment centres" (Fig. 6b-c). These features have both strong importance and estimated interaction effects, suggesting that feature interaction is central to these models' predictions. EFI also discovers a relatively significant interaction containing the feature "B" with the feature "DIS" (Fig. 6e-f). In the interaction between "B" and "DIS" the heat map (Fig. 7b) shows that low values of "DIS" and "B" are associated with lower predicted median house price. When the neighborhoods of this region are selected and their average neighborhood statistics calculated (Fig. 7d), a similar result emerges. These neighborhoods, marked with strong proximity to economic centers and low "B", are associated with generally worse neighborhood conditions.

In all three models, the engineered feature "B" occurs in relevant interactions, and is used to encode worse than usual neighborhood conditions. The models have learned that in some contexts, a high value of "B" is associated with higher housing prices, the use of these models in a real world setting (though it is hard to imagine this data set being used to train a contemporary decision-making system) could encourage racial segregation by assigning higher housing prices to more segregated neighborhoods.

Some of the complex interactions between features in these models that turned out to be important would not have been predictable in advance. This method efficiently finds the most meaningful combinations of features that make a difference to the model's classifications, suggesting a solution to the problem of deciding which intersections of features are the relevant ones to investigate when exploring intersectional fairness.

[5]The Boston Housing data set is in the *scikit-learn* library of bench-marking problems https://scikit-learn.org/stable/datasets/toy_dataset.html#boston-dataset

[6]https://medium.com/@docintangible/racist-data-destruction-113e3eff54a8

[7]Carlisle offers evidence that the data in the "B" column may also be inaccurate, and that the hypothesized parabolic relationship might not exist, making "B" a very troublesome feature.

5 CONCLUSION

In this study, we explore the application of interpretability in understanding the inner mechanisms of machine bias, or how systemic inequalities in our data are transferred into our predictive models. In order to achieve this, we introduce a novel, bio-inspired approach to post-hoc interpretability, EFI, that addresses several key limitations in the iML area, namely the trade-off between the descriptive accuracy and computational practicality of feature interaction based approaches, as well as the computational explosion faced by standard approaches to intersectional fairness. We address these challenges by leveraging the exploratory power of a multi-objective genetic algorithm, which we show is capable of efficiently searching the complex and exponential space of higher-order feature interactions in order to determine which ones are the most synergistic and relevant to the model.

In our experimental results, we provide through a combination of synthetic and real-world problem instances a diverse set of findings. In the context of sensitive ML applications such as crime and house price prediction, we show that harmful machine bias can materialize in several of the following forms. In one form, ML models can rely directly on protected class information, for example the MLP model using the value of "racepctblack", and that value alone, to predict violent crime. Alternatively, ML models can learn complex feature interactions that implicitly correspond to sensitive information. The latter result motivates the use of more descriptive post-hoc interpretability tools, such as measures of higher-order feature interaction, for the application of probing and regulating ML models in sensitive problem instances. The benefit of more descriptive probes is not limited to the ability to hold our ML algorithms more accountable. We believe that the probing of such ML algorithms with the goal of understanding the inner mechanisms of machine bias is crucial in the development of ML fairness, where a nuanced understanding of not only the data, but also the problem, the model, and a complex set of fairness objectives is required to produce an ML application that leads to more good than harm.

Although our methodology and results make several key contributions, they also open up several new avenues for future research, namely the MOGA as a useful tool to explore the space of higher-order feature interactions, and to address the computational complexity of investigating intersectional fairness. Although the genetic algorithm is relatively robust, other search mechanisms, such as Bayesian optimization could also be explored.

One limitation of our approach is the variance of Not-Perm, the permutation-based method proposed to measure the interaction strength of a candidate feature combination. The variance of this approach comes from the random nature of feature permutation. When many features are permuted at once, some permutations remove more or less signal from the data than others. This causes variance in the output of Not-Perm, which causes candidate feature combinations to sometimes receive unrepresentative fitness scores.

Another avenue of future research includes a more in-depth comparison between higher-order feature interactions in sensitive ML applications and quantitative measure of fairness. In this study, we mainly discussed those feature interactions in that have a relatively obvious relationship to machine bias. However, a more empirical analysis between feature interactions and their relationship to quantitative measures of fairness could yield surprising instances of machine bias. Another step for future research would be to apply this method to contemporary data sets such as are actually in use in decision-making contexts.

ACKNOWLEDGMENTS

We would like to acknowledge the organizations that provided the computing infrastructure and funding to support this research.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May* 23, 2016 (2016), 139–159.
[2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.
[3] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (Oct. 2001), 5–32.
[4] Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *u. Chi. Legal f.* (1989), 139.
[5] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and T. Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* 6, 2 (2002), 182–197. https://doi.org/10.1109/4235.996017
[6] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4, 1 (2018). https://doi.org/10.1126/sciadv.aao5580
[7] Agoston E. Eiben and James E. Smith. 2015. *Introduction to Evolutionary Computing* (2nd ed.).
[8] Jerome H. Friedman and Bogdan E. Popescu. 2008. Predictive learning via rule ensembles. *The Annals of Applied Statistics* 2, 3 (Sep 2008). https://doi.org/10.1214/07-aoas148
[9] David Harrison and Daniel Rubinfeld. 1978. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management* 5 (03 1978), 81–102. https://doi.org/10.1016/0095-0696(78)90006-2
[10] Anna Lauren Hoffmann. 2019. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society* 22, 7 (2019), 900–915.
[11] Jalil Kazemitabar, Arash Amini, Adam Bloniarz, and Ameet S. Talwalkar. 2017. Variable Importance Using Decision Trees. *Advances in Neural Information Processing Systems* 30 (2017), 426–435.
[12] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*. PMLR, 2564–2572.
[13] James W. Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences* 116, 44 (2019), 22071–22080. https://doi.org/10.1073/pnas.1900654116
[14] Sejong Oh. 2019. Feature Interaction in Terms of Prediction Performance. *Applied Sciences* 9, 23 (2019). https://doi.org/10.3390/app9235191
[15] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Michel Vincent., Bertrand Thirion, Oliver Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vinvent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, M. Perrot, and Duchesnay Edouard. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[16] Tim Räz. 2021. Group Fairness: Independence Revisited. *CoRR* abs/2101.02968 (2021). arXiv:2101.02968 https://arxiv.org/abs/2101.02968
[17] Olga Roudenko and Marc Schoenauer. 2004. A Steady Performance Stopping Criterion for Pareto-based Evolutionary Algorithms. In *6th International Multi-Objective Programming and Goal Programming Conference*. Hammamet, Tunisia. https://hal.inria.fr/hal-01909120
[18] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. 59–68.
[19] Nicholas Timme, Wesley Alford, Benjamin Flecker, and John M. Beggs. 2014. Synergy, redundancy, and multivariate information measures: an experimentalist's perspective. *Journal of Computational Neuroscience* 36, 2 (2014), 119–140. https://doi.org/10.1007/s10827-013-0458-4
[20] Forest Yang, Mouhamadou Cisse, and Sanmi Koyejo. 2020. Fairness with overlapping groups; a probabilistic perspective. *Advances in Neural Information Processing Systems* 33 (2020), 4067–4078.