Toward a Mutation-based Machine Learning Fairness Testing Approach to Uncover Bias Inducing Data Relations

ALFONSO CANNAVALE, University of Salerno, Italy
GIANMARIO VORIA, University of Salerno, Italy
ANTONIO DELLA PORTA, University of Salerno, Italy
GEMMA CATOLINO, University of Salerno, Italy
FABIO PALOMBA, University of Salerno, Italy

1 CONTEXT OF STUDY AND BACKGROUND

Machine Learning (ML)-enabled systems, i.e., software systems that include at least one component powered by ML algorithms [1], have found deployment across various critical domains, showcasing their potential efficacy and capabilities through recent applications in decision-making scenarios such as loan management [2] or hiring decisions [3].

Nevertheless, *every coin has a flip side*: employing ML-enabled solutions without being cautious of their implication poses a considerable risk. Numerous prior inquiries into the *ethical* dimensions of these systems, such as (1) discrimination against black people in medical costs previsioning [4], (2) biased evaluation of black people in criminal recidivism estimation [5], and (3) women discrimination in automated recruiting [6], have exposed their susceptibility to ML *fairness* issues [7]. In decision-making, *fairness* is the absence of prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics [7]. These worries often come from ML's reliance on data, which can make algorithms adopt biased assumptions leading to unfair results [8]. It is therefore relevant for the research community to define standards to treat *fairness* of ML systems as a first-class non-functional requirement [9].

Testing ML systems presents unique challenges due to their nature [10]. Unlike traditional systems, ML models are data-driven, meaning their decision-making process is derived from training data through algorithms, leading to behavior that may evolve as new data is provided. This dynamic nature contrasts with traditional systems, whose behavior remains relatively stable [11]. Therefore, testing these systems poses complicated challenges, particularly concerning the oracles [12]: since their task is to provide answers to questions for which no previous answer exists, it is difficult to establish a reference point for testing. Moreover, in the context of fairness testing, we face different definitions or metrics to measure fairness, some of which may also conflict with each other or be mathematically impossible to satisfy simultaneously [13]. This diversity of definitions makes the design of testing techniques a significant problem, as different definitions may require different testing oracles.

Recently, Chen et al. [14] summarized the works on fairness testing in a comprehensive survey. Examples are the work by Galhotra et al. [15] named THEMIS, which uses automated test suite generation by randomly perturbating the

Authors' Contact Information: Alfonso Cannavale, a.cannavale7@studenti.unisa.it, University of Salerno, Fisciano, Salerno, Italy; Gianmario Voria, gvoria@unisa.it, University of Salerno, Fisciano, Salerno, Italy; Antonio Della Porta, adellaporta@unisa.it, University of Salerno, Fisciano, Salerno, Italy; Gemma Catolino, gcatolino@unisa.it, University of Salerno, Fisciano, Salerno, Italy; Fabio Palomba, fpalomba@unisa.it, University of Salerno, Fisciano, Salerno, Italy.

input attributes, or the work by Aggarwal et al. [16], a black-box testing approach combining symbolical execution and local explainability to verify the fairness level of a model.

Chen et al. identified more than 100 research articles discussing different testing approaches in their survey. According to their findings, *mutation approaches* are one of the two main methods used to identify testing oracles. Mutation testing is a technique used to evaluate the effectiveness of a test suite by introducing small changes—mutations—to the source code and checking whether the tests can detect these changes. Concerning fairness testing, prior research primarily applies fairness-oriented metamorphic transformations to ML software's input or training data, aiming for these transformations to maintain the predictions unchanged or produce anticipated alterations.

2 OBJECTIVE OF THE STUDY

In mutation testing approaches for ML classification tasks, operators are applied on the *features* of the dataset [14]. However, these mutations usually concern the possible *values* of these attributes, i.e., the perturbations aim at modifying the instances of the data in order to test the outcome. Regarding tabular data, this involves examining pairs of instances possessing distinct sensitive attributes yet similar non-sensitive attributes, anticipating that they should result in identical classification outcomes. Hence, the oracle is defined as the expected *equity* in the prediction concerning the mutated samples. Nevertheless, the datasets employed in the training phase of ML systems often contain hidden information in the nuanced relationship among the values assumed by the features. This consideration drives the objective of our research.

> **Our Objective**
>
> *First, we aim to characterize a set of conditions that lead attribute relations to induce some form of bias in the data. Afterward, we aim to design and develop a set of mutation operators that act on these variables to test the fairness level of the model trained on such data. Finally, we will create an automated mutation testing approach using these operators.*

3 PLANNED METHOD

These variables, which we will define as *bias-inducing attribute relations*, will be computed starting from different definitions of fairness metrics such as Statistical Parity (SP), Equalized Odds (EO), and Average Odds (AO). [17, 18]. We will empirically evaluate different relations within the dataset and examine their impact on *fairness*. With these pieces of information, we will be able to define *testing oracles* based on the value of the variables.

One example is the *covariance* variable, which, when calculated between a sensitive attribute (e.g., race) and a non-sensitive attribute (e.g., income), may reveal patterns of association that could potentially reflect biased decision-making or systemic inequalities in the dataset.

After defining a set of relations, we will design mutation operators following guidelines defined in the literature [14]. These operators will act directly on the datasets' feature but with the purpose of modifying *relations' values*. This process will be supported by the knowledge of the impact of these relations, i.e., their positive or negative impact on fairness with a lower or a higher value. We plan to experiment with the development of the mutation operators with (1) static techniques and (2) the use of LLMs and prompt engineering. Finally, we will evaluate the performances of our testing approach by experimenting with different widely-employed datasets and for different ML tasks elicited from Fabris et al.'s ontology [19]. Moreover, we plan to compare our results with existing fairness testing approaches, such as Themis [15], to evaluate the efficacy of our approach in the identification of fairness issues.

REFERENCES

[1] Silverio Martínez-Fernández et al. "Software Engineering for AI-Based Systems: A Survey". In: *ACM Transactions on Software Engineering and Methodology* 31.2 (2022). ISSN: 1557-7392. DOI: 10.1145/3487043. URL: http://dx.doi.org/10.1145/3487043.

[2] Parmy Olson. "The algorithm that beats your bank manager". In: *CNN Money March* 15 (2011).

[3] Claire Cain Miller. "Can an algorithm hire better than a human". In: *The New York Times* 25 (2015).

[4] Ziad Obermeyer et al. "Dissecting racial bias in an algorithm used to manage the health of populations". In: *Science* 366.6464 (2019), pp. 447–453.

[5] Julia Angwin and Jeff Larson. *Machine Bias - There's software used across the country to predict future criminals. And it's biased against blacks.* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed: 2022-03-29. 2016.

[6] Jeffrey Dastin. *Amazon scraps secret AI recruiting tool that showed bias against women.* https://cutt.ly/VKWLqF1. Accessed: 2022-03-29. 2018.

[7] Ninareh Mehrabi et al. "A survey on bias and fairness in machine learning". In: *ACM Computing Surveys (CSUR)* 54.6 (2021), pp. 1–35.

[8] Solon Barocas, Moritz Hardt, and Arvind Narayanan. "Fairness in machine learning". In: *Nips tutorial* 1 (2017).

[9] Yuriy Brun and Alexandra Meliou. In: *ESEC/FSE'18*. 2018, pp. 754–759.

[10] Jie M. Zhang et al. "Machine Learning Testing: Survey, Landscapes and Horizons". In: *IEEE Transactions on Software Engineering* 48.1 (2022), pp. 1–36. DOI: 10.1109/TSE.2019.2962027.

[11] Saleema Amershi et al. "Software Engineering for Machine Learning: A Case Study". In: *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. 2019, pp. 291–300. DOI: 10.1109/ICSE-SEIP.2019.00042.

[12] Earl T. Barr et al. "The Oracle Problem in Software Testing: A Survey". In: *IEEE Transactions on Software Engineering* 41.5 (2015), pp. 507–525. DOI: 10.1109/TSE.2014.2372785.

[13] Suvodeep Majumder et al. "Fair Enough: Searching for Sufficient Measures of Fairness". In: *ACM Trans. Softw. Eng. Methodol.* 32.6 (2023). ISSN: 1049-331X. DOI: 10.1145/3585006. URL: https://doi.org/10.1145/3585006.

[14] Zhenpeng Chen et al. *Fairness Testing: A Comprehensive Survey and Analysis of Trends*. 2024. arXiv: 2207.10223 [cs.SE].

[15] 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*. ESEC/FSE 2017. Paderborn, Germany: Association for Computing Machinery, 2017, 498–510. ISBN: 9781450351058. DOI: 10.1145/3106237.3106277. URL: https://doi.org/10.1145/3106237.3106277.

[16] Aniya Aggarwal et al. "Black Box Fairness Testing of Machine Learning Models". In: ESEC/FSE 2019. Association for Computing Machinery, 2019.

[17] Moritz Hardt, Eric Price, and Nathan Srebro. *Equality of Opportunity in Supervised Learning*. 2016. arXiv: 1610.02413 [cs.LG].

[18] Cynthia Dwork et al. *Fairness Through Awareness*. 2011. arXiv: 1104.3913 [cs.CC].

[19] Alessandro Fabris et al. "Algorithmic fairness datasets: the story so far". In: *Data Mining and Knowledge Discovery* 36.6 (Sept. 2022), 2074–2152. ISSN: 1573-756X. DOI: 10.1007/s10618-022-00854-z. URL: http://dx.doi.org/10.1007/s10618-022-00854-z.