open proceedings

MIMOSA: A Tool for Fairness Exploration Through Explanations

Vasiliki Papanikou
University of Ioannina
Archimedes/Athena RC
Athens, Greece
v.papanikou@athenarc.gr

Danae Pla Karidi
Archimedes/Athena RC
Athens, Greece
danae@athenarc.gr

Evaggelia Pitoura
University of Ioannina
Archimedes/Athena RC
Athens, Greece
pitoura@uoi.gr

Emmanouil Panagiotou
Freie Universität, Berlin
Universität der Bundeswehr
München, Germany
emmanouil.panagiotou@fu-berlin.de

Eirini Ntoutsi
Universität der Bundeswehr
München, Germany
eirini.ntoutsi@unibw.de

Abstract

As Artificial Intelligence (AI) is increasingly used in areas that impact human lives, concerns about fairness and transparency have grown, especially for protected groups. To better understand such concerns, explainability techniques can be leveraged not only for model interpretation but also to assess potential biases. The MIMOSA[1] tool utilizes both individual and group explanation methods as bias detectors. It allows users to compare group fairness metrics with explanation findings, identify which features contribute to biased outcomes, visualize explanations through multiple perspectives and apply fairness interventions while tracking how feature contributions change. The tool is designed to be accessible to a wide audience of users, including sociologists, domain experts and machine learning practitioners.

Keywords

Explainable AI, algorithmic fairness

1 Introduction

As AI becomes embedded in critical domains such as healthcare and education, concerns about fairness and transparency have grown, especially for protected groups defined by gender or race. For example, AI-based hiring systems have been shown to disadvantage women, reflecting historical inequalities [5]. Explainable AI (XAI) [10] can be used as a critical tool for tackling these challenges, by shedding light into the relationships between protected attributes and target outcomes. To support such exploration, we present MIMOSA, a tool that utilizes explanation methods as bias detectors through an interactive pipeline.

Related work. Explanations have been used to uncover unfair model behavior before (e.g., see [6] for a survey). For example, LimeOut [2] uses LIME to uncover and mitigate bias, while SHAP-based approaches measure demographic parity as SHAP differences between protected and non-protected groups [1] or interpret reduced SHAP attributions as improved fairness after mitigation [3]. Regarding counterfactual explanations, PreCoF [8] identifies unfairness by analyzing feature changes across negatively classified groups, while FACEGroup [7] generates group counterfactuals and proposes fairness metrics. In addition, data-based methods [16, 18] have also been proposed for detecting bias. All such approaches focus on using specific explanation methods without a standardized pipeline posing questions about their generality [4].

Also, several open-source libraries exist for fairness[2] and explainability[3] independently, but there is a lack of tools that integrate explanation techniques directly for fairness analysis.

To address the lack of a unifying approach, MIMOSA[4], based on our work in [15], provides an interactive tool that follows a clearly defined pipeline for fairness evaluation while offering flexibility to explore different configurations within each stage of the workflow.

Users can define focus groups using protected attributes and select specific outcome slices (e.g., true positives, false negatives). Several *group fairness* metrics can be examined alongside explanations to understand where unfairness originates, whether from protected features or proxy attributes (i.e., attributes correlated with the protected ones) with different contributions across groups. MIMOSA supports multiple explanation methods and configurable explanation settings. Each explanation method offers a different perspective of the model behavior. To address the concerns about the robustness and consistency of explanation techniques, the tool enables users to evaluate explanation quality using the AOPC metric and select the most appropriate method. MIMOSA also offers diverse visualization options to examine bias from different perspectives and to address the fact that the way information is presented can influence fairness perception [11]. Furthermore, MIMOSA incorporates fairness mitigation, allowing users to apply interventions, compare metrics and explanations before and after mitigation and observe how feature contributions shift.

2 Explanations as Bias Detectors

Algorithmic fairness ensures that model outcomes do not systematically disadvantage individuals based on sensitive characteristics. These characteristics, known as protected attributes, often include gender, race, age, and sexual orientation. *Group fairness* evaluates whether the predictions and errors of a model are equitably distributed across groups defined by protected attributes. Model outcomes can be described using true positives (TP), true negatives (TN), false positives (FP), and false negatives

[1] The mimosa flower symbolizes purity, innocence and sensitivity, values that are essential in the pursuit of truth, justice and fairness.

[2] https://github.com/Trusted-AI/AIF360
[3] https://github.com/Trusted-AI/AIX360
[4] Source code: GitHub Video presentation: Video

EDBT '26, Tampere (Finland)

10.48786/edbt.2026.67

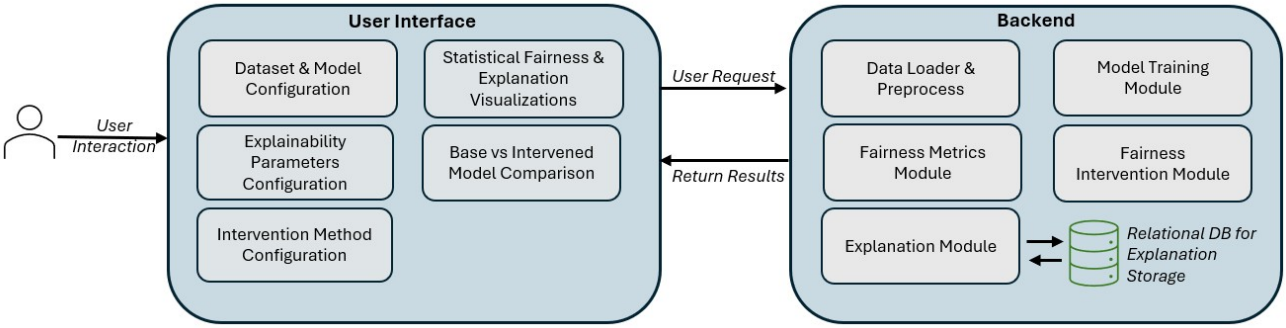


Figure 1: The MIMOSA tool architecture.

(FN). *Group fairness* metrics include *demographic parity*, which assesses whether the overall positive prediction rate is the same across different protected groups, *equal opportunity* that focuses on equality of the TP rate across groups, and *predictive equality* that focuses on FP rate.

MIMOSA offers results on *group fairness* metrics and users can then explore whether fairness disparities arise because the model relies heavily on a protected attribute or on proxy features correlated with it. This examination of how features are used in the decision-making process is also referred to as *process fairness* [9]. To support this analysis, MIMOSA integrates post-hoc, black-box explanation methods that clarify model behavior after training without requiring access to internal model details. Users can define groups based on one or more protected attributes, focus on specific outcome slices (e.g., TP, FP) and compare feature contributions across groups. This enables the exploration of how the model uses features when it succeeds versus when it fails, revealing different fairness insights. For example, explanations for FP can show why certain groups receive favorable predictions without meeting the criteria, while FN explanations reveal that some groups need stronger evidence for a positive outcome. MIMOSA provides both individual and group explanations to analyze model behavior on specific subpopulations.

For individual-level explanations, we provide attributional and counterfactual methods, offering complementary perspectives [12]. Attribution-based methods indicate which features influenced a decision, capturing the sufficiency of feature values. Counterfactual methods show the minimal changes needed to alter a prediction, capturing the necessity and cost of feature changes. For attribution based methods we use LIME (Local Interpretable Model-agnostic Explanations) [17] and SHAP (SHapley Additive exPlanations) [13]. For individual counterfactual explanations, we incorporate DiCE (Diverse Counterfactual Explanations) [14] and at the group level, we utilize FACEGroup [7], a group counterfactual method which identify how a group of instances, could alter their features to achieve favorable outcomes.

Users can apply fairness interventions and assess their impact by comparing model performance and *group fairness* before and after. Explanations reveal how feature contributions change and whether the influence of protected attributes shifts to proxy features. Supported interventions include removing the protected attribute and adversarial debiasing [19].

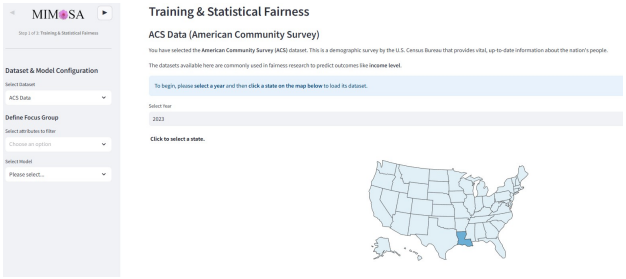


Figure 2: Interactive state map for ACS PUMS data.

3 System Architecture

The system architecture in Figure 1 consists of a backend engine that performs data loading, model training, *group fairness* computation, explanation generation, fairness interventions and a user interface that guides users through the fairness exploration workflow.

The backend is responsible for all core system operations. The Data Loader Module handles both preconfigured datasets and user-uploaded data, performing preprocessing, categorical encoding and optional removal of protected attributes for intervention scenarios. The Model Training Module that supports multiple predictive models trains the selected model using GridSearchCV and generates predictions. The Fairness Metrics Module computes group fairness metrics and reports disparities between the focus group and the rest of population including statistical significance tests. The Explanation Module provides different explanation types with parameter configuration, subgroup-specific explanations and computation of contribution-disparity summaries. Explanations are stored in a relational database to enable efficient access and comparisons. Finally, the Intervention Module implements fairness mitigation strategies and comparisons with the base model.

4 MIMOSA Functionality and User Interface

The functionality and interface of MIMOSA is organized into three views: (1) *Select Dataset, Model & Group Fairness Metrics*, (2) *Explain* and (3) *Fairness Interventions*.

In the first view, users choose from a set of datasets or upload their own. The predefined collection includes fairness benchmarks such as Adult, COMPAS, German Credit, and ACS PUMS data for the ACSIncome task. For ACS, users can select a survey year, interact with a state-level map (Figure 2) and download the corresponding subset, enabling fairness analysis across different

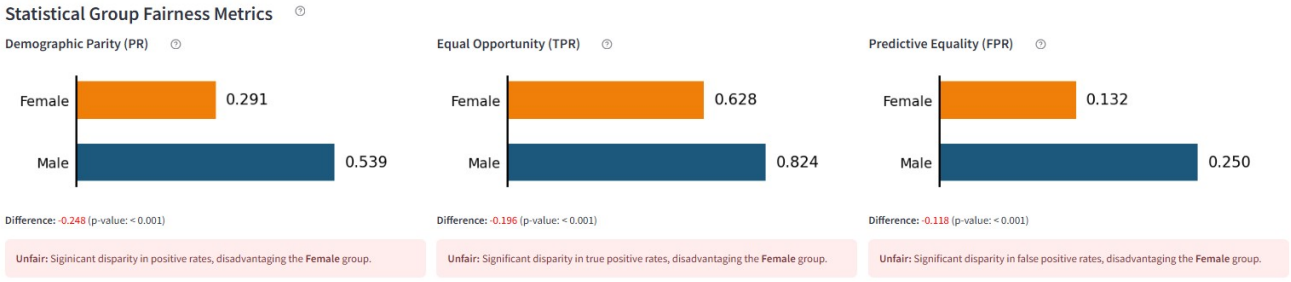
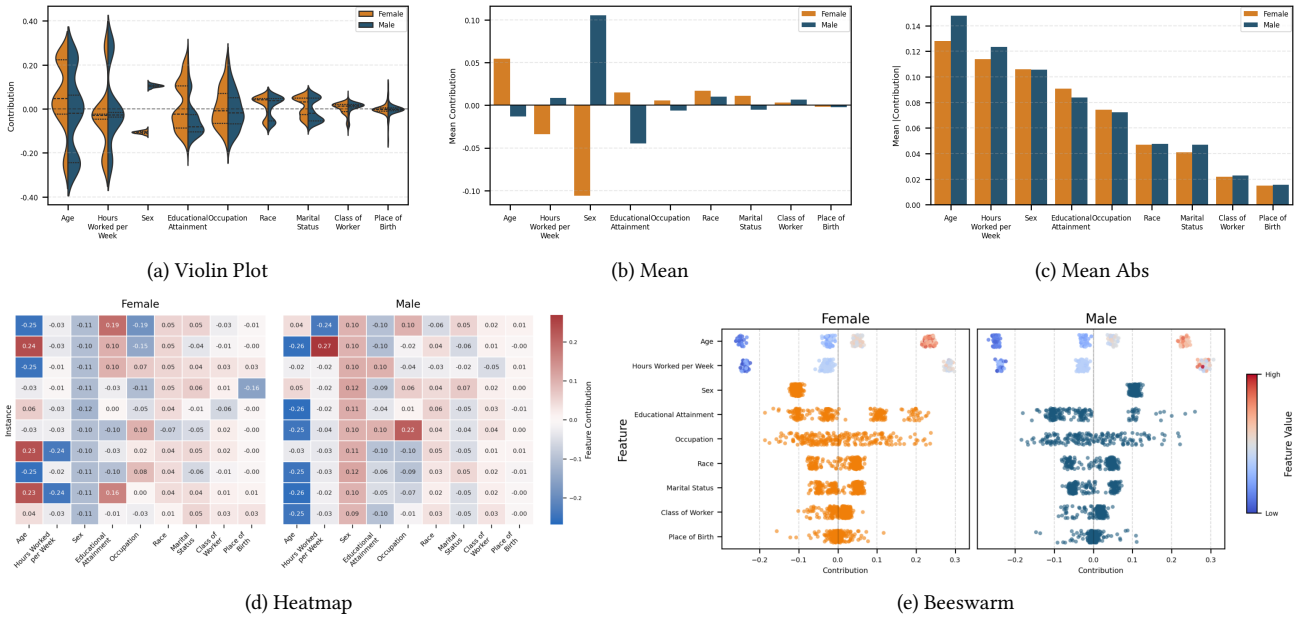


Figure 3: Fairness metrics among groups, reporting differences, statistical significance and fairness assessment messages.



(a) Violin Plot (b) Mean (c) Mean Abs

(d) Heatmap (e) Beeswarm

Figure 4: Explanation visualization options comparison across focus and rest groups.

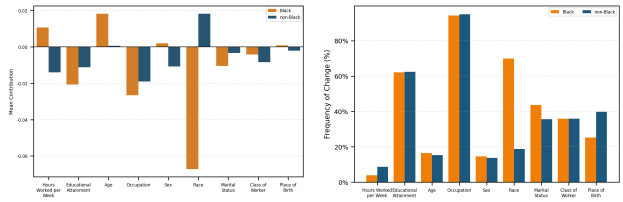
cultural, demographic, and political contexts. Users then define a focus group using one or more attributes. After selection, the interface provides a dataset preview, description, and basic group visualizations. Users select and train a predictive model, after which model performance and group fairness metrics are displayed. Differences between groups are reported with statistical significance tests: significant disparities appear in red as "Unfair", while non-significant ones appear in green as "Fair", as shown in Figure 3.

The *Explain* view helps users understand why bias occurs by analyzing how different features contribute to model predictions. Users first select the explanation method they want to apply. Next, users select the specific type of model outcome for which explanations should be generated (TP, TN, FP, FN).

We provide multiple visualization options. Figure 4 shows different visualizations for attribution methods. Violin plots (a) show the distribution of contributions per feature and group, helping users compare them easily. Mean (b) and mean abs. (c) plots provide aggregated views of contributions: the signed mean reveals the direction of contribution, while the absolute mean highlights the magnitude. Heatmaps (d) display contributions for selected individuals, due to the colors user can easily see patterns of positive and negative contributions per group. Beeswarm plots (e) combine contribution distributions with feature values, showing how the actual values of features relate to their contribution levels. For counterfactual explanations, users can see how often each feature must change to flip predictions and also the direction and magnitude of changes, with color indicating original values. Each visualization is paired with a numerical summary of contribution disparities.

Users can also compare attribution and counterfactual methods, since they offer complementary insights. For example, in Figure 5, SHAP indicates that race contributes differently for Black and non-Black groups, while counterfactual analysis shows that over 60% of Black individuals would need to change their race to receive a favorable outcome. The interface also supports selecting the most suitable explanation method using the AOPC metric, where higher scores indicate features with stronger influence on predictions.

In the *Fairness Interventions* view (Figure 6), users apply a mitigation strategy and compare the new model with the original. The system reports updated performance, *group fairness* metrics and explanations, allowing users to observe how the intervention affects fairness and model behavior.



(a) SHAP (b) DiCE

Figure 5: SHAP vs. DiCE across race groups.

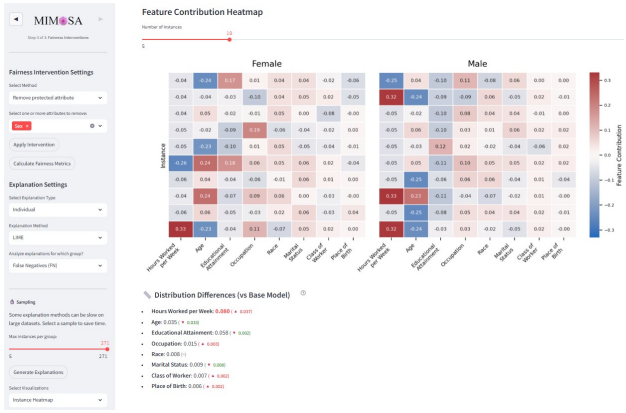


Figure 6: Fairness Intervention view.

5 Demonstration Outline

A typical user interaction include the following steps: (1) **Select Dataset and Focus Group.** User begins by selecting dataset. Then they define a focus group (e.g., *sex = female*) and inspect basic visualizations of the group distribution relative to the rest of the population. (2) **Train a Model and Inspect Fairness.** Users select a model to train. The system computes model performance metrics along with *group fairness* metrics and their disparities as shown in Figure 3 for the Louisiana 2023 ACS dataset and XGBoost model with female group as the focus group. (3) **Generate Explanations.** User explores the *Explain* view, selecting an explanation type, an explanation method and a specific subpopulation (e.g., FN). Then they can visualize how feature contributions differ between groups using a variety of visualizations as shown in Figure 4 or examine DiCE counterfactual changes. Across these visualizations, users can observe contrasting attribution patterns, for instance, sex contributes negatively for females and positively for males. MIMOSA also reports quantitative disparity summaries. (4) **Apply Fairness Interventions.** In the *Fairness Interventions* view (Figure 6) users apply mitigation strategies. MIMOSA automatically retrains the model and reports the updated performance, fairness metrics and explanations while each result is accompanied by a comparison with the baseline model.

6 Conclusion

In this work, we introduce MIMOSA, a fairness exploration tool that uses explanations to help users understand the root causes of bias in machine learning systems. Designed for a broad audience, it integrates multiple explanation methods, enables subgroup analysis and provides diverse visualizations. MIMOSA also supports fairness interventions and allows comparison of model performance, fairness metrics and explanation behavior before and after mitigation.

Acknowledgments

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

References

[1] Tom Begley, Tobias Schwedes, Christopher Frye, and Ilya Feige. 2020. Explainability for fair machine learning. *ArXiv* (2020).
[2] Vaishnavi Bhargava, Miguel Couceiro, and Amedeo Napoli. 2020. LimeOut: An Ensemble Approach to Improve Process Fairness. In *ECML PKDD Workshops*. Springer.
[3] Juliana Cesaro and Fábio Gagliardi Cozman. 2019. Measuring Unfairness Through Game-Theoretic Interpretability. In *ECML PKDD Workshops*.
[4] Luca Deck, Jakob Schoeffer, Maria De-Arteaga, and Niklas Kühl. 2024. A Critical Survey on Fairness Benefits of Explainable AI. In *FAccT '24*.
[5] Alessandro Fabris, Nina Baranowska, Matthew J Dennis, David Graus, Philipp Hacker, Jorge Saldivar, Frederik Zuiderveen Borgesius, and Asia J Biega. 2024. Fairness and bias in algorithmic hiring: A multidisciplinary survey. *ACM TIST* (2024).
[6] Christos Fragkathoulas, Vasiliki Papanikou, Danae Pla Karidi, and Evaggelia Pitoura. 2024. On Explaining Unfairness: An Overview. In *ICDEW '24*. IEEE.
[7] Christos Fragkathoulas, Vasiliki Papanikou, Evaggelia Pitoura, and Evimaria Terzi. 2025. FACEGroup: Feasible and Actionable Counterfactual Explanations for Group Fairness. In *ECML PKDD '25*.
[8] Sofie Goethals, David Martens, and Toon Calders. 2023. PreCoF: counterfactual explanations for fairness. *Machine Learning* (2023).
[9] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. 2016. The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making. In *NIPS ML & Law Workshop*.
[10] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM (CSUR)* (2018).
[11] Zhanna Kaufman, Madeline Endres, Cindy Xiong Bearfield, and Yuriy Brun. 2025. Your Model Is Unfair, Are You Even Aware? Inverse Relationship Between Comprehension and Trust in Explainability Visualizations of Biased ML Models. *arXiv preprint arXiv:2508.00140* (2025).
[12] Ramaravind Kommiya Mothilal, Divyat Mahajan, Chenhao Tan, and Amit Sharma. 2021. Towards unifying feature attribution and counterfactual explanations: Different means to the same end. In *AIES '21*.
[13] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *NeurIPS* (2017).
[14] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *FAccT '20*.
[15] Vasiliki Papanikou, Danae Pla Karidi, Evaggelia Pitoura, Emmanouil Panagiotou, and Eirini Ntoutsi. 2025. Explanations as Bias Detectors: A Critical Study of Local Post-hoc XAI Methods for Fairness Exploration. *arXiv:2505.00802* (2025).
[16] Romila Pradhan, Jiongli Zhu, Boris Glavic, and Babak Salimi. 2022. Interpretable data-based explanations for fairness debugging. In *SIGMOD '22*.
[17] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *KDD '16*.
[18] Tanmay Surve and Romila Pradhan. 2025. Explaining Fairness Violations using Machine Unlearning. In *EDBT '25*.
[19] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *AIES '18*.