Calibrated Explanations: Trustworthy uncertainty for every prediction and explanation

Helena Löfström[2,*], Tuwe Löfström[1,*]

[2]Jönköping International Business School, Jönköping University, Sweden
[1]School of Engineering, Jönköping University, Sweden

Abstract

We present a live browser-based demonstrator for Calibrated Explanations, a calibration-first XAI framework. A four-stage workflow guides users through project loading, model calibration, instance-level explanation in factual and alternative rule views, and batch prediction with conformal intervals and threshold-event probabilities. The workflow enforces a calibration gate: the Explain and Predict pages remain locked until calibration completes. The web application features a walkthrough of the California Housing regression scenario, illustrating interval prediction and threshold-event probabilities. The live deployment is available online.[1]

Keywords

Uncertainty-aware XAI, Calibrated Explanations, Conformal Prediction, Venn-Abers Calibration, Demonstrator, Reproducibility

1. Introduction

Raw model scores often look decision-ready long before they are empirically reliable. That calibration gap matters in high-stakes settings because a confident score can still be systematically wrong [1]. If explanations are attached directly to such uncalibrated outputs, the explanation may be faithful to the model's error rather than to the risk faced by the user.

Calibrated Explanations addresses that problem by making calibration the first step in the explanation pipeline. The framework builds on conformal prediction for uncertainty quantification [2] and on Venn-Abers calibration for probabilistic outputs [3]. Prior work establishes the methodological basis of Calibrated Explanations including its extension to regression [4, 5].

This paper contributes a live web demonstrator. Three aspects of the design are specific to this system:

- *Calibration gate.* The Explain and Predict pages are unavailable until calibration completes. The calibration requirement is explicit: users cannot reach Explain or Predict without completing the calibration step.
- *Unified workflow.* Calibration, instance-level explanation, and reject-policy configuration are integrated in one browser session without switching tools.
- *Interactive rule editing.* Conjunctions and custom rules can be inserted live, with immediate updates to the calibrated uncertainty display.

The repository exposes three built-in demo pages: *Diabetes (Binary Classification)*, *Iris (Multiclass Classification)*, and *California Housing (Regression)*. The walkthrough focuses on California Housing because it exercises conformal prediction intervals, threshold-event probabilities, and interactive explanation refinement in a single workflow.

[1]https://calibrated-explanations-demo.ju.se

Late-breaking work, Demos and Doctoral Consortium, colocated with the 4th World Conference on eXplainable Artificial Intelligence: July 01–03, 2026, Fortaleza, Brazil

*Corresponding author, equal contribution.
helena.lofstrom@ju.se (H. Löfström); tuwe.lofstrom@ju.se (T. Löfström)
0000-0001-9633-0423 (H. Löfström); 0000-0003-0274-9026 (T. Löfström)

2. Calibration-First Workflow in the Demonstrator

The system consists of a React (TypeScript) frontend communicating with a Python FastAPI backend over a REST API. All calibration, explanation, and prediction computation runs on the backend. The backend hosts the dataset/model registry. The backend uses the `calibrated-explanations` library to calibrate estimators with Venn-Abers calibration [3, 6, 5] for probabilistic outputs and Conformal Predictive Systems (CPS, inductive variant) [2, 5] for interval prediction. The library also produces factual and alternative rule lists and supports conjunctions for exploring feature interactions. A configurable backend URL means the walkthrough runs against either the public deployment or a local instance.

The workbench is organized as four pages: *Project*, *Prepare & Calibrate*, *Explain*, and *Predict*. In the built-in regression path, *Project* loads the California Housing split and target metadata; *Prepare & Calibrate* wraps the selected estimator and calibrates it on a held-out calibration set; *Explain* operates on *X_test* rows and shares the same regression controls as the prediction page; and *Predict* provides the batch view for conformal intervals or threshold probabilities.

In regression mode, the default state presents a conformal prediction interval for y. When *Use Threshold (probabilistic regression)* is enabled, the same calibrated model is queried for the event probability $P(Y \leq T)$. The demonstrator then asks the explanation engine to explain the currently displayed calibrated output rather than a raw score.

The demonstrator becomes interactive after project loading. The built-in demo page loads a dataset/-model pair from the server-side registry, while calibration, explanation generation, batch prediction, reject-policy handling, and plot rendering are executed on demand through the backend API. In the current implementation, the explanation page also exposes *Add Conjunctions*, custom rule insertion, and subset filtering; the batch-side *Predict* page exposes *Reject Policy* controls [7].

The application offers two entry points. The *Guided Demo* targets new users: after selecting a task (classification or regression), the system automatically loads a built-in dataset, runs calibration, and presents a small set of pre-selected representative instances across factual, alternative, and batch views without requiring manual configuration. The *Expert Hub* provides the full four-page workflow for practitioners who wish to control dataset splits, calibration settings, and instance selection directly, and allows users to upload *csv* files for datasets of moderate size.

2.1. Calibration Gate

The *Explain* and *Predict* pages are disabled until the backend returns a completed calibration status. No quantitative pass/fail threshold is enforced. After fitting completes, the page shows a progress log and a completion status; no per-bin coverage diagnostics are currently displayed for the regression path. The gate therefore functions as a workflow checkpoint rather than a substantive quality check.

For split-conformal CPS at significance level ϵ, a non-trivial interval requires at least $\lceil 1/\epsilon \rceil$ calibration instances; below this bound the interval collapses to the full target range. The current UI does not warn when the calibration set approaches this limit; uniformly wide intervals should be read as a calibration-data signal rather than a model signal.

2.2. Methods and Defaults

The California Housing built-in demo uses a fixed random seed (42) for all splits. The dataset is partitioned into approximately 54 % training, 26 % calibration, and 20 % test instances (80/20 train+cal/test split, then 67/33 proper/calibration within the training portion). The default significance level is $\epsilon = 0.10$, giving nominal 90 % coverage; this can be adjusted via the UI before calibration.

The interval prediction method is CPS (inductive variant) [2, 5]. The system also supports Mondrian CPS, which partitions the calibration set by a user-specified categorical variable and provides per-category rather than marginal-only coverage guarantees. Threshold-event probability calibration in regression uses Venn-Abers [3, 5]. For a fixed threshold T, the threshold probability $P(Y \leq T \mid x)$ is computed by the Venn-Abers-calibrated model; the associated *Probability Interval* is the Venn-Abers

output interval for that probability, i.e. the range of calibrated probability estimates returned by the Venn-Abers procedure for the threshold event rather than a new CPS interval for y. This validity statement is therefore per fixed T; if a user inspects several thresholds and then selects one based on the displayed outputs, that multiplicity is not accounted for by the Venn-Abers certificate; see Section 6.

The system accepts any scikit-learn-compatible black-box regressor and does not require native probabilistic outputs; calibration wraps the point-prediction model. Numerical and binary-encoded categorical features are supported, consistent with the `calibrated-explanations` library defaults [5].

Throughout this paper, *probabilistic regression mode* refers to enabling *Use Threshold* in the UI; the CE library documentation uses *probabilistic regression* as the canonical user-facing term for this capability [5].

3. Demo Claims, Scope, and Guarantees

The demonstrator supports the following observations and claims.

Implemented and observable in the demo:

- The calibration-first gate is enforced: explanation and prediction are unavailable until calibration completes.
- Inductive CPS (conformal regression) provides approximately exact marginal coverage under exchangeability [2, 5].
- Venn-Abers produces a multiprobability interval $[p_0, p_1]$ for the threshold event that is calibrated in the Venn-Abers sense under exchangeability [3].
- In threshold mode, rule weights are computed on the midpoint of the Venn-Abers interval, while the displayed weight interval is induced by the lower and upper calibrated estimates for the perturbed instances rather than by raw model confidence [6, 5].

What the system does not guarantee:

- Explanation quality under small or unrepresentative calibration sets.
- Feasibility or actionability of alternative rules.
- Causal relationships.

Conditional (per-subgroup) coverage is supported via Mondrian CPS, which partitions the calibration set by a categorical mondrian variable and provides separate coverage guarantees per category [2], distinct from marginal CPS.

The exchangeability assumption underlying CPS holds when calibration and test data are drawn from the same distribution. Under covariate shift, empirical coverage can fall below $1 - \epsilon$; the current UI performs no automated shift test. Users must verify distributional similarity through external means before applying intervals to deployment data.

A terminological note: *prediction interval* denotes the CPS interval for y (a set with a frequentist coverage guarantee); *probability interval* denotes the Venn-Abers output interval for $P(Y \leq T)$ (an interval over a calibrated probability, not over y). The UI labels these objects distinctly; conflating them leads to incorrect decision thresholds.

4. Live Demonstrator

Figures 1 and 2 illustrate the two most important interface states: explicit calibration before explanation, and the switch from interval prediction to probabilistic regression explanation. The following steps walk through the Expert Hub path using California Housing as the anchor dataset.

1. On *Project*, select *California Housing (Regression)* from Quick Start; the dataset preview and target metadata load and *Explain*/*Predict* remain gated. The default model is *Random Forest* but *TabPFN* can be chosen.

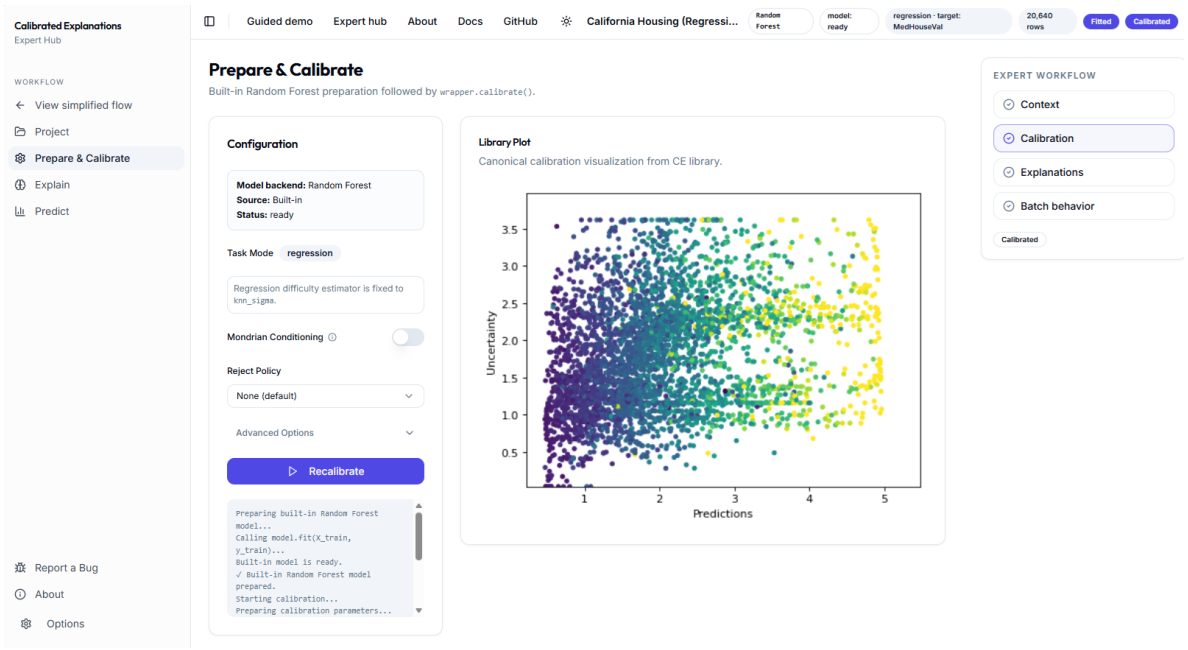

Figure 1: Prepare and calibrate options for regression: After a one-click California Housing selection in *Project*, the calibration gate must complete before *Explain*/*Predict* unlock.

2. On *Prepare & Calibrate*, run *Prepare [model] & Calibrate*; a progress log and completion status appear and the gated pages unlock. Models are fitted or initialized automatically before calibration.
3. On *Explain*, select any row from *Test Set Instances*; the instance header shows $\hat{y}$ with its CPS interval and the *Factual* tab populates with signed rules.
4. Switch between *Factual* and *Alternatives* on the same instance; the factual side shows the current calibrated output and the alternatives side lists candidate condition changes with updated outputs and uncertainty intervals.
5. In *Regression Options*, enable *Use Threshold (probabilistic regression)* and enter a threshold T; the header changes from $\hat{y}$ to $P(Y \leq T)$ and the bracket becomes the Venn-Abers probability interval (see Fig. 2).
6. Click *Add Conjunctions*; conjunctive rules append to the rule list and the plot refreshes, surfacing pairwise feature interactions.
7. On *Predict*, toggle threshold mode off and on; the page alternates between per-row *Conformal Interval* values and per-row probabilities with a *Probability Interval* column.

5. Reading the Regression Explanations

The UI explicitly distinguishes the conformal prediction interval for y (when threshold mode is off) from the Venn-Abers probability interval for $P(Y \leq T)$ (when threshold mode is on). For conformal prediction, CPS provides approximately exact marginal coverage at significance level ϵ.

Each factual rule is displayed with a signed *weight* and a bracketed uncertainty interval. The weight represents the rule's signed effect on the currently displayed calibrated output, not a raw model coefficient. In threshold mode, this weight is computed on the midpoint of the Venn-Abers interval, while the bracketed uncertainty interval is induced by the lower and upper calibrated estimates obtained for the perturbed instances [6, 5]. Table 1 shows an illustrative factual explanation for a California Housing test instance: the first two rules have intervals that do not cross zero and therefore have a clear direction, while the third has uncertain direction and is flagged accordingly in the UI.

Alternative rules are presented as candidate condition changes with an updated predicted output and its uncertainty interval. Among the alternatives listed, those flagged as *ensured* [8] satisfy a stricter

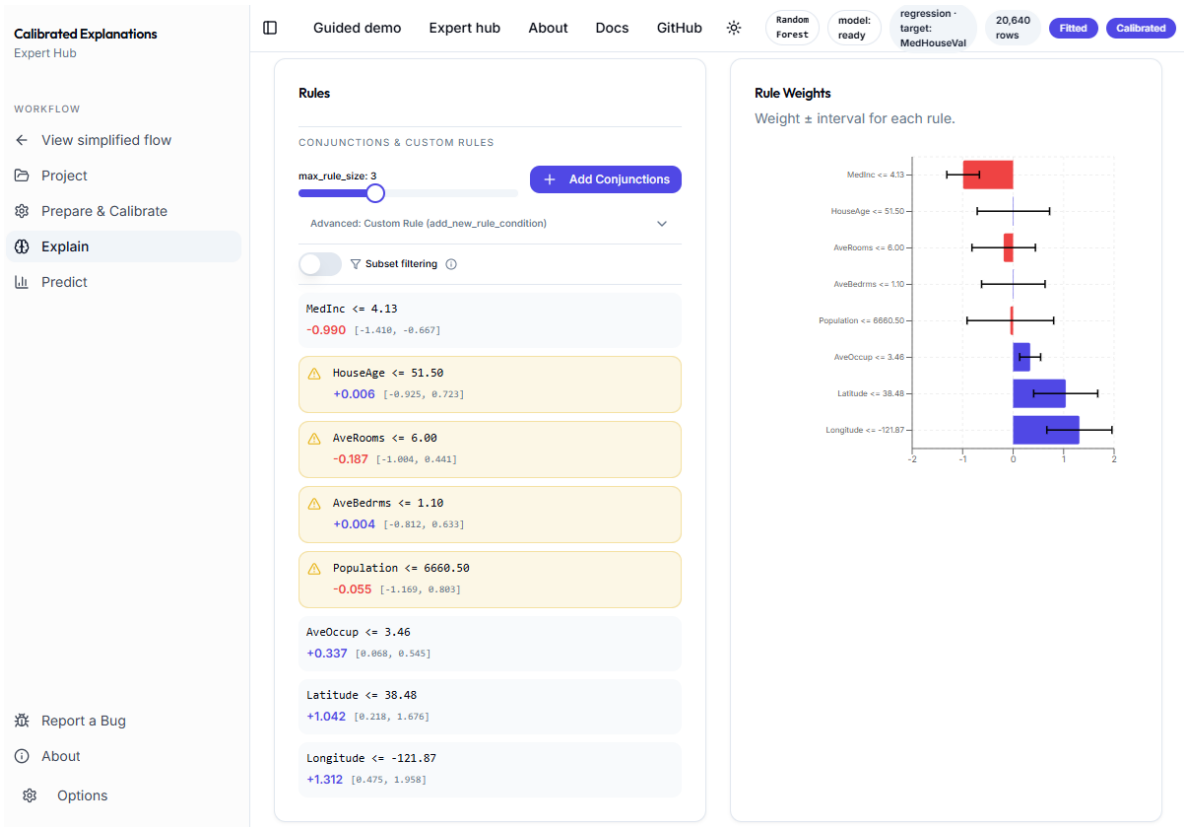

Figure 2: Probabilistic regression mode (*Use Threshold* enabled): the *Explain* page shows a factual explanation for a test instance, with the output header displaying $P(Y \leq T)$ and its Venn-Abers probability interval in place of the conformal prediction interval for y. Rule weights reflect the signed effect on the midpoint of that calibrated probability interval, while the displayed weight interval is induced by the lower and upper perturbed estimates.

Table 1
Illustrative factual rules for a California Housing test instance ($\hat{y} = 2.83$, CPS interval $[2.41,\ 3.25]$ at $\epsilon = 0.10$). A weight interval crossing zero indicates direction uncertainty.

Condition	Weight	Interval
$MedInc > 5.00$	$+0.38$	$[+0.25,\ +0.51]$
$Latitude > 37.80$	-0.22	$[-0.34,\ -0.11]$
$AveRooms \leq 6.40$	$+0.07$	$[-0.03,\ +0.17]$

criterion: the uncertainty interval of the predicted output is no wider under the alternative condition than under the factual condition.

An alternative rule is a modified version of a factual rule in which one or more feature conditions are changed. The alternative reports the calibrated output and its uncertainty interval that would result if the instance satisfied the modified conditions. Candidates are generated by perturbing the factual conditions within the observed feature range [6, 5].

Feasibility and actionability constraints are not enforced. An alternative may describe a condition that is physically impossible, causally inconsistent, or practically unachievable. Users should treat alternative rules as conditional sensitivity queries rather than actionable recommendations; domain knowledge is required to filter them.

Conjunctions extend a rule by adding a second feature condition, surfacing pairwise interaction effects. Clicking *Add Conjunctions* appends conjunction-extended rules to the current list; the number of candidates is bounded by a configurable cap to avoid display overload.

6. Failure Modes, Limitations, and Misuse Risks

CPS intervals become uninformatively wide when the calibration set contains fewer than $\lceil 1/\epsilon \rceil$ instances; the UI does not prevent calibration in this case, meaning uniformly wide intervals signal a data-size problem. The marginal coverage guarantee also requires exchangeability: if the test distribution differs from the calibration distribution, empirical miscoverage may exceed ϵ, and the current UI performs no automated shift test.

Two reading pitfalls arise with the regression output. A weight interval crossing zero does not indicate that the feature has no effect; it indicates that the calibrated evidence is insufficient to determine direction. Additionally, confusing the *Probability Interval* with a prediction interval for y leads to incorrect interval widths in threshold-based decisions.

Causal effect estimation and actionability filtering for alternatives are not implemented. A well-calibrated model with poor predictive accuracy will produce valid but wide intervals; the demo does not validate predictive quality beyond calibration.

7. Related Work

Explanation methods. Post-hoc explanation methods for black-box models can generally be divided into feature-attribution, rule-based, and counterfactual approaches. In several surveys a broad overview of the area is provided, such as those by Guidotti et al. [9] and Molnar [10]. Among local attribution methods, LIME [11], SHAP [12], and Integrated Gradients [13] explain individual predictions through feature relevance scores. Rule-based approaches such as Anchors [14] instead aim to produce human-readable local decision conditions. A complementary line of work studies counterfactual explanations, where explanations are given as minimal changes to the input that would alter the prediction [15, 16]. These categories of explanations differ in explanatory form and intended use, but most primarily target interpretability, actionability, or local fidelity rather than formal calibration of predictive confidence.

Uncertainty quantification and calibration. A separate literature focuses on uncertainty quantification for predictions. Conformal prediction provides distribution-free prediction sets or intervals under exchangeability [2], with later extensions including Mondrian conformal prediction and conformalized quantile regression [17]. More recent overviews have clarified the role of conformal methods as practical uncertainty wrappers for black-box models [18]. For probabilistic calibration, Venn-Abers predictors produce calibrated probability intervals for classification and related threshold-event queries [3]. While these methods are not explanation techniques in themselves, they are highly relevant for explanation interfaces because they can communicate when a prediction is well supported and when explanatory claims should be presented more cautiously.

Systems and tool support. Several software systems expose explanation methods in practical workflows. Alibi [19], Captum [20], and the What-If Tool [21] support multiple explanation modalities and interactive inspection of model behaviour. Some of these tools expose method-specific confidence-like quantities or prediction scores, but they do not attach calibrated uncertainty estimates to the explanations themselves, nor use such uncertainty to govern when explanations are shown and how alternatives are prioritized.

Positioning of our demo. Our demo builds on prior work on calibrated explanations [5, 6], which connects local explanatory outputs with calibrated predictive uncertainty. In contrast to systems that present explanations without an explicit reliability layer, the demo emphasizes how calibrated intervals or probabilities can support explanation presentation, abstention, and the ranking of alternative conditions. This positions the contribution at the intersection of explainability, uncertainty quantification, and interactive decision support.

8. Conclusion

The demonstrator makes calibration-first explanations directly inspectable in an interactive workflow. The three contributions are: a calibration gate that requires calibration to complete before explanations are available; a unified session spanning calibration, explanation inspection, and reject-policy configuration; and interactive rule editing with live uncertainty updates. The browser-based walkthrough makes the calibration-to-decision workflow observable rather than merely described: users move from conformal prediction intervals to threshold-event queries and compare factual, alternative, and conjunctive rules in one coherent session with the California Housing dataset. The live deployment and reference implementation are publicly available.

Acknowledgements

Tuwe Löfström acknowledges the Swedish Knowledge Foundation for financially supporting the research and education environment on Knowledge Intensive Product Realization SPARK at Jönköping University, Sweden, project PREMACOP, grant no. 20220187. Helena Löfström acknowledges support from the Kamprad Family Foundation for the project *Förbättrad livskvalitet för äldre genom testning för allvarliga sjukdomar*, grant reference no. 20250309.

Declaration on Generative AI

During the preparation of this work, the authors used LLMs for grammar and spelling check and to improve writing style. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: Proceedings of the 34th International Conference on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 2130–2143.

[2] V. Vovk, A. Gammerman, G. Shafer, Algorithmic Learning in a Random World, Springer, 2005.

[3] V. Vovk, I. Petej, Venn–abers predictors, in: Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence, 2014, pp. 731–755.

[4] H. Löfström, Trustworthy explanations: Improved decision support through well-calibrated uncertainty quantification, Ph.D. thesis, Jönköping International Business School, Jönköping University, 2023.

[5] T. Löfström, H. Löfström, U. Johansson, C. Sönströd, R. Matela, Calibrated explanations for regression, Machine Learning 114 (2025). doi:10.1007/s10994-024-06642-8.

[6] H. Löfström, T. Löfström, U. Johansson, C. Sönströd, Calibrated explanations: with uncertainty information and counterfactuals, Expert Systems with Applications (2024) 123154. doi:10.1016/j.eswa.2024.123154.

[7] J. Hallberg Szabadváry, T. Löfström, U. Johansson, C. Sönströd, E. Ahlberg, L. Carlsson, Classification with reject option: Distribution-free error guarantees via conformal prediction, Machine Learning with Applications 20 (2025) 100664. doi:10.1016/j.mlwa.2025.100664.

[8] H. Löfström, T. Löfström, J. Hallberg Szabadvary, Ensured: Explanations for decreasing the epistemic uncertainty in predictions, 2024. arXiv:2410.05479.

[9] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Computing Surveys 51 (2018) 1–42. doi:10.1145/3236009.

[10] C. Molnar, Interpretable Machine Learning, 2 ed., 2022. URL: https://christophm.github.io/interpretable-ml-book.

[11] M. T. Ribeiro, S. Singh, C. Guestrin, "why should I trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 1135–1144. doi:10.1145/2939672.2939778.

[12] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017, pp. 4765–4774.

[13] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: Proceedings of the 34th International Conference on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 3319–3328.

[14] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, volume 32, AAAI Press, 2018, pp. 1527–1535.

[15] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, Harvard Journal of Law & Technology 31 (2017) 841–887.

[16] R. K. Mothilal, A. Sharma, C. Tan, Explaining machine learning classifiers through diverse counterfactual explanations, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, ACM, 2020, pp. 607–617. doi:10.1145/3351095.3372850.

[17] Y. Romano, E. Patterson, E. J. Candes, Conformalized quantile regression, in: Advances in Neural Information Processing Systems, volume 32, Curran Associates, Inc., 2019, pp. 3538–3548.

[18] A. N. Angelopoulos, S. Bates, Conformal prediction: A user-friendly introduction, Foundations and Trends in Machine Learning 16 (2023) 494–591. doi:10.1561/2200000106.

[19] J. Klaise, A. Van Looveren, G. Vacanti, A. Coca, Alibi explain: Algorithms for explaining machine learning models, Journal of Machine Learning Research 22 (2021) 1–7. doi:10.5555/3546258.3546315.

[20] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, et al., Captum: A unified and generic model interpretability library for PyTorch, 2020. arXiv:2009.07896.

[21] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, J. Wilson, The what-if tool: Interactive probing of machine learning models, IEEE Transactions on Visualization and Computer Graphics 26 (2019) 56–65. doi:10.1109/TVCG.2019.2934619.