Demonstration Paper

GLOVES: Global Counterfactual-based Visual Explanations

Panagiotis Gidarakos
Athena Research Center
Athens, Greece
pgidar@athenarc.gr

Nikolas Theologitis
Athena Research Center
Athens, Greece
n.theologitis@athenarc.gr

Stavros Maroulis
Athena Research Center
Athens, Greece
stavmars@athenarc.gr

Loukas Kavouras
Athena Research Center
Athens, Greece
kavouras@athenarc.gr

Giorgos Giannopoulos
Athena Research Center
Athens, Greece
giann@athenarc.gr

George Papastefanatos
Athena Research Center
Athens, Greece
gpapas@athenarc.gr

ABSTRACT

We introduce Global Counterfactual-based Visual Explanations (GLOVES), a visualization platform designed to enhance the explainability of decision-making systems through global counterfactual explanations (GCE). GLOVES focuses on visualizing global counterfactuals, calculated on top of a classifier's decisions on an examined population, enabling users to explore and compare different configurations of GCE algorithms interactively. The platform allows users to upload their own datasets and models or use preloaded options, providing comparative analyses of solution quality and explanation consistency across configurations. Through detailed visualizations, including diagrams and tables, users can examine counterfactual actions and their impact on populations. By integrating advanced visualization tools with global counterfactual methods, GLOVES supports deeper understanding, debugging, and refinement of decision-making systems.

1 INTRODUCTION

Counterfactual explanations are a powerful tool for understanding machine learning models, offering insights by answering "what-if" questions about how changes to input features can alter predictions. They empower users to interpret model decisions and identify actionable adjustments, such as increasing income to gain loan approval. Counterfactuals are categorized into local approaches, which focus on individualized recourse, and global approaches, which provide strategies for groups or entire populations.

A local counterfactual identifies a *minimum-cost* set of changes to an individual's features that can induce a desired outcome in the classifier's decision, such as transitioning from a "reject" to an "accept" decision (i.e., achieve recourse). Typically, counterfactuals are examined for the *affected* population, i.e., individuals who have received an adverse outcome. The *minimum-cost* objective can incorporate various considerations, such as different l_p-norms of the feature vector, feasibility, and validity of the counterfactual, depending on the requirements of the application.

In contrast, global counterfactuals focus on providing recourse to groups or entire populations, introducing challenges in balancing multiple objectives. These include minimizing the cost of feature changes, maximizing effectiveness—the proportion of the population achieving recourse—and maintaining interpretability by limiting the size of counterfactual action sets. A key concept is the *counterfactual action*, representing consistent feature changes applied to a subgroup of individuals. These actions are designed to be generalizable and interpretable, enabling all individuals within a subgroup to achieve recourse through a single strategy. Different notions of counterfactual actions are discussed in Section 2.

While considerable progress has been made in algorithmic methods for local counterfactual explanations and their visualization, the importance of global counterfactuals has only recently gained attention in the literature [7, 9–12, 15]. To the best of our knowledge, no frameworks currently exist for the dedicated visualization and analysis of global counterfactuals. Visual exploration [3, 19] can drastically improve the interpretability of global counterfactuals by enabling users to analyze the distribution of features across populations, apply dimensionality reduction techniques to better understand high-dimensional data, and interactively explore and refine counterfactual actions.

In this work, we introduce GLOVES, a *visualization platform* designed to facilitate the exploration and analysis of global counterfactual explanations. GLOVES takes as input a dataset and a trained machine learning classification model, and applies a global counterfactual algorithm to the model's predictions to generate global counterfactuals. Users can upload their own datasets and models or choose from preloaded options (datasets and trained classifiers) to run the algorithm. In this version, we support models from scikit-learn version 1.1.3 with Python 3.9.20, requiring them to be in a pickle format (version 4.0) for compatibility.

GLOVES main contribution is its *experimentation-driven analysis*, which empowers users to interactively explore and experiment with different configurations for global counterfactual generation. The platform offers *intuitive tools* to inspect and compare counterfactual actions, evaluate their population-wide impact, and refine solutions dynamically. Through *detailed visualizations* and *interactive interfaces*, GLOVES transforms the traditionally complex process of analyzing global counterfactuals into an *accessible and actionable experience*, making it an indispensable tool for researchers and practitioners.

Related Work. Visualization techniques are extensively utilized in explainable AI (XAI) to intuitively communicate complex model explanations. Local counterfactual explanations, which identify minimal changes to an instance's input to alter a model's prediction, have been widely explored through interactive visual systems. These tools integrate counterfactual methods with visual analytics frameworks to enhance interpretability and improve user understanding of predictive outcomes [1, 5]. They often allow users to analyze model behavior at the instance level, providing personalized and actionable insights [4, 5].

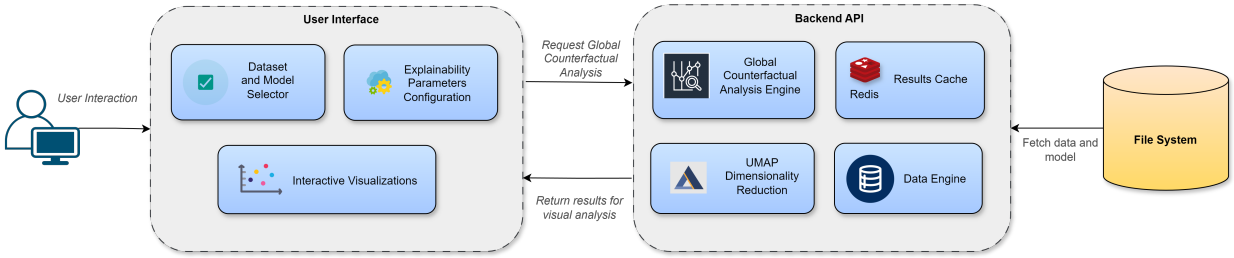

Figure 1: GLOVES: Framework Architecture

Recent research has emphasized the utility of counterfactual visualizations in exploratory data analysis. These visualizations help users avoid incorrect assumptions, identify confounding variables, improve causal inference, and support decision-making processes [8, 17]. Additionally, counterfactual reasoning has been applied to specialized domains such as education [2] and recommendation systems, where feature manipulation enhances user engagement and outcomes [6]. Furthermore, visualizing decision boundaries has enabled "what-if" analyses, offering deeper insights into model behavior [16].

2 GLOBAL COUNTERFACTUAL METHODS

Global counterfactual explanation (GCE) methods provide population-wide recourse strategies by offering solutions applicable across groups. These methods vary in their approaches and outputs, addressing priorities such as scalability, cost, and interpretability.

Rule-based methods like `AReS` [15] and `Fast AReS` [12] generate hierarchical rule sets that summarize recourse strategies, with minimal feature changes required for individuals in subgroups. Partition-based methods, such as `CET` [7], divide the feature space into partitions and assign a single counterfactual action to each, ensuring consistent recourse within each partition.

Clustering-based methods, such as `GLANCE` [9], group individuals into representative subgroups based on feature and action spaces and generate concise sets of counterfactual actions. Other approaches, such as `Group-CF` [18], aim to maximize the proportion of the population receiving recourse, while directional methods like `GLOBE-CE` [13] define recourse as movement along specific action directions in the feature space.

Evaluation of GCE methods typically involves three metrics:

- **Effectiveness:** Measures how well the counterfactual actions achieve the desired outcomes for the population, indicating the breadth of recourse provided.
- **Cost:** Assesses the magnitude of changes required to implement the recourse, reflecting the practicality and feasibility of the solutions.
- **Size (Interpretability):** Refers to the number of distinct actions or rules generated, with smaller sets being more interpretable and easier to communicate to stakeholders.

Our platform, GLOVES, is designed to explore, compare, and analyze global counterfactual explanation (GCE) methods. In its first version, we demonstrate its capabilities using the `GLANCE` algorithm [9], providing users with interactive visualizations and detailed analyses of counterfactual actions. The platform features an extensible architecture, allowing for the seamless integration of additional GCE methods in future updates while ensuring a consistent user experience and adaptable visualization tools.

The currently supported method for GCE generation, `GLANCE`, operates by clustering individuals in both feature and action spaces to identify representative subgroups, ensuring that counterfactual actions are meaningful and applicable to each group. For each cluster, it generates counterfactual actions that balance the key metrics of effectiveness, cost, and interpretability. The algorithm uses the l_1-norm (Manhattan distance) to quantify cost, measuring the magnitude of feature changes required for recourse. This approach, commonly adopted in global counterfactual methods, offers a straightforward and interpretable way to assess recourse across populations. However, as the l_1-norm does not account for the varying difficulty of modifying specific features, future updates will incorporate customizable cost functions, enabling users to assign weights to features based on context-specific factors.

3 GLOVES ARCHITECTURE

Figure 1 illustrates the architecture of the GLOVES system, which consists of a backend API for executing explainability methods and a web-based user interface for conducting counterfactual analyses and visualizing results. The frontend provides users with tools like the *Dataset and Model Selector* for selecting or uploading datasets and ML models, and the *Explainability Parameters Configuration* module for defining analysis scenarios and parameters. Once a dataset and model are selected, the frontend sends a request to the backend API to process the data.

The backend reads the data and interacts with the selected ML model to dynamically compute the predicted label for each instance and determine the affected population. Additionally, the system offers optional dimensionality reduction functionality using UMAP [14], which computes a low-dimensional representation of the data for visualization purposes.

Through the user interface, users can configure multiple global counterfactual analysis scenarios tailored to the selected dataset and model and compare their respective performance. The configurations are processed by the *Global Counterfactual Analysis Engine*, which computes counterfactual explanations for the given data and model. Additionally, the backend employs a lightweight *Results Cache* to store previously computed results, minimizing redundant computations and enhancing performance.

The computed explanations are returned to the frontend, where they are visualized interactively in the *Interactive Visualization and Insights* module, enabling users to explore the data, understand suggested actions, and analyze the impact of counterfactual explanations.

4 GLOVES USER INTERFACE

This section introduces the visual interface of GLOVES (Fig. 2 and 3), designed to facilitate the exploration and analysis of global counterfactual explanations. The interface is structured into three distinct views, *Select Dataset and Model*, *Explore Dataset* and *Analyze Counterfactuals*, each tailored to different stages of the analysis process. Users can upload their datasets and models or select from preloaded options, configure counterfactual analysis parameters, and explore results through intuitive visualizations.

The *Explore Dataset* view (Fig. 2) focuses on understanding the dataset and its features. Users can preview the dataset in a tabular format, with options to filter and sort attributes. Instances affected by adverse outcomes are highlighted, and users can toggle between affected and non-affected populations for focused exploration. Additionally, a scatter plot visualization allows users to analyze relationships between two selected features or apply dimensionality reduction techniques, such as UMAP, to understand high-dimensional data. These visualizations dynamically update based on user interactions, aiding in uncovering patterns and understanding the distribution of outcomes.

The *Analyze Counterfactuals* view supports experimentation with global counterfactual explanations. Users can define parameters for counterfactual analysis, such as the number of actions and selected features to calculate counterfactuals upon. Once configurations are set, the system computes results and presents metrics like total cost and effectiveness to enable comparisons across strategies. Results are displayed in both tabular and visual formats, with charts enabling users to visually compare configurations. Dynamic visualizations, including scatter plots and bar charts, illustrate cost-effectiveness trade-offs and changes in the population after applying the actions. For each configuration, users can explore detailed insights, such as the specific actions generated and their respective impacts on the affected population. Additionally, users can visually examine how the affected population changes after applying the suggested counterfactual actions. The interface also allows for more granular analysis by supporting the application of individual actions. Users can apply a single action across all affected instances to observe its isolated impact, aiding in deeper evaluation and refinement of counterfactual strategies.

Availability. The tool and its functionalities are available online at *http://gloves.imsi.athenarc.gr*. A video demonstration is available at *https://vimeo.com/1037427238*.

5 DEMONSTRATION OUTLINE

In this section, we outline our demonstration scenario. Attendees will interact with GLOVES to analyze datasets from various domains using global counterfactual explanations. For example, they can focus their analysis on datasets like the *HELOC (Home Equity Line of Credit)* dataset, which contains financial attributes used to predict the likelihood of an individual repaying a loan.

Initially, attendees will be introduced to the platform and its key components. They will then interact with the tool to perform the following operations:

- Select a dataset and a trained ML model (e.g., XGBoost) or upload their own data and models.
- Preview the dataset in a tabular format, explore relationships between features using scatter plots, and apply dimensionality reduction techniques (e.g., UMAP) to gain insights into the affected and non-affected populations.

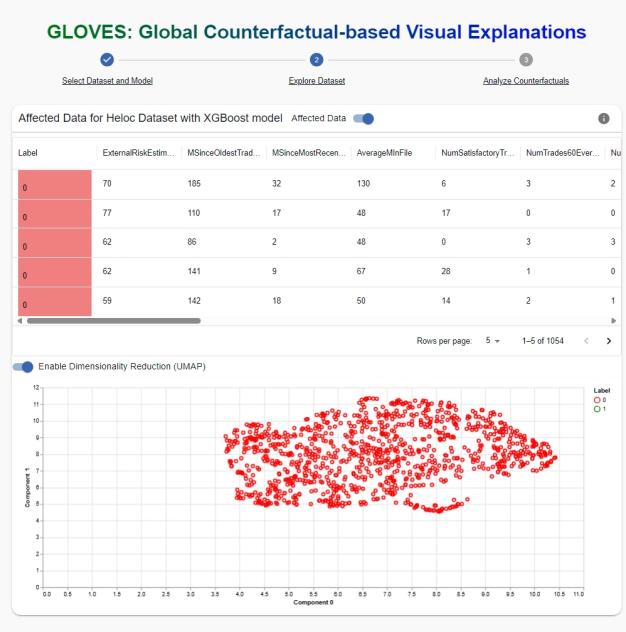

Figure 2: The *Explore Dataset* view in GLOVES, allowing users to preview the dataset, examine feature relationships via scatter plots, and analyze the affected population.

- Configure global counterfactual analysis by selecting parameter values, such as the number of actions or feature subsets to include in the analysis.
- Generate global counterfactual explanations and compare configurations based on metrics such as cost, effectiveness, and interpretability.
- Examine detailed counterfactual actions, assess their impact on the affected population, and visually inspect transformations in the feature space (Fig. 3).
- Perform fine-grained analysis by applying individual actions to the dataset to observe their isolated effects and evaluate their effectiveness.

By engaging with these scenarios, attendees will gain hands-on experience with GLOVES and understand how global counterfactual explanations can provide actionable and interpretable insights for improving decision-making systems.

6 CONCLUSION

In this work, we introduced GLOVES, a visualization platform for exploring global counterfactual explanations. By combining explainability algorithms with visualizations, GLOVES helps users analyze model behavior, evaluate fairness, and debug decision-making systems. Future developments will focus on extending GLOVES by incorporating additional global counterfactual algorithms, extending this way its applicability to a wider range of scenarios and user needs. Plans include enabling users to define custom weights in the cost function to reflect the varying difficulty of modifying specific features in different contexts. Additionally, we aim to enhance scalability, usability, and introduce additional visualizations for comparative analysis, ensuring the platform's adaptability for large datasets and complex decision-making models.

Figure 3: The *Analyze Counterfactuals* view in GLOVES, showcasing a single configuration's details, including generated counterfactual actions, their impact on the affected population, and visual insights.

ACKNOWLEDGMENTS

This work was partially supported by AutoFair project (EU Horizon program, GA 101070568) and ExtremeXP project (EU Horizon program, GA 101093164).

REFERENCES

[1] Furui Cheng, Yao Ming, and Huamin Qu. 2021. DECE: Decision Explorer with Counterfactual Explanations for Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (Feb. 2021), 1438–1447. https://doi.org/10.1109/TVCG.2020.3030342 Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[2] Germain Garcia-Zanabria, Daniel A Gutierrez-Pachas, Guillermo Camara-Chavez, Jorge Poco, and Erick Gomez-Nieto. [n.d.]. SDA-Vis: A Visualization System for Student Dropout Analysis Based on Counterfactual Exploration. ([n. d.]).

[3] Giorgos Giannopoulos, George Papastefanatos, Dimitris Sacharidis, and Kostas Stefanidis. 2021. Interactivity, Fairness and Explanations in Recommendations. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21)*. 157–161. https://doi.org/10.1145/3450614.3462238

[4] Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. 2020. ViCE: Visual Counterfactual Explanations for Machine Learning Models. https://doi.org/10.48550/arXiv.2003.02428 arXiv:2003.02428.

[5] Victor Guyomard, Françoise Fessant, Thomas Guyet, Tassadit Bouadi, and Alexandre Termier. 2023. Interactive Visualization of Counterfactual Explanations for Tabular Data. In *Machine Learning and Knowledge Discovery in Databases: Applied Data Science and Demo Track*, Gianmarco De Francisci Morales, Claudia Perlich, Natali Ruchansky, Nicolas Kourtellis, Elena Baralis, and Francesco Bonchi (Eds.). Springer Nature Switzerland, Cham, 330–334. https://doi.org/10.1007/978-3-031-43430-3_25

[6] Gyewon Jeon, Sangyeon Kim, and Sangwon Lee. 2024. Interactive Feedback Loop with Counterfactual Data Modification for Serendipity in a Recommendation System. *International Journal of Human–Computer Interaction* 40, 19 (Oct. 2024), 5585–5601. https://doi.org/10.1080/10447318.2023.2238369 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2023.2238369.

[7] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Yuichi Ike. 2022. Counterfactual explanation trees: Transparent and consistent actionable recourse with decision trees. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1846–1870.

[8] Smiti Kaul, David Borland, Nan Cao, and David Gotz. 2022. Improving Visualization Interpretation Using Counterfactuals. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (Jan. 2022), 998–1008. https://doi.org/10.1109/TVCG.2021.3114779 Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[9] Loukas Kavouras, Eleni Psaroudaki, Konstantinos Tsopelas, Dimitrios Rontogiannis, Nikolaos Theologitis, Dimitris Sacharidis, Giorgos Giannopoulos, Dimitrios Tomaras, Kleopatra Markou, Dimitrios Gunopulos, Dimitris Fotakis, and Ioannis Emiris. 2024. GLANCE: Global Actions in a Nutshell for Counterfactual Explainability. arXiv:cs.LG/2405.18921 https://arxiv.org/abs/2405.18921

[10] Loukas Kavouras, Konstantinos Tsopelas, Giorgos Giannopoulos, Dimitris Sacharidis, Eleni Psaroudaki, Nikolaos Theologitis, Dimitrios Rontogiannis, Dimitris Fotakis, and Ioannis Emiris. 2023. Fairness Aware Counterfactuals for Subgroups. *arXiv preprint arXiv:2306.14978* (2023).

[11] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2019. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 131–138.

[12] Dan Ley, Saumitra Mishra, and Daniele Magazzeni. 2022. Global Counterfactual Explanations: Investigations, Implementations and Improvements. In *ICLR Workshop on Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*.

[13] Dan Ley, Saumitra Mishra, and Daniele Magazzeni. 2023. GLOBE-CE: A Translation Based Approach for Global Counterfactual Explanations. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.), Vol. 202. PMLR, 19315–19342. https://proceedings.mlr.press/v202/ley23a.html

[14] Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:stat.ML/1802.03426 https://arxiv.org/abs/1802.03426

[15] Kaivalya Rawal and Himabindu Lakkaraju. 2020. Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses. *Advances in Neural Information Processing Systems* 33 (2020), 12187–12198.

[16] Jan-Tobias Sohns, Christoph Garth, and Heike Leitte. 2023. Decision Boundary Visualization for Counterfactual Reasoning. *Computer Graphics Forum* 42, 1 (2023), 7–20. https://doi.org/10.1111/cgf.14650 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14650.

[17] Arran Zeyu Wang, David Borland, and David Gotz. 2024. A framework to improve causal inferences from visualizations using counterfactual operators. *Information Visualization* (Aug. 2024), 14738716241265120. https://doi.org/10.1177/14738716241265120 Publisher: SAGE Publications.

[18] Greta Warren, Mark T Keane, Christophe Gueret, and Eoin Delaney. 2023. Explaining groups of instances counterfactually for XAI: a use case, algorithm and user study for group-counterfactuals. *arXiv preprint arXiv:2303.09297* (2023).

[19] Lingyun Yu, Nikos Bikakis, Panos K. Chrysanthis, Guoliang Li, and George Papastefanatos. 2025. Data Exploration & Visual Analytics in AI Era. *Big Data Research* (2025).