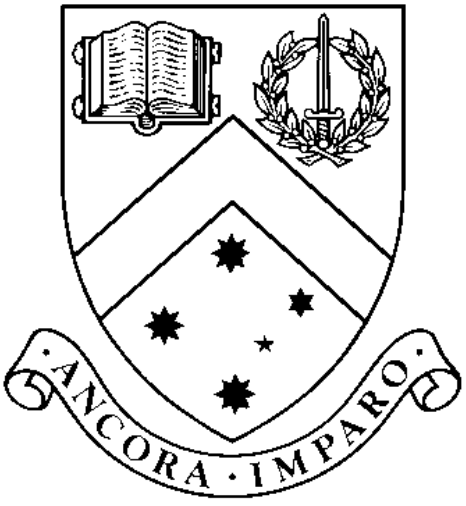
Enhancing Fairness Testing with Bias-Revealing Feature Analysis

Timothy Jordan
Bachelor of Computer Science Advanced (Honours)
31479391

Supervised by Dr. Aldeida Aleti and Dr. Chetan Arora

FIT4443/FIT4444
Thesis
Word Count: 7176



Faculty of Information Technology
Monash University
Clayton Campus
Australia

CONTENTS

Test Input Generation Using Bias-Revealing Features In Fairness Testing

Timothy Jordan
Faculty of Information Technology
Monash University
Melbourne, Australia
tjor0005@student.monash.edu

***Abstract*—Discrimination in artificial intelligence (AI) systems poses significant ethical concerns, particularly given their widespread application in critical decision-making contexts such as healthcare and criminal justice. These systems' potential biases can have severe implications for individuals affected. Fairness testing algorithms are essential to ensure that AI-based systems operate impartially. One critical aspect of fairness testing is test input generation, which uncovers demographic biases through systematically exploring the input space. Current test input generation methods face challenges, primarily due to the time-intensive process required to explore the extensive input space of machine learning (ML) models, which often have numerous features. This research assesses the effectiveness of leveraging significant non-sensitive, bias-revealing features to streamline the test input generation process. By employing Instance Space Analysis (ISA), this study identifies non-sensitive attributes highly correlated with biased outcomes. We enhance the search-based generation methods of fairness testing algorithms—Aequitas, KOSEI, and ExpGA (explanation-guided fairness testing approach utilising a genetic algorithm)—to focus on these identified attributes. Our empirical results indicate that ISA is unable to effectively explain how specific non-sensitive features contribute to bias due to a lack of a linear trend. Adjustments to the fairness testing algorithms to concentrate on sets of identified bias-revealing features yielded mixed outcomes: in 11 cases, there was a notable increase in the success rate of identifying discriminatory inputs, while in 7 cases, performance was unchanged or degraded. Further investigations across diverse datasets and models are necessary to conclusively determine whether focusing on bias-revealing features can enhance the efficacy of fairness testing. This thesis postulates that a more targeted approach in test input generation may potentially increase the detection of discriminatory instances, thereby establishing greater causal fairness in AI-based applications.**

***Index Terms*—Fairness Testing, Input Generation, Machine Learning, Instance Space Analysis**

I. Introduction

Fairness testing is crucial to ensuring that AI-based systems operate without prejudicial bias, identifying and mitigating "fairness bugs"—flaws that cause discrepancies between predicted outcomes and those that satisfy fairness conditions [4]. Such bugs may arise from skewed dataset distributions [29], flawed algorithmic design [11], or societal biases embedded within training data [9]. A critical component of fairness testing algorithms is the generation of samples that elicit discriminatory outputs known as the test input generation phase [4]. Producing a higher number of discriminatory samples provides software developers with critical insights into model behavior, helping to pinpoint and rectify the root causes of fairness bugs.

In assessing the fairness of AI systems, we focus on sensitive attributes—traits we aim to protect from unfair treatment within the system [4]. These typically include characteristics such as sex, race, and age.

Our thesis explores individual fairness within classifier models, specifically addressing causal fairness, which examines the causal relationship between predicted outcomes and sensitive attributes [17]. Consider a classifier model f, designed to predict the likelihood of recidivism, using an attribute vector X for predictions:

$$X = (A_1, X_1, X_2, X_3, X_4)$$

where A_1 represents the sensitive attribute (e.g., age, gender, race), and X_1, X_2, X_3, and X_4 are non-sensitive attributes (e.g., prior convictions, length of last sentence, type of crime).

We define causal fairness as the condition where the classifier's output remains unchanged when only the sensitive attribute A_1 is altered, while all other attributes are held constant. Given two input vectors:

$$x = (a_1, x_1, x_2, x_3, x_4)$$
$$x' = (a_1', x_1, x_2, x_3, x_4)$$

where $a_1 \neq a_1'$, a violation of causal fairness occurs if:

$$f(x) \neq f(x')$$

The process of evaluating test inputs to determine if they are discriminatory is often the most time-consuming task within fairness testing workflows [4]. While search-based test input generation is a well-documented method for measuring causal fairness, there remains substantial scope for enhancing the efficiency and effectiveness of these methods.

This research proposes the utilisation of bias-revealing features—non-sensitive features that, when interacting with sensitive attributes, are likely to induce discriminatory outputs. Our methodology leverages ISA [22], to pinpoint such features in ML classifier models. This approach provides a nuanced understanding of how non-sensitive attributes like zip code, income, and employment status interact with sensitive attributes such as race, potentially leading to disparities due to underlying socioeconomic and demographic patterns [3].

Our research aims to address the following key questions:

RQ1*: How can we identify significant non-sensitive features of discriminatory instances?*

RQ2*: Do significant non-sensitive features help identify bias-revealing inputs more effectively?*

To investigate RQ1, we apply selected fairness testing algorithms—Aequitas, KOSEI, and ExpGA—across commonly used machine learning models, utilising well-known fairness testing benchmark datasets Census Income [2] and Bank Marketing [15]. We record and label discriminatory and non-discriminatory samples within these datasets. These samples are subsequently analysed using ISA to identify attributes strongly linked to discriminatory outcomes. Each identified bias-revealing feature is ranked, and the most impactful features are aggregated across the three algorithms.

For RQ2, we refine the input generation processes of the fairness testing algorithms to focus on the identified significant non-sensitive features. This strategic focus is expected to reduce the exploration of less impactful attributes, enhancing the efficiency and efficacy of the testing process.

We evaluate the impact of these modifications through metrics such as Total Samples Generated (TSN), Total Discriminatory Samples Generated (DSN), Average Time Consumed to Generate a Discriminatory Sample (DSS), and the Success Rate of Generating a Discriminatory Sample (SUR). We also assess the execution times of the modified algorithms compared to their previous versions. Statistical testing like paired t-tests is performed to validate the improvements.

This study aims not only to enhance the performance of generating discriminatory samples but also to offer greater explainability regarding the interactions between non-sensitive features and demographic biases. Additionally, we discuss data anomalies and propose further research to optimise the testing of significant features in fairness testing.

II. Background

In this section, we briefly review relevant background, which includes fairness bugs in AI, current state of literature and its research gaps. We then discuss the importance of non-sensitive features in biased outcomes and provide an overview of ISA

A. Fairness Bugs in AI-based Systems

AI-based systems are increasingly deployed in domains characterised by their high sensitivity and ethical stake. These systems are extensively utilised in healthcare, where machine learning models are leveraged to diagnose conditions such as cardiovascular diseases [1] and liver disorders [10]. In the financial sector, logistic regression models are used to assess credit risks [5]. Moreover, AI is integrated into criminal justice systems to predict recidivism rates [25]. Each application area involves significant ethical concerns, particularly regarding privacy, fairness, and accountability.

The decisions rendered by AI-based systems can often be opaque, complicating efforts to ensure they are free of bias. Consequently, there is a pressing demand for increased transparency and explainability in these systems. The necessity for these measures is underscored by the potential impact of these technologies on individual lives and societal structures, making the ethical dimensions of AI applications an area of paramount concern.

B. Current State of Literature

Robust fairness testing methods are implemented to ensure AI-based systems do not perpetuate prejudiced biases in society. Current state-of-the-art fairness testing algorithms, notably Aequitas [27], KOSEI [18], and ExpGA [7], employ a two-phase search technique to generate discriminatory instances, a methodology critical for uncovering biases in model predictions [4].

1) Aequitas: The two-phase search technique begins with the global search phase, where Aequitas systematically explores the input space to create an initial set of potentially discriminatory instances [27]. It involves indiscriminately selecting inputs within input space and evaluating their discriminatory potential [27].

Following this, the local search phase focuses on searching the neighborhood of these globally-searched instances to further identify discriminatory outcomes [27]. Consider an instance x_1 used as input in the local phase of the Aequitas framework:

$$x_1 = [5, 27, 5, 60, a_1]$$

where a_1 represents the sensitive attribute value. Initially, each non-sensitive attribute is equally likely to be selected for perturbation. Assume the first attribute (with value 5) is chosen for perturbation. The perturbation process adjusts this attribute by either incrementing or decrementing its value by 1, based on a probabilistic decision. Suppose the value is incremented by 1, resulting in a new instance x':

$$x' = [6, 27, 5, 60, a_1]$$

The evaluation of x' determines whether it is classified as a discriminatory instance. Based on this assessment, the algorithm is updated to either increase or decrease the likelihood of selecting this non-sensitive feature for future iterations [27].

2) KOSEI: Building on the local search strategy of Aequitas, KOSEI enhances the search process by allowing for the perturbation of all non-sensitive feature by increments of +1 and -1 for each iteration [18]. This approach enables a more exhaustive exploration of the neighborhood around a global discriminatory instance, increasing the likelihood of identifying additional discriminatory samples [18].

3) ExpGA: ExpGA introduces a different paradigm consisting of an initialisation phase and an optimisation phase. The initialisation phase begins with random generation of samples from the input space, followed by the use of interpretable methods to rank the importance of each feature in influencing the model's output [7]. Features with higher importance scores from the explanation results suggest that minor perturbations could significantly alter the predicted outcome [7]. ExpGA

selects instances in which the interpreter has identified the sensitive attribute as highly ranked [7]. These identified instances then form a set of seed samples for further analysis.

The optimisation phase leverages the strengths of a genetic algorithm to heighten the diversity and quality of the test inputs [7]. This phase encompasses three key processes:

1) **Selection:** From the initial set of seed samples, high-quality instances are selected to generate the next population [7]. This selection process optimises a fitness function [7]. For a given sample x, two derivative inputs x' and x'' are created, differing only in the value of their sensitive attribute. The fitness function $f(x)$ is defined as the absolute difference in predicted outcomes of these two inputs, i.e.,

$$f(x) = |\text{Model}(x') - \text{Model}(x'')|$$

where $\text{Model}(x')$ and $\text{Model}(x'')$ represent the model's outputs. If the objective function value is above a certain threshold, then the sample is selected.

2) **Crossover and Mutation:** These genetic operators are utilised to introduce diversity within the population. The application of crossover and mutation is governed by their respective probability rates. During crossover, a pair of samples is randomly selected, and a set of features is chosen at random to be exchanged between these samples [7]. In the mutation process, selected features of a sample are randomly altered to create new genetic variations [7].

These processes collectively enhance the genetic diversity and fitness of the population, facilitating the identification of potentially discriminatory instances by exploring a wider space of input modifications.

C. Limitations & Research Gaps with Current Fairness Testing Algorithms

Despite the pivotal role of test input generation in fairness testing, the scholarly exploration of this area is in its infancy where initial methodologies, such as the one used in THEMIS, which employs random instance generation [4], [8], published only in 2017. A research gap exists in the identification of non-sensitive attributes that significantly influence biased outcomes within ML models. As these models increasingly incorporate a diverse array of features into their decision-making processes, understanding the interaction between non-sensitive attributes and machine learning outcomes becomes crucial, particularly in sensitive applications where these attributes may correlate indirectly with sensitive characteristics.

The strategic identification of bias-revealing features can illuminate new avenues for generating discriminatory outcome insights. Fairness testing algorithms like Aequitas and KOSEI conduct a uniform search across the input space to identify discriminatory instances [18], [27]. However, if a model's biased behavior is localised within specific regions of the input space, such uniform search strategies may prove inadequate for uncovering all discriminatory samples. By narrowing the focus of test input generation to specific non-sensitive features known to impact biased outcomes significantly, it is possible to refine the search and more effectively target the nuances of machine learning model behavior that induce bias. This targeted approach not only increases the percentage of discriminatory samples identified but also enhances the overall effectiveness of the testing process.

Currently, there is no universally applicable method to enhance the efficiency of test input generation across fairness testing algorithms. The process of fairness testing, particularly the generation and evaluation of test inputs, can be resource-intensive, especially when the model under test incorporates a substantial number of features [28]. By concentrating the test input generation on a specific subset of significant non-sensitive features, we can substantially reduce the time required to generate discriminatory samples. This reduction in execution time enables software developers to conduct fairness testing more frequently, thereby promoting the development of more equitable AI-based systems.

D. Connection of Non-Sensitive Features to Bias

The interaction between bias-revealing features and sensitive attributes often indirectly leads to disparities in model decision-making processes. For instance, consider a model designed to assess credit-worthiness for loan approvals, which incorporates both non-sensitive attribute zip code, and sensitive attribute race. The interplay between zip code and race may inadvertently cause disparities due to underlying socio-economic and demographic patterns linked to historical redlining, where zip codes can be strongly correlated with racial demographics [3]. Consequently, regions characterised by specific zip codes, predominantly housing minority communities, may exhibit lower average incomes, higher unemployment rates, and reduced access to educational resources, factors that could influence assessments of creditworthiness [3].

These hidden interactions can be propagated into the machine learning model during training, as the model may recognise patterns where higher loan default rates are prevalent in certain zip codes, which coincidentally align with higher populations of specific racial groups due to historical and socio-economic factors [30]. Should the model assign significant weight to the zip code in predicting creditworthiness, it may inadvertently enforce discriminatory biases against individuals from those areas, disproportionately affecting racial minorities.

The primary aim of our research is to uncover which non-sensitive attributes effectively reveal discriminatory patterns within various ML models. Additionally, we seek to assess whether modifying the test input generation of modern fairness testing algorithms to prioritise these bias-revealing features enhances their efficiency and effectiveness.

E. Instance Space Analysis (ISA)

Originally developed by Kate Smith-Miles et al., ISA aims to enhance the objective evaluation of algorithms through a detailed exploration of the instance space associated with algorithm performance [22]. ISA is fundamentally designed to reveal how varying feature properties impact algorithmic

efficacy, primarily through the creation of visualisations that map the entire space of potential test instances [22]. These visualisations help identify "algorithm footprints," which are regions within the instance space where the algorithm is likely to perform optimally.

The versatility of ISA has facilitated its application across diverse domains. It has been employed to gain nuanced performance insight to known optimisation challenges such as the knapsack problem [20] and the job-shop scheduling problem [23], and to evaluate the performance of various software testing algorithms [16].

ISA consists of 4 major tasks to perform objective performance analysis. Within these tasks, we can perform modifications to target causal fairness performance:

- **PRELIM *Preparation for Learning of Instance Metadata*:** This initial phase involves standardising the metadata, preparing the data for feature selection and principle component analysis (PCA) [22].
- **SIFTED *Selection of Instance Features to Explain Difficulty* :** In this phase, automated feature subset selection is applied to eliminate irrelevant non-sensitive features—specifically, those that are either redundantly correlated with other features or exhibit weak Pearson correlation with the target outcome [22]. This stage focuses on retaining features that are uncorrelated with each other yet strongly associated with discriminatory outputs, ensuring the model's focus on relevant predictors [22].
- **PILOT *(Projecting Instances with Linearly Observable Trends)***: Implements a Principle Component Analysis (PCA) dimension reduction technique which allows one to project instances onto a 2D instance space visualisation [22]. A novel method is implemented to maximise the linear relationship in non-sensitive features and causal fairness performance [22].
- **CLOISTER *(Correlated Limits of the Instance Space's Theoretical or Experimental Regions)***: This phase consists of outlining the theoretical boundaries of the instances space where model behaviour has a high likelihood of inducing bias [22]. The theoretical upper and lower boundaries of each feature are used to define these boundaries [22]. We also factor in correlation scores calculated in SIFTED stage to exclude areas in which there is strong correlation between features in the instance space boundary [22]

These components of ISA are integral to our methodology, enabling us to precisely target and address the nuances of discrimination within machine learning algorithms.

III. Details of the Purposed System

A. Experimental Design

This section outlines the methodology adopted to address each of the research questions posed in this study.

Addressing RQ1: Initially, we execute the selected state-of-the-art fairness testing algorithms—namely Aequitas, KOSEI, and ExpGA—on various machine learning classifiers, including Decision Trees (DT), Random Forests (RF), and Multi-Layer Perceptron Classifiers (MPLC). The primary objective is to assess the causal unfairness of these models by quantifying the discriminatory samples generated by each algorithm. During these tests, both discriminatory and non-discriminatory instances are recorded. Subsequently, these datasets undergo Instance Space Analysis (ISA) to extract insights concerning the association of non-sensitive features with bias across each classifier model. A non-sensitive attribute is identified as a bias-revealing feature if its Pearson correlation coefficient, as determined by ISA, exceeds 0.3, thereby indicating at least a weak association with bias. **Addressing RQ1:** Our approach commenced with the deployment of state-of-the-art fairness testing algorithms—specifically, Aequitas, KOSEI, and ExpGA—across a variety of machine learning classifiers, including Decision Trees (DT), Random Forests (RF), and Multi-Layer Perceptron Classifiers (MPLC). The primary objective of this phase was to evaluate the causal unfairness exhibited by these models. This evaluation was achieved by quantifying the number of discriminatory samples produced by each algorithm. Throughout these evaluations, we meticulously recorded both discriminatory and non-discriminatory instances generated during the testing process.

Following the initial testing phase, the collected datasets were subjected to Instance Space Analysis (ISA). This analysis was aimed at deriving deeper insights into how non-sensitive features relate to biased outcomes within each classifier. Specifically, we focused on identifying non-sensitive attributes that significantly correlate with bias. An attribute was designated as a bias-revealing feature if it ranked within the top ten based on its Pearson correlation coefficient, indicating a strong association with discriminatory outcomes.

Addressing RQ2: utilising the bias-revealing features identified through ISA, we refine the input search generation mechanisms for each fairness testing algorithm. This ISA-guided approach involves averaging the Pearson correlation coefficients of each feature across all fairness testing algorithms for a given classifier. For example, with the DT classifier, correlation coefficients from the discriminatory samples generated by Aequitas, KOSEI, and ExpGA are averaged to generalise findings and the top ten significant non-sensitive attributes influencing biased outcomes are chosen. Subsequent modifications to the fairness testing algorithms involve restricting the input boundaries of non-bias revealing features, allowing for a concentrated focus on more impactful features.

To rigorously assess the improvements in the fairness testing algorithms modified in RQ2, we employ the following evaluation metrics:

1) **Total Samples Generated (TSN)**: Represents the total number of samples produced during the testing phase.
2) **Total Discriminatory Samples Generated (DSN)**: Denotes the count of samples that resulted in discrimination when only the sensitive attribute was altered.
3) **Average Time Taken to Generate a Discriminatory**

Sample (DSS): Calculated as

$$\text{DSS} = \frac{\text{Execution Time}}{\text{DSN}}$$

where "Execution Time" encompasses the complete duration taken by the fairness testing algorithm.

4) **Success Rate of Generating a Discriminatory Sample (SUR)**: This effectiveness metric is computed as

$$\text{SUR} = \frac{\text{DSN}}{\text{TSN}}$$

,reflecting the proportion of discriminatory samples out of the total generated.

Furthermore, the average execution time, represented as "Time", of the revised fairness testing algorithms is analysed to assess enhancements in efficiency.

To ensure the robustness of our results, each fairness testing algorithm, classifier, and benchmark dataset combination undergoes 25 iterations. Paired t-tests are utilised for statistical analysis to determine significant enhancements in the performance of fairness testing algorithms.

From our conducted experiment, we hypothesise a reduction in TSN and DSN, attributed to the constrained input boundaries stemming from the ISA-directed focus, which limits the diversity of inputs generated. This consequently impacts the total number of unique discriminatory samples able to be detected, reducing DSN. However, improvements are expected in DSS, Time, and SUR metrics, as the input generation is more targeted around features that have a higher likelihood of inducing a biased outcome, potentially increasing the success rate of detecting discriminatory samples. The focused input search also implies reduced time spent exploring less promising areas of the input space, thereby enhancing both DSS and Time metrics.

B. Datasets

As outlined in Table I, our experimental design applies two datasets: the Census Income dataset and the Bank Marketing dataset. The Census Income dataset comprises demographic and employment-related data to predict whether an individual's annual income exceeds $50,000 [2]. The Bank Marketing dataset, on the other hand, consists of customer data from a banking institution, used to predict the likelihood of a customer subscribing to a bank term deposit [15]. These datasets are frequently employed in fairness testing research to evaluate the implications of classifier models within financially sensitive domains [7], [12], [27].

C. Uncovering Bias-Revealing Features Using ISA

ISA can be tailored to analyse bias within machine learning models by focusing on the performance of instances in terms of causal unfairness induction.

1) ISA Execution: ISA is employed to scrutinise each set of discriminatory samples produced by the fairness testing algorithms. This analysis is performed consistently across each classifier, dataset and fairness testing algorithm combination. The Github repository InstanceSpace [21] is leveraged perform analysis and parameters are customised to suit the specific requirements of our experiments.

2) Metadata Construction and Labeling: The metadata provided to ISA consists of test instances generated by the fairness testing algorithms. Each instance is labeled to indicate whether it has produced a discriminatory outcome—that is, whether a variation in the sensitive attribute has led to a change in the output of the classifier model under test. Discriminatory instances are labeled as 1, while non-discriminatory instances are labeled as 0.

3) Parameter Configurations for PRELIM Stage: In the PRELIM stage of ISA, initial data pre-processing is conducted. A Box-Cox transformation is applied to transform the dataset to resemble a normal distribution [22]. This step addresses potential skewness in the data distribution, which could bias subsequent analyses such as k-means clustering within the SIFTED stage [22] . Subsequently, z-score normalization is performed, scaling all features to a mean of zero and a standard deviation of one. This standardisation ensures that features are evaluated on an equal basis, allowing the correlation calculations to accurately reflect true associations with the discriminatory label, independent of the original data scales [22].

The configuration of ISA for this stage sets the performance measure to be absolute, indicating that any discriminatory instances (labeled as '1') are considered errors. The settings aim to maximise the detection of discriminatory samples.

4) Parameter Configurations for SIFTED Stage: The SIFTED stage is critical for automated feature selection, aiming to eliminate irrelevant features through Pearson cross-correlation and k-means clustering analyses [22]. This stage leverages the Pearson correlation coefficient to evaluate the relationship between features and the discriminatory label. This process ensures that only features with strong correlations to the discriminatory label are retained for further analysis [22]. During our experiment, SIFTED is configured to maximise the classifier causal unfairness performance, that is, focusing on finding the maximum theoretical boundary of areas that are discriminatory in the instance space. The performance of an instance is measured in absolute, where any instance labelled as 1 is classified as discriminatory. Features are ranked based on the absolute values of their correlation coefficients, with those falling below a predefined threshold of 0.01 labelled as insignificant [22]. The list of feature correlation values generated from this stage is extracted and analysed to determine the features that have the strongest correlation to the discriminatory label and thus significantly impact bias.

Parameters for PILOT and CLOISTER stages remain in their optimised default configurations. We extract the visualisation generated by the PILOT and CLOISTER stages to gain more insight as to algorithm footprints of the instance space where discriminatory instances are more prevalent.

Dataset	Description	Sensitive Attribute	Number of Features	Total Input Combinations	URL
Adult	Income prediction	Sex	14	2.71×10^{15}	[2]
Bank Marketing	Deposit subscription prediction	Age	16	1.91×10^{16}	[15]

TABLE I: Datasets used for fairness testing

D. Enhancing Input Test Generation with Bias-Revealing Features

Building on the insights gathered from the application of ISA, this section elaborates on the strategic modification of the test input generation phase for fairness testing algorithms Aequitas, KOSEI, and ExpGA. The primary objective is to refine these algorithms and augment their capability in generating discriminatory samples effectively and efficiently. It is important to mention that the modifications and comparative analysis testing were conducted using the existing GitHub codebases for Aequitas [26], KOSEI [19], and ExpGA [6], which provided a foundational framework for our enhancements and evaluations.

Input Boundaries and Search Mechanisms: Search-based test input generation methodologies operate within defined input boundaries to explore the input space [4]. This exploration occurs in two phases: initial input space exploration and subsequent focused searches around neighborhoods of previously identified discriminatory samples [4]. To optimise this process, we adjust the input boundaries for Aequitas, KOSEI, and ExpGA, emphasising the enhancement of input variability for bias-revealing features while restricting the variability for features deemed insignificant by ISA.

Feature Constraints: Let $\mathcal{F} = \{f_1, f_2, \ldots, f_n\}$ denote the entire feature space, where each f_i represents a feature within the dataset. Each feature f_i is bound by input constraints defined by an interval $[x_i, y_i]$, representing the minimum and maximum permissible values respectively:

$$f_i \in [x_i, y_i], \quad \forall i \in \{1, 2, \ldots, n\}$$

Upon determining the relative significance of each feature concerning unfairness, the input range $[x_i, y_i]$ for critical features is preserved to maintain their exploratory potential. Conversely, for features not identified by ISA as significantly correlated with causal unfairness, this range is contracted.

Constraining Non-Significant Features: Consider $\mathcal{F}_{\text{ns}}$ as the set of non-significant features. For each non-significant feature f_i, let δ_i denote the smallest permissible range, and m_i the midpoint of the initial range $[x_i, y_i]$. The revised constraints for these features are then specified as follows:

$$f_i \in [m_i, m_i + \delta_i], \quad \forall f_i \in \mathcal{F}_{\text{ns}}$$

Here, δ_i represents the minimal range width, centered around m_i, effectively restricting the variability of non-significant features. This focused approach minimises unnecessary exploration within the input space, therefore channeling the test input generation efforts towards evaluating combinations of the more impactful features. These modifications are designed to enhance the likelihood of discovering biased outcomes in the ML model, facilitating a more efficient and targeted detection of potential discriminatory instances.

E. Tolerance for Errors

For search-based fairness testing algorithms, randomness is factored in when generating the initial set of discriminatory samples [4]. Due to this, there is inherent variability in the test input generation performance. Hence it is crucial to implement measures that ensure the reliability and accuracy of our performance metrics. To mitigate the impact of randomness, anomalies, and inherent errors—such as fluctuating detection rates of discriminatory samples or inconsistencies in execution times—we conduct multiple iterations of each fairness testing algorithm across all classifiers. The mean performance is the computed across iterations.

Subsequently, we employ paired t-testing to determine whether the observed differences in performance metrics between the original and modified test input generation are statistically significant. This analysis provides a rigorous statistical foundation to validate the efficacy of the modifications implemented in the fairness testing algorithms.

This structured approach to handling errors not only enhances the reliability of our findings but also substantiates the statistical significance of any observed improvements, therefore reinforcing the validity of our research outcomes.

IV. Results and Discussion

A. Achieved Results from ISA Analysis

The empirical results depicted in Figures 1 and 2 indicate a generally poor performance by ISA in identifying strong correlations in both the Census and Bank Marketing datasets. For the Census dataset, as illustrated in Figure 1, ISA struggled to identify features with significant correlations to discriminatory outcomes in the RF and MLPC models. Notably, only two features—'hours per week' and 'capital loss'—exhibited correlation coefficients above 0.1 when analysing the MLPC model with the ExpGA dataset. Similar trends were observed in the RF model with the KOSEI dataset, where only 'capital gain', 'final weight', 'marital status', and 'relationship' reached correlation values of 0.1, 0.14, 0.12, and 0.1, respectively.

An exception was noted in the DT model with the KOSEI dataset, where ISA successfully identified 'final weight', 'occupation', and 'education' as features demonstrating strong correlations to the discriminatory output, with coefficients of 0.59, 0.53, and 0.54, respectively. Additionally, there was a convergence of correlation values between the ExpGA, KOSEI, and Aequitas datasets for 'capital gain' and 'relationship',

with 'capital gain' exhibiting correlations of 0.46 and 0.48 for ExpGA and KOSEI, respectively, and 'relationship' showing correlations of 0.39 and 0.35 for KOSEI and Aequitas, respectively.

The underlying causes for the observed discrepancies in correlation performance across different models and algorithms are explored further in the discussion section. These findings underscore the complex and variable nature of feature significance within machine learning models subjected to fairness testing, as indicated by the ISA methodology.

For the Bank dataset shown in in Figure 2, ISA has identified only a limited number of features with notable correlations to discriminatory outcomes across the RF, MLPC, and DT models within the fairness testing algorithm generated samples. Notably, the test inputs generated by ExpGA for the MLPC model highlights 'previous', 'campaign', and 'pdays' as significant bias-revealing features, exhibiting correlation values of 0.32, 0.44, and 0.55 respectively. However, these features do not show similar correlation strengths in KOSEI and Aequitas, indicating a divergence in feature significance across different fairness testing algorithms.

For the DT classifier tested using Aequitas, the features 'duration', 'month', and 'balance' were identified as having strong bias-revealing features with correlation values of 0.76, 0.64, and 0.45 respectively. These results are somewhat supported by the findings from the ExpGA generated samples, which also recognised the substantial impact of the 'duration' feature with a strong correlation of 0.48 to discriminatory outcomes.

Despite these findings, Figure 2 generally indicates a scarcity of non-sensitive features in the Bank dataset that significantly influence discriminatory behavior for all the models under study. This observation suggests limitations in the predictive capacity of non-sensitive features to consistently reveal bias across various classifiers and fairness testing algorithms, pointing to a complex and potentially model-specific interaction between feature significance and discriminatory outcomes.

Figures 3 and 4 display the rankings of each non-sensitive feature based on the average correlation value calculated across various fairness testing algorithms for the Census and Bank Marketing datasets respectively. It is important to mention, the Decision Tree (DT) classifier's data has been plotted twice in each figure to facilitate a more comprehensive visual analysis.

In the Census dataset shown in Figure 3, the features 'occupation' and 'final weight' consistently maintain high importance across the three classifier models, varying between ranks two and four. The DT and MLPC classifiers exhibit the most similar sets of significant feature rankings. Here, five features—namely 'occupation', 'final weight', 'relationship', 'hours per week', and 'native country'—either share the same ranking or differ by only one rank. Conversely, the DT and RF models show the least similarity in feature rankings, where 'marital status' and 'relationship' are labelled in the top 5 significant feature for DT but are classified as the least significant features for RF.

As depicted in Figure 4 , the variability in feature rankings is notably higher in Bank dataset compared to the Census dataset, indicating unique classifier discriminatory responses to the features in the Bank Marketing benchmark.. The maximum number of features showing similar rankings (difference of at most 1 in ranking) between any two classifier models is only two. Despite the general variability, certain features such as 'month' and 'poutcome' consistently appear within the top six most significant features across all classifiers, suggesting their pivotal role in influencing the discriminatory output. However, substantial differences in the ranked importance of other features are evident, particularly when comparing the DT and MLPC models. Notably, 'pdays', 'campaign', and 'previous' are ranked as the top three most significant bias-revealing features for the MLPC model, yet these features are positioned at much lower ranks in the DT model at ranks 9, 12, and 14 respectively.

This pronounced disparity in feature rankings across different classifiers signifies the complex nature of feature interactions within machine learning models and highlights the challenges in generalising feature significance across different ML model architectures. The results suggest that the effectiveness of bias-revealing features in detecting discriminatory outcomes can vary significantly depending on the classifier used, thus necessitating a model-specific approach in fairness testing to adequately address potential biases.

B. Instance Space of Feature Significance

The results in Figures 1 and 2 illustrates the challenges in identifying substantial correlation values above 0.3 for the RF and MLPC models. The 2D instance spaces generated during the CLOISTER stage of ISA provide deeper insights into why certain features do not significantly influence discriminatory outcomes. Let us compare the instance spaces from the KOSEI generated samples on DT classifier Census dataset in contrast to the Aequitas generated samples on RF classifier Bank Marketing dataset.

The instance space for the KOSEI testing on the DT classifier within the Census dataset offers valuable perspectives on feature significance. As depicted in Figure 5, the three instance spaces highlights the two highest-ranking significant features, 'education' and 'occupation', and the discriminatory labels. Figure 5c shows a clear separation between discriminatory and normal instances, where biased instances are clustered together in a specific area of the instance space. Notably, as illustrated by the red boxes in Figures 5a and 5b, a significant majority of the values within this cluster are uniform, suggesting a strong correlation between these feature values and the biased outcomes.

In contrast, the instance spaces generated from the other experiments show unclear patterns. For instance, Aequitas testing for the RF classifier against the Bank Marketing dataset, as shown in Figure 6, exhibit a stark difference in the instance space generated. Figure 6c reveals no distinct separation between discriminatory and non-discriminatory instances, with a substantial overlap evident between the two. This lack of clear

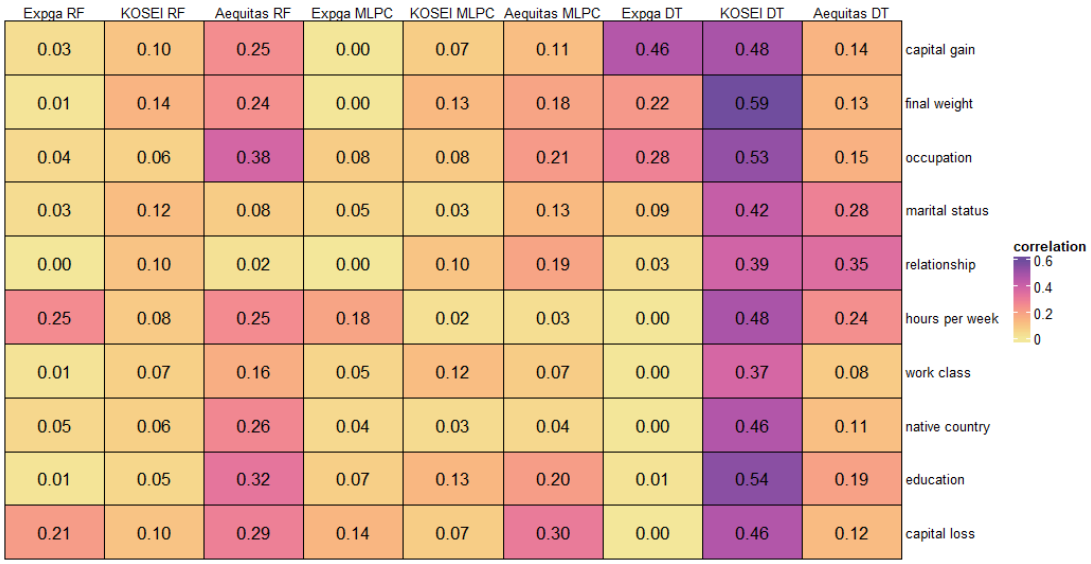


Fig. 1: Pearson correlation of census data features against discriminatory output. Calculated for each fairness testing algorithm under each classifier

delineation is further emphasised in the value distributions for ‘contact‘ and ‘housing‘, displayed in Figures 6a and 6b. The scattering of these feature values across the instance space shows no discernible pattern, indicating the complexity and randomness in how these features interact with the model’s biased behaviour.

The visualisations generated by ISA not only enhance our understanding of feature behaviour but also suggest strategic directions for focusing test input searches. Specifically, by targeting specific areas in the instance space where discriminatory instances are more likely, such as the regions highlighted in Figure 5a and Figure 5b, modifications in feature values can be effectively employed to probe and address potential biases. This focused approach could significantly improve the efficiency of fairness testing algorithms by honing in on the most impactful features and instance characteristics.

C. Performance On Test Input Generation Enhancement

Tables II, III, and IV present the comparative performance of the fairness testing algorithms KOSEI, Aequitas, and ExpGA on the MLPC, RF, and DT classifiers respectively. These tables illustrate the normal performance of these algorithms compared with their performance when test input generation methods have been modified to target bias-revealing features as directed by ISA.

1) TSN Comparative Analysis: In contrast to our hypothesis, 7 out of the 18 comparative tests the ISA-guided approach outperformed its original counterpart. The most significant increase in Total Samples Generated (TSN) was observed in Table IIa, where the ISA-Directed KOSEI generated 33,540 more samples than its original version on average. Contrary to expectations, this increase occurred despite the ISA-directed approach reducing the number of possible input combinations by constraining the boundaries of non-impactful features.

Conversely, 11 of the 18 tests showed a decrease in TSN, aligning with initial thesis assumptions. For instance, the ExpGA testing the RF model under the Census benchmark showed a decrease of 25,324 in TSN as indicated in Table IIIc.

2) DSN Comparative Analysis: In 11 of the tests, an increase in Discriminatory Sample Numbers (DSN) was observed using the ISA-directed strategy. Remarkably, five of these instances also showed a decrease in TSN. For instance, the Aequitas DT classifier on the Bank dataset (see Table IVb) demonstrated an increase in average DSN from 5,086 to 8,438, with a reduction in average TSN from 55,760 to 49,437. This highlights the effectiveness of bias-revealing features in enhancing the proportion of discriminatory samples generated.

3) DSS Comparative Analysis: DSS value remained relatively consistent across most datasets and classifiers when comparing the modified fairness testing algorithms to their original counterparts. However, some areas showed worse performance under the ISA-directed approach. Notably, Aequitas testing the MLPC classifier saw an increase in DSS from 0.2179 seconds to 0.6614 seconds, a statistically significant difference as verified by paired t-testing (see Table IIb).

4) Time Comparative Analysis: In only half of the experiments a decrease in average execution time was observed after implementing the ISA-directed fairness testing algorithms; the remainder surprisingly showed an increase. Notable outliers include the Aequitas-tested MLPC and DT classifiers under the Bank dataset (Tables IIb and IVb), where the average execution times increased significantly. We observed the initial time to be 127.51 sec and 181.29 sec to a worse performance of 456.18 sec and 410.14 sec, which is 3.58x and 2.26x times the original time taken for MLPC and DT classifiers respectively. This variability leaves it unclear whether the ISA-directed approach consistently enhances the efficiency of

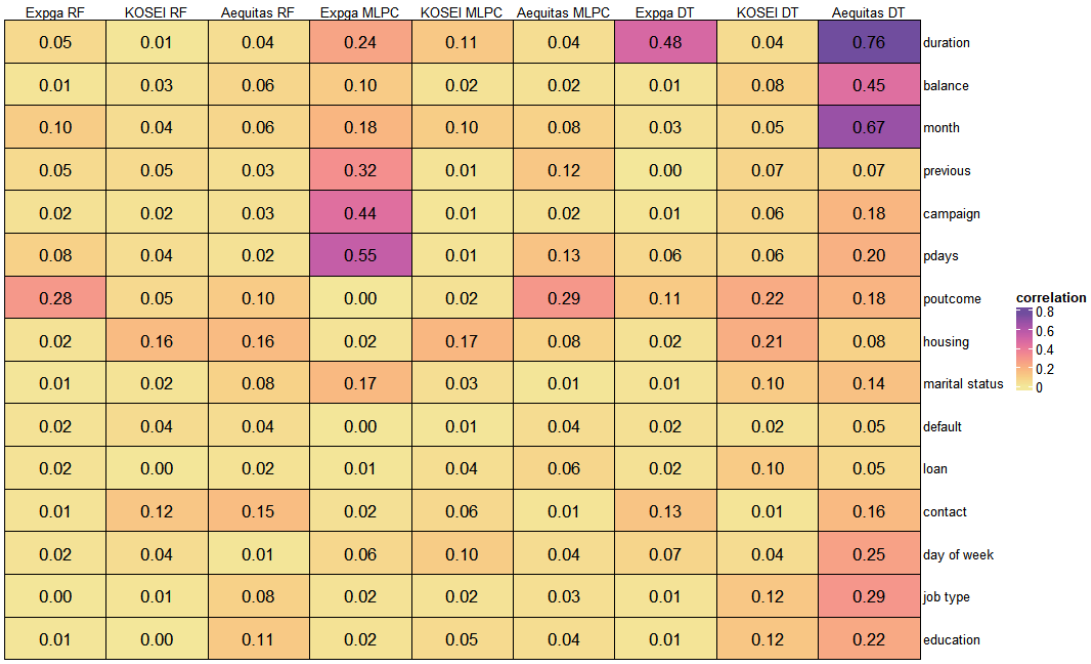


Fig. 2: Pearson correlation of Bank Marketing features against discriminatory output. Calculated for each fairness testing algorithm under each classifier

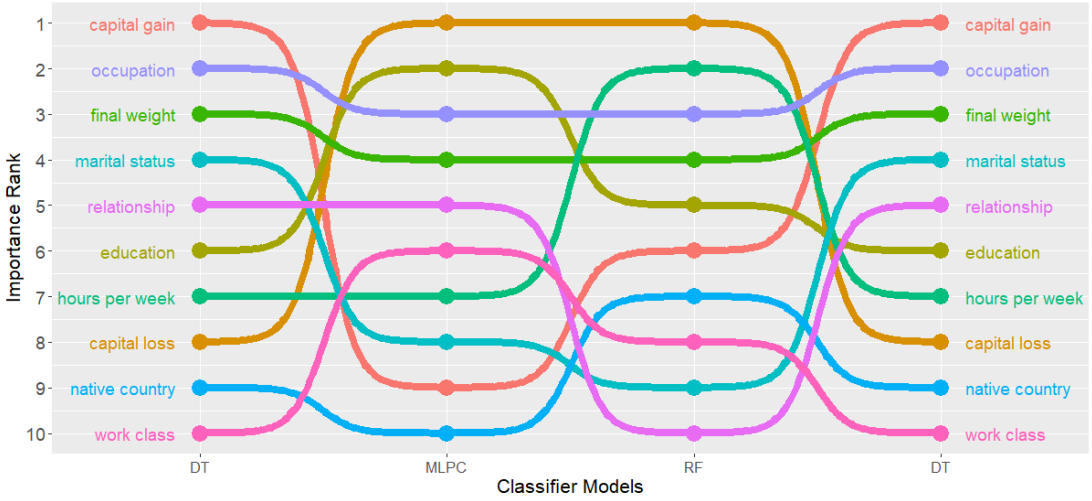


Fig. 3: Ranking of most significant bias-revealing features in Census dataset across classifiers

fairness testing algorithms.

5) SUR Comparative Analysis: A significant improvement was noted in SUR performance in 15 of the 18 experiments. Paired t-tests on the SUR differences indicated that 11 of these improvements were statistically significant, marked with an asterisk (*) and a p-value less than 0.05. The largest increase was observed with Aequitas under the MLPC on the Census benchmark, showing a 26.62% increase in SUR, from 22.98% to 49.59%.

D. Trends in Comparative Performance

Our empirical results displayed a trend where the ISA-directed strategy resulted in an increased SUR performance but decreased DSS and overall execution time. This phenomenon is observed primarily within Aequitas testing the MLPC and DT classifiers on the Census and Bank Marketing datasets, as highlighted by the † symbol shown in Table IIb and Table IVb respectively. The paired t-tests confirm significant mean differences between the modified and unmodified performances.

Further analysis reveals that the unmodified version of Aequitas generated an average of 22 global discriminatory

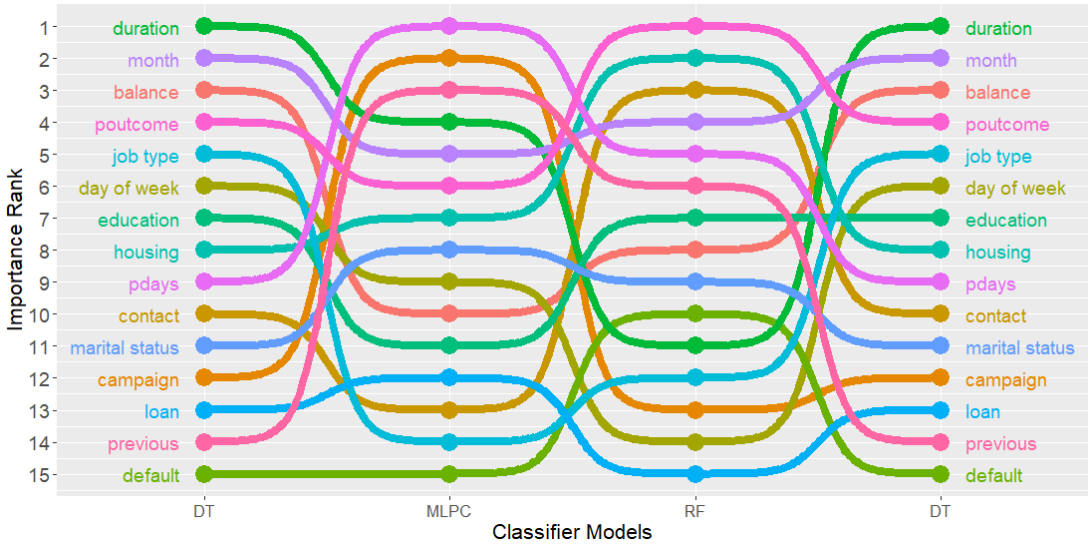


Fig. 4: Ranking of most significant bias-revealing features in Bank Marketing dataset across classifiers

TABLE II: Comparison of test input generation for MLPC model

(a) KOSEI comparison

Dataset	KOSEI					KOSEI ISA-Directed				
	TSN	DSN	DSS	Time	SUR	TSN	DSN	DSS	Time	SUR
Census (gender)	43,862	27,560	0.0041	112.25	62.78%	**44,746**	**28,440**	**0.0041**	**114.71**	**63.56%**
Bank (age)	47,011	28,295	0.0006	17.96	60.18%	**80,551**	**50,690**	**0.0004**	**22.16**	**62.92%***

(b) Aequitas comparison

Dataset	Aequitas					Aequitas ISA-Directed				
	TSN	DSN	DSS	Time	SUR	TSN	DSN	DSS	Time	SUR
Census (gender)	22,779	5,057	0.2179	6.74	22.98%	**14,276**	**6,476**	**0.6614 †**	**8.55 †**	**49.59%***
Bank (age)	4,843	1,531	0.0843	127.51	31.98%	**3,870**	**1,329**	**0.3461 †**	**456.18 †**	**34.51%***

(c) ExpGA comparison

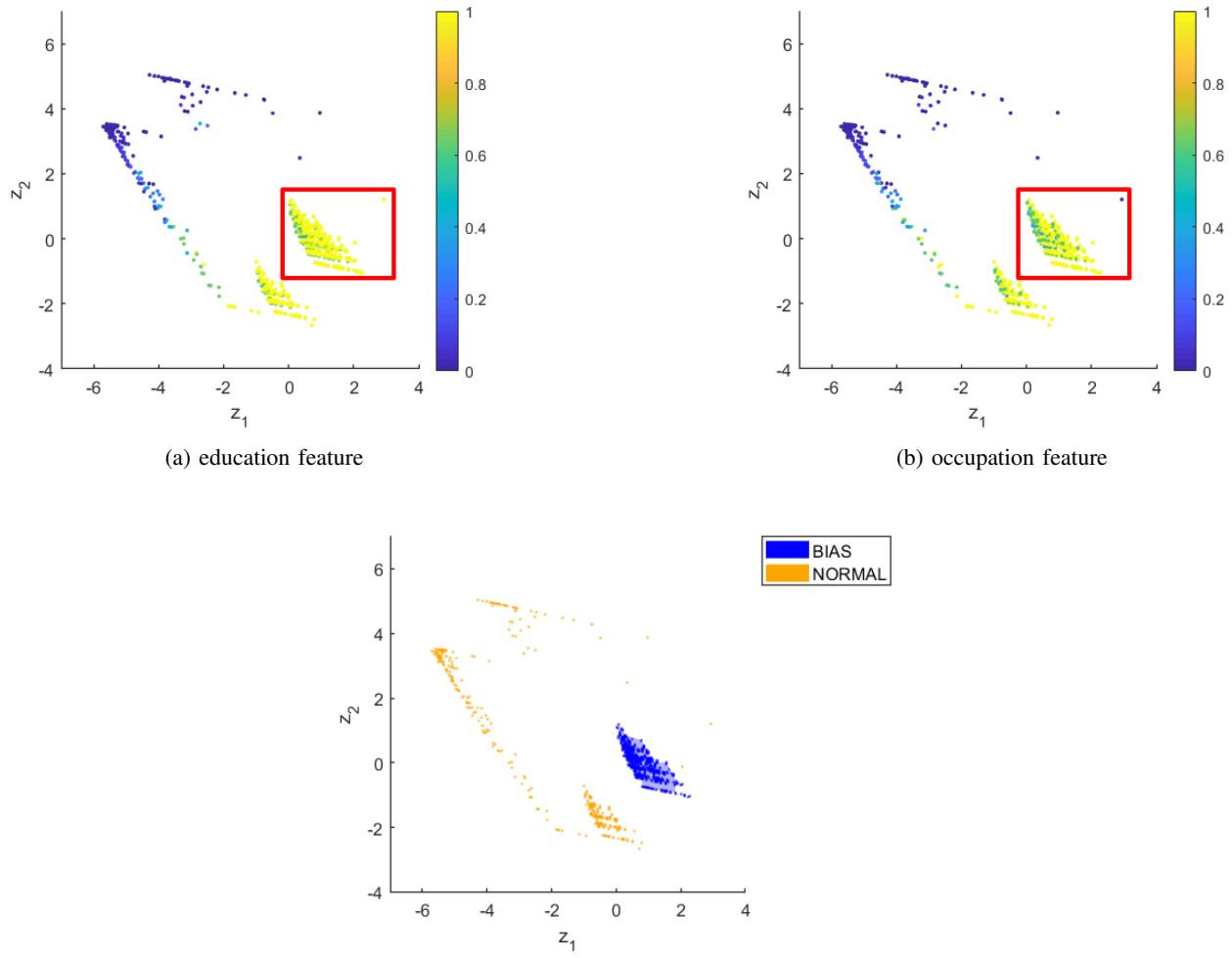
Dataset	ExpGA					ExpGA ISA-Directed				
	TSN	DSN	DSS	Time	SUR	TSN	DSN	DSS	Time	SUR
Census (gender)	75,597	42,141	0.0104	440.16	55.74%	**72,405**	**43,416**	**0.0102**	**441.52**	**59.96%***
Bank (age)	71,535	14,804	0.0137	196.89	20.70%	**70,675**	**14,335**	**0.0158**	**204.24**	**20.29%**

samples, whereas the ISA-directed version uncovered an average of 58 global biased instances. This substantial increase by over 2.6x times potentially influences the average Time and DSS performance negatively. Aequitas's local search mechanism, designed to generate more neighboring samples for this increased amount of global discriminatory samples, results in a higher number of local instances that need to be evaluated [27]. Evaluating test inputs is time-intensive as it requires the classifier model to execute twice per instance when the sensitive attribute value varies [4]. This process significantly extends the time Aequitas requires to run the local search phase, thus leading to increased average DSS and Time.

Similar phenomena are observed in some of the KOSEI experiments, specifically when testing the MLPC classifier under the Census and Bank datasets as shown in Figure IIa. Given KOSEI's local search scheme is fundamentally based on Aequitas's, this result is expected. However, KOSEI is less affected due to its mechanism for recognising previously generated inputs [18], which prevents redundant evaluations and therefore does not suffer as significantly from increased execution times and DSS performance.

From these observations, it is evident that while the utilisation of bias-revealing features in search-based fairness testing can improve the success rate of identifying discriminatory samples, it may lead to decreased performance efficiency. As more global discriminatory samples are discovered in the initial phase, the fairness testing algorithm must evaluate a larger number of neighboring inputs. Without efficient mechanisms to evaluate these inputs, a decrease in average DSS and execution time is observed, indicating a trade-off between effectiveness and efficiency.

Notably, there are outliers where the SUR value decreases,



(a) education feature

(b) occupation feature

(c) discriminatory outcome

Fig. 5: Distribution of education and occupation features in the instance space from KOSEI DT Census. (c) shows discriminatory instances within theoretical boundaries

TABLE III: Comparison of test input generation for RF model

(a) KOSEI comparison

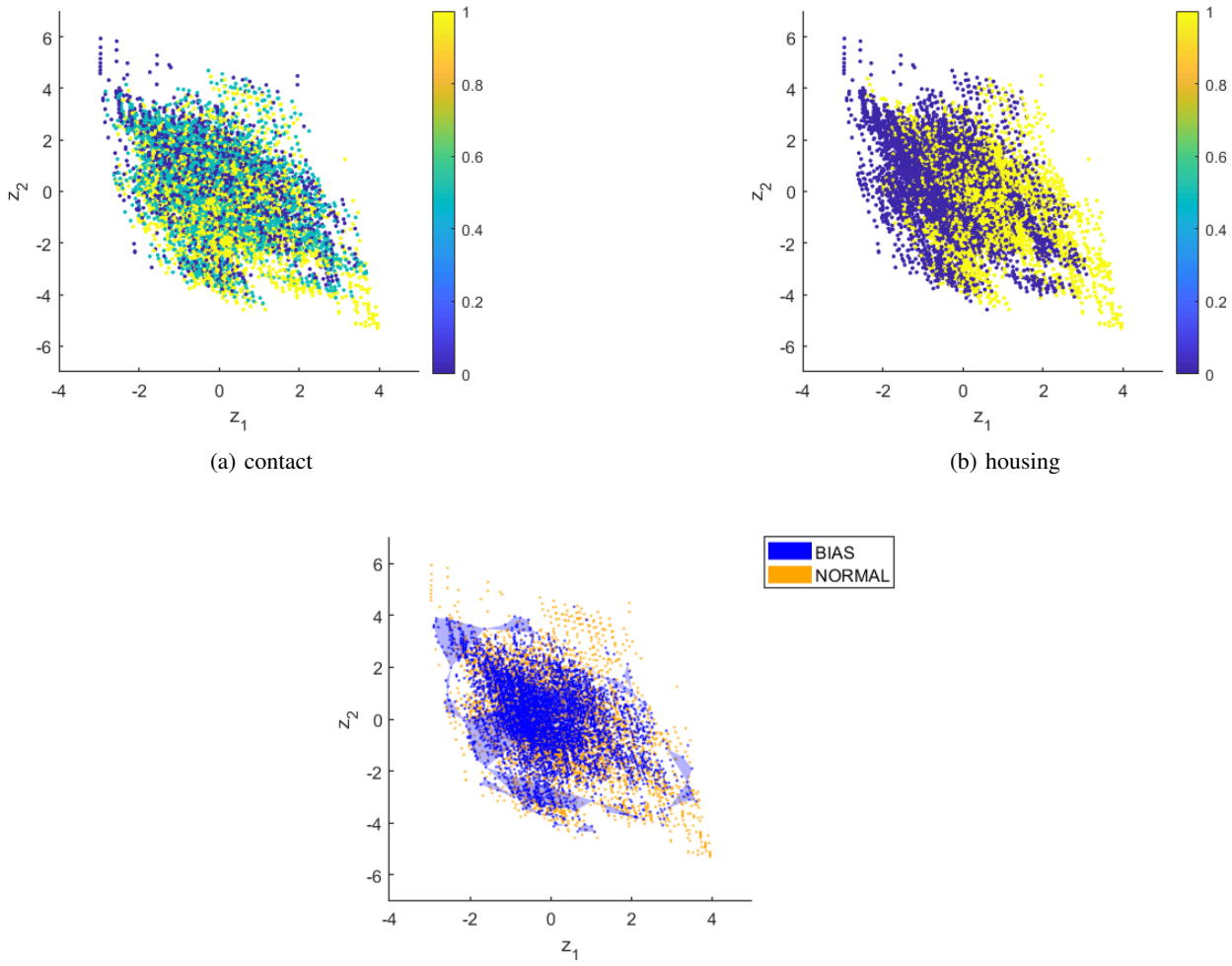
Dataset	KOSEI					KOSEI ISA-Directed				
	TSN	DSN	DSS	Time	SUR	TSN	DSN	DSS	Time	SUR
Census (gender)	33,394	19,863	0.0101	435.57	59.55%	**33,473**	**20,462**	**0.0100**	**434.40**	**60.39%***
Bank (age)	256,828	159,559	0.0028	450.58	62.12%	**255,556**	**161,165**	**0.0028**	**447.35**	**63.07%***

(b) Aequitas comparison

Dataset	Aequitas					Aequitas ISA-Directed				
	TSN	DSN	DSS	Time	SUR	TSN	DSN	DSS	Time	SUR
Census (gender)	6,904	1,955	0.1428	266.66	28.01%	**7,652**	**2,055**	**0.1532**	**300.55**	**26.67% †**
Bank (age)	13,234	6,834	0.2682	1,788.58	51.71%	**13,120**	**6,725**	**0.2533**	**1,686.69**	**51.36% †**

(c) ExpGA comparison

Dataset	ExpGA					ExpGA ISA-Directed				
	TSN	DSN	DSS	Time	SUR	TSN	DSN	DSS	Time	SUR
Census (gender)	72,101	50,909	0.0101	435.57	70.61%	**46,777**	**38,240**	**0.0100**	**434.40**	**81.75%***
Bank (age)	75,501	43,218	0.0085	433.70	57.24%	**72,237**	**43,624**	**0.0113**	**433.46**	**60.39%***



(a) contact

(b) housing

(c) discriminatory outcome

Fig. 6: Distribution of contact and housing features in the instance space from Aequitas RF Bank. (c) shows discriminatory instances within theoretical boundaries

TABLE IV: Comparison of test input generation for DT model

(a) KOSEI comparison

Dataset	KOSEI					KOSEI ISA-Directed				
	TSN	DSN	DSS	Time	SUR	TSN	DSN	DSS	Time	SUR
Census (gender)	1,940	1,254	0.0014	4.02	41.28%	**2,660**	**1,851**	**0.0013**	**3.53**	**45.79%**
Bank (age)	65,732	42,733	0.0003	14.47	64.94%	**65,878**	**42,979**	**0.0003**	**14.40**	**65.20%**

(b) Aequitas comparison

Dataset	Aequitas					Aequitas ISA-Directed				
	TSN	DSN	DSS	Time	SUR	TSN	DSN	DSS	Time	SUR
Census (gender)	1,789	383	0.0187	6.74	11.74%	**2,361**	**395**	**0.0305** †	**8.55** †	**13.38%**
Bank (age)	55,760	5,086	0.0373	181.29	9.42%	**49,437**	**8,438**	**0.0524** †	**410.14** †	**17.57%***

(c) ExpGA comparison

Dataset	ExpGA					ExpGA ISA-Directed				
	TSN	DSN	DSS	Time	SUR	TSN	DSN	DSS	Time	SUR
Census (gender)	75,641	50,279	0.0009	45.72	66.47%	**72,350**	**48,301**	**0.0009**	**45.11**	**66.76%**
Bank (age)	72,151	49,929	0.0007	32.50	69.20%	**68,851**	**47,901**	**0.0006**	**29.52**	**69.57%***

such as Aequitas testing the RF classifier under both datasets in Table IIIb, underscored by the † symbol. These results, which contradict the general trend that bias-revealing features at least increase SUR performance, are statistical anomalies. Further testing is necessary to determine the precise causes of these deviations.

E. Limitations in Experimental Results

The experimental setup faced analytical limitations due to the computational constraints encountered when handling large volume of instances generated by the fairness testing algorithms, particularly from ExpGA and KOSEI. On average, these algorithms generated 74,274 samples, a volume that proved too substantial for comprehensive analysis using the full capabilities of Instance Space Analysis (ISA).

The SIFTED and PILOT stages of ISA, which are critical for feature selection and generating the instance space, could not be conducted within the available computational resources due to the excessive memory demands. As a result, we placed a maximum limit on the number of samples analysed to 30,000. This restriction significantly reduced the scope of our analysis as more than half of the samples generated from KOSEI and ExpGA are removed from the instance space we intended to examine.

This limitation likely impacted our ability to discern linear relationships between non-sensitive features and the discriminatory label, potentially contributing to the low correlation scores observed in the samples generated by KOSEI and ExpGA, as depicted in Figures 1 and 2. The inability to fully utilise ISA's analytical potential may have concealed important insights into the behavior of these features within the model's decision-making process.

V. Future Work & Considerations

The evaluation of our ISA bias-revealing feature analysis and input test generation enhancement has highlighted several opportunities for further research and improvement. While our approach has demonstrated potential, it is evident that significant advancements are necessary in both the detection of bias-revealing features and their effective implementation. A potential research pathway is leveraging the unique properties of each test input generation mechanism to to uncover more nuanced and effective applications of bias-revealing features in fairness testing.

A. Expanding Dataset Utilisation

Our study validated the bias-revealing feature approach using two benchmark datasets, which led to mixed results in performance enhancements across state-of-the-art test input generation mechanisms. This variability in performance can be potentially attributed to the absence of a linear relationship between non-sensitive attributes and the discriminatory outcomes observed in the Census and Bank datasets. To broaden the scope of our research, it is advisable to evaluate additional fairness testing benchmark datasets. One such dataset, the Mep16 dataset, which is used for predicting healthcare needs [14], could provide valuable insights. Analysing ISA's capability to elucidate the relationships and interactions among non-sensitive attributes may offer a deeper understanding of potential biases common in the healthcare sector.

B. Exploring Alternative Fairness Testing Methods

Further investigations are also warranted to determine whether bias-revealing inputs can enhance test input generation methods beyond the two-phase search-based approach. For instance, the RULER fairness testing algorithm by Tao et al., which assesses causal fairness through perturbations on both sensitive and non-sensitive attributes [24], presents a promising alternative. Since non-sensitive attributes have direct influence on the biased outcome based on RULER's fairness metric, we can potentially observe more significant non-sensitive features that induce bias from ISA analysis. This algorithm also facilitates testing on Deep Neural Networks (DNNs), which are increasingly employed in the healthcare domain [13]. ISA's analytical depth could potentially reveal how DNNs may inherently foster bias through their internal mechanisms.

C. Leveraging the Iterative Nature of ISA

Moreover, the iterative nature of ISA provides a systematic framework for continuous improvement in fairness testing. By enabling repeated rounds of testing, ISA allows for the ongoing refinement of test inputs and the exploration of new bias-revealing features in successive iterations of the testing lifecycle. This not only enhances the adaptability of the testing process but also deepens the comprehensiveness of fairness assessments over time.

VI. Conclusion

This research has advanced the understanding of model unfairness by innovatively leveraging non-sensitive attributes to reveal and explain biased behavior in classifier models. Utilising ISA, our study identified some impactful features and visualised them within a 2D instance space. This visualisation allowed a direct comparison of fairness weaknesses across classifiers and highlighted specific areas within the instance space with a heightened likelihood of generating discriminatory samples.

Furthermore, we developed and implemented a novel method to enhance fairness testing algorithms. This method guides the test input generation processes to prioritise significant bias-revealing features identified by ISA. Our extensive testing yielded promising results, demonstrating substantial improvements in the effectiveness of fairness testing. 61% of our comparative tests showed a significant statistical increase in both the success rate of generating discriminatory samples and the total number of discriminatory samples generated.

Despite these advancements, our methodology encountered limitations in balancing effectiveness with efficiency. The additional time required to evaluate a larger pool of discriminatory instances has proven to be a considerable drawback in average execution time of fairness testing. These findings suggest that

while our approach enhances the success rate of identifying bias, it does so at the cost of increased computational effort and time.

Additionally, the experimental results highlighted a lack of linear relationships between non-sensitive features and discriminatory outcomes. Given these challenges and the critical insights gained, there is a clear need for further research. Future work should focus on refining the application of bias-revealing features in fairness testing to overcome the current limitations. This future research will be crucial in developing more nuanced and efficient methods for detecting and mitigating bias in machine learning models.

VII. References

[1] Md Manjurul Ahsan and Zahed Siddique. Machine learning-based heart disease diagnosis: A systematic literature review. *Artificial Intelligence in Medicine*, 128:102289, 2022.

[2] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[3] S. A. Berkowitz, C. Y. Traore, D. E. Singer, and S. J. Atlas. Evaluating area-based socioeconomic status indicators for monitoring disparities within health care systems: results from a primary care network. *Health Services Research*, 50(2):398–417, 2015.

[4] Zhenpeng Chen, Jie M Zhang, Max Hort, Federica Sarro, and Mark Harman. Fairness testing: A comprehensive survey and analysis of trends. *arXiv preprint arXiv:2207.10223*, 2022.

[5] Elena Dumitrescu, Sullivan Hué, Christophe Hurlin, and Sessi Tokpavi. Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, 297(3):1178–1192, 2022.

[6] Ming Fan, Wenying Wei, Wuxia Jin, Zijiang Yang, and Ting Liu. Expga. https://github.com/waving7799/ExpGA.git, 2022.

[7] Ming Fan, Wenying Wei, Wuxia Jin, Zijiang Yang, and Ting Liu. Explanation-guided fairness testing through genetic algorithm. In *Proceedings of the 44th International Conference on Software Engineering*, pages 871–882, 2022.

[8] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint meeting on foundations of software engineering*, pages 498–510, 2017.

[9] Bertrand K Hassani. Societal bias reinforcement through machine learning: a credit scoring perspective. *AI and Ethics*, 1(3):239–247, 2021.

[10] Rayyan Azam Khan, Yigang Luo, and Fang-Xiang Wu. Machine learning based liver disease diagnosis: A systematic review. *Neurocomputing*, 468:492–509, 2022.

[11] Danielle Li, Lindsey R Raymond, and Peter Bergman. Hiring as exploration. Technical report, National Bureau of Economic Research, 2020.

[12] Minghua Ma, Zhao Tian, Max Hort, Federica Sarro, Hongyu Zhang, Qingwei Lin, and Dongmei Zhang. Enhanced fairness testing via generating effective initial individual discriminatory instances. *arXiv preprint arXiv:2209.08321*, 2022.

[13] Pawan Kumar Mall, Pradeep Kumar Singh, Swapnita Srivastav, Vipul Narayan, Marcin Paprzycki, Tatiana Jaworska, and Maria Ganzha. A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities. *Healthcare Analytics*, 4:100216, 2023.

[14] The Mep16 dataset. https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-192, 2016. Accessed: 2024-05-25.

[15] S. Moro, P. Rita, and P. Cortez. Bank Marketing. UCI Machine Learning Repository, 2012. DOI: https://doi.org/10.24432/C5K306.

[16] Neelofar Neelofar, Kate Smith-Miles, Mario Andrés Muñoz, and Aldeida Aleti. Instance space analysis of search-based software testing. *IEEE Transactions on Software Engineering*, 49(4):2642–2660, 2023.

[17] Drago Plecko and Elias Bareinboim. Causal fairness analysis. *arXiv preprint arXiv:2207.11385*, 2022.

[18] Shinya Sano, Takashi Kitamura, and Shingo Takada. An efficient discrimination discovery method for fairness testing. In *SEKE*, pages 200–205, 2022.

[19] Shinya Sano, Takashi Kitamura, and Shingo Takada. Kosei. https://github.com/sskeiouk/KOSEI.git, 2022.

[20] Kate Smith-Miles, Jeffrey Christiansen, and Mario Andrés Muñoz. Revisiting where are the hard knapsack problems? via instance space analysis. *Computers & Operations Research*, 128:105184, 2021.

[21] Kate Smith-Miles and Mario Andrés Muñoz. Instancespace. https://github.com/andremun/InstanceSpace.git, 2019.

[22] Kate Smith-Miles and Mario Andrés Muñoz. Instance space analysis for algorithm testing: Methodology and software tools. *ACM Computing Surveys*, 55(12):1–31, 2023.

[23] Simon Strassl and Nysret Musliu. Instance space analysis and algorithm selection for the job shop scheduling problem. *Computers & Operations Research*, 141:105661, 2022.

[24] Guanhong Tao, Weisong Sun, Tingxu Han, Chunrong Fang, and Xiangyu Zhang. Ruler: discriminative and iterative adversarial training for deep neural network fairness. In *Proceedings of the 30th acm joint european software engineering conference and symposium on the foundations of software engineering*, pages 1173–1184, 2022.

[25] Guido Vittorio Travaini, Federico Pacchioni, Silvia Bellumore, Marta Bosia, and Francesco De Micco. Machine learning and criminal justice: A systematic review of advanced methodology for recidivism risk prediction. *International journal of environmental research and public health*, 19(17):10594, 2022.

[26] Sakshi Udeshi, Pryanshu Arora, and Sudipta Chattopadhyay. Aequitas. https://github.com/sakshiudeshi/Aequitas.git, 2018.

[27] Sakshi Udeshi, Pryanshu Arora, and Sudipta Chattopadhyay. Automated directed fairness testing. In *Proceedings of the 33rd ACM/IEEE international conference on automated software engineering*, pages 98–108, 2018.

[28] B Venkatesh and J Anuradha. A review of feature selection and its methods. *Cybernetics and information technologies*, 19(1):3–26, 2019.

[29] Yu Yang, Aayush Gupta, Jianwei Feng, Prateek Singhal, Vivek Yadav, Yue Wu, Pradeep Natarajan, Varsha Hedau, and Jungseock Joo. Enhancing fairness in face detection in computer vision systems by demographic bias mitigation. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 813–822, 2022.

[30] Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. Towards fair classifiers without sensitive attributes: Exploring biases in related features. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1433–1442, 2022.