

EVALUATING EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR  
HIGH-STAKES DECISION MAKING:  
AN EMPIRICAL ASSESSMENT OF PREDICTIVE ACCURACY,  
ALGORITHMIC FAIRNESS, AND MODEL INTERPRETABILITY

Saba Jabbar<sup>\*1</sup>, Gulzar Ali Brohi<sup>2</sup>, Muhammad Mudaser<sup>3</sup>, Atif Saeed<sup>4</sup>, Nazia Azim<sup>5</sup>

<sup>\*1</sup>PhD Artificial Intelligence,

<sup>2</sup>M.Phil Computer Science, University of Sindh Jamshoro

<sup>3</sup>National College of Business Administration and Economics (NCBA&E), Multan Campus

<sup>4</sup>Department of CS, COMSATS University Islamabad, Lahore campus.

<sup>5</sup>Lecturer department of computer science Abdul Wali Khan University Mardan

<sup>\*1</sup>sabajabbarmalik@gmail.com, <sup>2</sup>gulzar.brohi@usindh.edu.pk, <sup>3</sup>mudasirmalik6763@gmail.com

<sup>4</sup>asaheed@cuilahore.edu.pk, <sup>5</sup>n.azim@awkum.edu.pk

DOI: <https://doi.org/10.5281/zenodo.22842942>

**Keywords**

explainable artificial intelligence;  
XAI; algorithmic fairness;  
interpretability; COMPAS;  
SHAP; calibration; high-stakes  
decision making; responsible AI

**Article History**

Received: 12 June 2026

Accepted: 03 September 2026

Published: 19 September 2026

Copyright @Author

Corresponding Author: \*

Saba Jabbar

**Abstract**

High-stakes artificial intelligence requires evaluation beyond predictive accuracy because errors, calibration, demographic disparities, and explanation quality can have consequential effects. This study evaluated logistic regression, a shallow decision tree, random forest, XGBoost, and an interpretable spline generalized additive model (GAM) using the ProPublica COMPAS two-year recidivism dataset (N=7,214). A leakage-conscious feature set included age, sex, prior-offense count, juvenile-offense counts, and charge degree, while race was reserved for fairness auditing. The primary 75:25 holdout was supplemented by repeated stratified five-fold cross-validation with three repeats. Mean cross-validated ROC-AUC was 0.724 for logistic regression, 0.718 for the decision tree, 0.729 for random forest, and approximately 0.731 for both XGBoost and the spline GAM. Calibration was broadly comparable (held-out ECE 0.020-0.037), and threshold analysis from 0.20 to 0.80 showed that measured racial disparities changed materially with the operating point. At threshold 0.50, all models produced higher positive-prediction and false-positive rates for African-American than Caucasian observations. A secondary sex-based analysis also identified non-trivial disparities. XGBoost explanations were dominated by prior-offense count and age; global SHAP rankings were highly stable across bootstrap refits (mean pairwise Spearman rho=0.973; mean top-five overlap=1.00). These findings show that model complexity produced limited discrimination gains and did not eliminate demographic disparities. High-stakes model evaluation should therefore jointly examine discrimination, calibration, subgroup error burdens, threshold sensitivity, and explanation stability rather than relying on accuracy or explanation display alone.

**1. Introduction**

Artificial intelligence increasingly supports consequential decisions in criminal justice,

health care, credit, insurance, employment, education, and public administration. In these settings, predictive validity is necessary but

insufficient because errors can affect liberty, access to services, financial opportunity, or institutional scrutiny. Trustworthy-AI frameworks therefore treat validity, reliability, fairness, transparency, explainability, and accountability as related but distinct properties. [1,21,22,36]

Explainable artificial intelligence (XAI) makes aspects of model behavior more inspectable. Intrinsically interpretable models expose their structure directly, whereas post-hoc methods such as LIME and SHAP approximate or attribute the behavior of more complex predictors. [4,5,6] Explanation quality is itself multidimensional: fidelity, stability, usability, stakeholder relevance, and task appropriateness may diverge, and recent evaluation frameworks argue that explanations should be assessed rather than merely displayed. [13,14,15,16,17,18,26,27,28,29,30,32,33,34,35]

Algorithmic fairness is similarly multidimensional. Demographic parity, equal opportunity, predictive-value criteria, and equalized odds encode different distributional goals and can conflict when outcome prevalence differs across groups. [7,8,9,10,19,20,25,31] This study therefore evaluates predictive performance, calibration, race- and sex-based fairness, threshold sensitivity, model-agnostic importance, SHAP behavior, and explanation stability in a common empirical pipeline. The objective is not to identify a universally preferable algorithm, but to make the relevant trade-offs visible and reproducible.

## 2. Literature Review and Conceptual Framework

### 2.1 Explainability in high-stakes AI

Foundational work on LIME and SHAP established widely used local and additive attribution approaches. [4,5] Subsequent scholarship has emphasized that post-hoc explanations can be unstable, can be sensitive to modeling choices, and should not be interpreted as causal effects. [13,18,23,24,28,32]

Functionally grounded, human-grounded, and application-grounded evaluations address different questions, while usability-oriented frameworks emphasize whether explanations are understandable and actionable for intended users. [18,26,27,29,30,33,34]

### 2.2 Algorithmic fairness

Fairness metrics quantify different properties of predictions and errors. Demographic parity concerns positive prediction rates; equal opportunity focuses on true-positive rates; and equalized odds considers both true-positive and false-positive behavior. [7,8,9,10] COMPAS has become a prominent benchmark because prior analyses demonstrated that apparently reasonable performance summaries can coexist with materially different group error profiles. [2,3,8,25] XAI can support auditing, but explanation does not by itself establish fairness. [19,20]

### 2.3 Accuracy-fairness-interpretability trade-offs

In high-stakes applications, a small gain in discrimination may not justify substantially greater opacity or larger error disparities, while transparent models should not be preferred automatically if they are materially less reliable. Contemporary work increasingly treats XAI evaluation as a combination of technical, human-centered, and governance criteria. [13,17,18,21,22,26,30,36]

## 3. Data and Methods

### 3.1 Dataset and analytic sample

The analysis used the publicly released ProPublica COMPAS two-year dataset. [2,3] The analytical dataset comprised 7,214 observations and 53 variables. The binary outcome was two\_year\_recid; 3,251 observations (45.1%) had a positive outcome. Race was excluded from the primary predictive feature set and retained for fairness auditing. The original held-out test sample contained 1,804 observations, including 933 African-American and 588 Caucasian observations.

*Table 1. Descriptive characteristics of the analytical sample (N=7,214).*

| Characteristic | Value        |
|----------------|--------------|
| Age, years     | 34.8 ± 11.9  |
| Male sex       | 5819 (80.7%) |

| Characteristic                           | Value        |
|------------------------------------------|--------------|
| Female sex                               | 1395 (19.3%) |
| African-American race                    | 3696 (51.2%) |
| Caucasian race                           | 2454 (34.0%) |
| Prior offenses, median (IQR)             | 2 (0-5)      |
| Juvenile felony count, median (IQR)      | 0 (0-0)      |
| Juvenile misdemeanor count, median (IQR) | 0 (0-0)      |
| Felony charge degree                     | 4666 (64.7%) |
| Two-year recidivism                      | 3251 (45.1%) |

### 3.2 Predictors, preprocessing, and leakage control

Primary predictors were age, sex, priors\_count, juvenile felony count, juvenile misdemeanor count, juvenile other-offense count, and charge degree. COMPAS score fields, post-index recidivism information, later case outcomes, and custody variables that could encode future information were excluded. Numeric variables were median-imputed. Logistic regression standardized numeric predictors and used one-hot encoding for categorical predictors; tree-based models used median imputation and one-hot encoding without scaling. Logistic regression used L2 regularization ( $C=1.0$ ), the default lbfgs solver, and max\_iter=3000.

### 3.3 Models and validation

Five models were fitted: logistic regression; a decision tree (max\_depth=4, min\_samples\_leaf=40); random forest (200 trees, max\_depth=7, min\_samples\_leaf=8); XGBoost (160 estimators, max\_depth=3, learning\_rate=0.04, subsample=0.85, colsample\_bytree=0.90); and a spline GAM. The GAM used cubic spline bases (degree=3, five knots, no bias term) for age and priors\_count, standardized linear terms for the three juvenile-count predictors, one-hot encoded sex and charge degree, and L2-regularized logistic regression ( $C=1.0$ ).

The primary held-out analysis used a 75:25 stratified split (random\_state=42). Robustness was assessed with repeated stratified five-fold cross-validation with three repeats (15 held-out folds; random\_state=2026). Fold-level paired t-tests and Wilcoxon tests are reported only as exploratory dependence-sensitive comparisons because repeated cross-validation folds are not statistically independent. To reduce multiplicity

in the t-test family, Holm-adjusted p-values are also reported; practical magnitude and fold-wise differences are emphasized over binary significance declarations.

### 3.4 Performance, calibration, and fairness metrics

Predictive performance was evaluated with ROC-AUC, accuracy, precision, recall, F1 score, and Brier score. Calibration intercept and slope were estimated by regressing the binary outcome on the logit of predicted probability. Expected calibration error (ECE) used 10 equal-frequency (quantile) bins and was calculated as the sample-size-weighted absolute difference between mean predicted probability and observed event rate within bins. Calibration curves used the same 10-bin quantile strategy.

Fairness was evaluated primarily for African-American versus Caucasian observations using selection rate, demographic-parity difference, selection-rate ratio, true-positive rate (TPR), false-positive rate (FPR), false-negative rate (FNR), positive predictive value (PPV), equal-opportunity difference, and equalized-odds difference. A secondary sensitivity analysis evaluated Male minus Female differences. The default threshold was 0.50; threshold sensitivity was evaluated from 0.20 to 0.80 in increments of 0.05. Three hundred bootstrap resamples of the held-out test set (seed=123) quantified uncertainty for key metrics.

### 3.5 Interpretability analysis

Permutation importance was calculated for every model on the held-out test set using eight repeats (random\_state=42) and decrease in ROC-AUC as the scoring criterion. XGBoost was additionally evaluated with TreeSHAP using shap.TreeExplainer on the fitted tree ensemble.

SHAP values were computed for the first 1,000 held-out observations. Dependence plots used original age in years and original `priors_count` on the horizontal axes while plotting the corresponding SHAP attribution. Explanation stability was assessed by refitting XGBoost on five bootstrap resamples (`seed=77`), computing

global mean absolute SHAP values on a fixed 500-observation test subset, and comparing rankings using pairwise Spearman correlation and top-five overlap. SHAP values describe fitted-model behavior and are not causal effects. [4,23,24]

#### 4. Results

##### 4.1 Predictive performance and repeated validation

*Table 2. Predictive performance on the original held-out test sample (n=1,804).*

| Model               | ROC-AUC | Accuracy | Precision | Recall | F1    | Brier |
|---------------------|---------|----------|-----------|--------|-------|-------|
| Logistic regression | 0.724   | 0.676    | 0.679     | 0.533  | 0.597 | 0.212 |
| Decision tree       | 0.717   | 0.667    | 0.670     | 0.513  | 0.581 | 0.213 |
| Random forest       | 0.704   | 0.667    | 0.655     | 0.552  | 0.599 | 0.218 |
| XGBoost             | 0.725   | 0.682    | 0.667     | 0.588  | 0.625 | 0.210 |
| Spline GAM          | 0.728   | 0.683    | 0.660     | 0.613  | 0.635 | 0.209 |

*Table 3. Repeated stratified five-fold cross-validation results (three repeats; 15 held-out folds).*

| Model               | AUC mean | AUC SD | Accuracy mean | F1 mean | Brier mean |
|---------------------|----------|--------|---------------|---------|------------|
| Logistic regression | 0.724    | 0.012  | 0.677         | 0.597   | 0.212      |
| Decision tree       | 0.718    | 0.011  | 0.675         | 0.607   | 0.211      |
| Random forest       | 0.729    | 0.010  | 0.679         | 0.619   | 0.209      |
| XGBoost             | 0.731    | 0.010  | 0.680         | 0.622   | 0.207      |
| Spline GAM          | 0.731    | 0.011  | 0.682         | 0.634   | 0.208      |

Repeated validation produced mean AUCs of 0.724 (SD 0.012) for logistic regression, 0.718 (0.011) for the decision tree, 0.729 (0.010) for random forest, 0.731 (0.010) for XGBoost, and 0.731 (0.011) for the spline GAM. Thus, the single-split ordering was not fully preserved, but the principal conclusion was stable: discrimination differences among the strongest models were small.

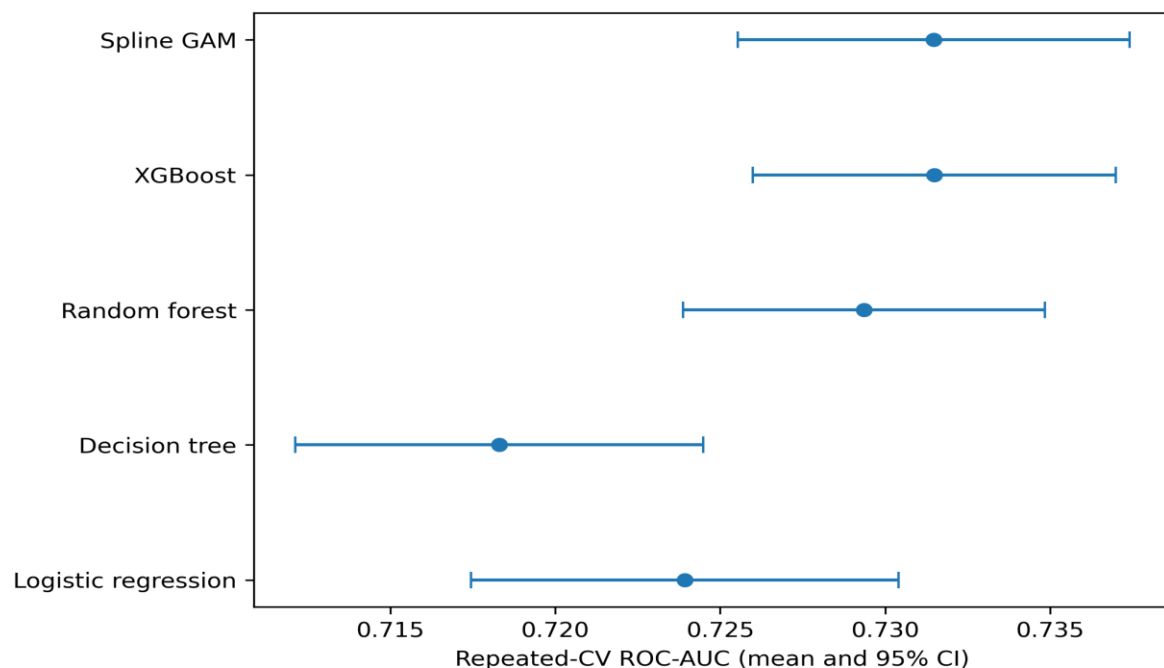


Figure 1. Mean repeated-cross-validation ROC-AUC with 95% *t*-based confidence intervals across the 15 validation folds.

Table 4. Exploratory fold-level paired comparisons against the spline GAM. Repeated-CV folds are not independent; *p*-values are descriptive and Holm adjustment is applied to the paired *t*-test family.

| Comparator          | Reference  | Mean AUC difference | 2.5pct fold diff | 97.5pct fold diff | Paired <i>t</i> <i>p</i> | Holm-adjusted <i>p</i> | Wilcoxon <i>p</i> |
|---------------------|------------|---------------------|------------------|-------------------|--------------------------|------------------------|-------------------|
| Logistic regression | Spline GAM | 0.008               | 0.002            | 0.014             | 0.000                    | 0.000                  | 0.000             |
| Decision tree       | Spline GAM | 0.013               | 0.007            | 0.021             | 0.000                    | 0.000                  | 0.000             |
| Random forest       | Spline GAM | 0.002               | -0.002           | 0.008             | 0.018                    | 0.036                  | 0.030             |
| XGBoost             | Spline GAM | -0.000              | -0.004           | 0.003             | 0.976                    | 0.976                  | 1.000             |

The mean GAM-XGBoost AUC difference was effectively zero (-0.00002), whereas the GAM-logistic difference was 0.0075. Because repeated-CV folds overlap in training data, these tests are

not treated as independent-sample confirmatory inference; effect sizes and consistency across folds are the primary evidence.

#### 4.2 Calibration

Table 5. Held-out calibration metrics. ECE uses 10 equal-frequency bins.

| Model               | Brier | Calibration intercept | Calibration slope | ECE   |
|---------------------|-------|-----------------------|-------------------|-------|
| Logistic regression | 0.212 | 0.003                 | 0.979             | 0.037 |
| Decision tree       | 0.214 | -0.011                | 0.873             | 0.020 |
| Random forest       | 0.210 | 0.020                 | 1.027             | 0.020 |

| Model      | Brier | Calibration intercept | Calibration slope | ECE   |
|------------|-------|-----------------------|-------------------|-------|
| XGBoost    | 0.210 | -0.006                | 0.911             | 0.029 |
| Spline GAM | 0.209 | 0.002                 | 0.937             | 0.028 |

Held-out Brier scores ranged from 0.209 for the spline GAM to 0.214 for the decision tree. Calibration slopes ranged from 0.873 to 1.027

and intercepts remained close to zero. ECE ranged from 0.020 to 0.037, indicating broadly similar calibration performance across models.

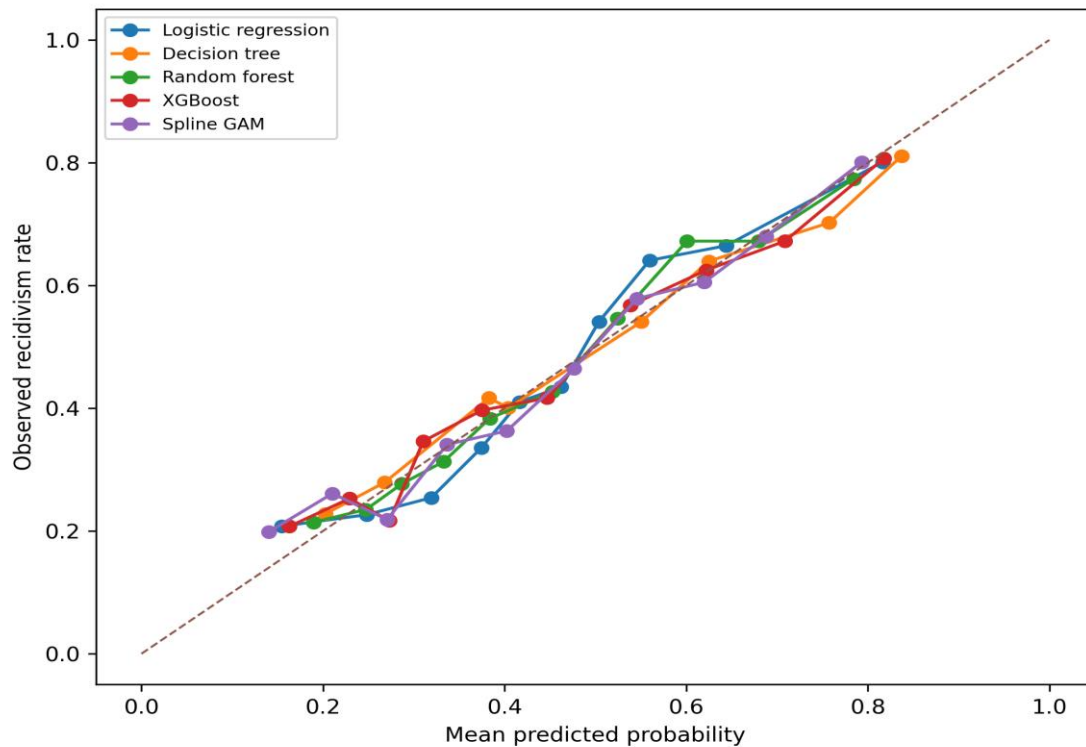


Figure 2. Calibration curves on the held-out test sample using 10 equal-frequency bins

#### 4.3 Race-based fairness and threshold sensitivity

Table 6. Group-specific race fairness metrics at threshold 0.50. Rates are reported separately to make the direction and magnitude of disparities transparent.

| Model               | Race group       | n   | Selection rate | TPR   | FPR   | FNR   | PPV   |
|---------------------|------------------|-----|----------------|-------|-------|-------|-------|
| Logistic regression | African-American | 933 | 0.472          | 0.648 | 0.292 | 0.352 | 0.693 |
| Logistic regression | Caucasian        | 588 | 0.226          | 0.386 | 0.130 | 0.614 | 0.639 |
| Decision tree       | African-American | 933 | 0.494          | 0.660 | 0.325 | 0.340 | 0.675 |
| Decision tree       | Caucasian        | 588 | 0.284          | 0.450 | 0.185 | 0.550 | 0.593 |
| Random forest       | African-American | 933 | 0.498          | 0.679 | 0.314 | 0.321 | 0.688 |
| Random forest       | Caucasian        | 588 | 0.257          | 0.418 | 0.160 | 0.582 | 0.609 |

| Model      | Race group       | n   | Selection rate | TPR   | FPR   | FNR   | PPV   |
|------------|------------------|-----|----------------|-------|-------|-------|-------|
| XGBoost    | African-American | 933 | 0.512          | 0.701 | 0.320 | 0.299 | 0.690 |
| XGBoost    | Caucasian        | 588 | 0.270          | 0.427 | 0.177 | 0.573 | 0.591 |
| Spline GAM | African-American | 933 | 0.541          | 0.730 | 0.348 | 0.270 | 0.681 |
| Spline GAM | Caucasian        | 588 | 0.281          | 0.445 | 0.182 | 0.555 | 0.594 |

At threshold 0.50, all five models produced higher selection rates and FPRs for African-American than Caucasian observations. For example, XGBoost selection rates were 0.512 versus 0.270 and FPRs were 0.320 versus 0.177,

while TPRs were 0.701 versus 0.427. These group-specific values clarify that a single disparity difference can reflect simultaneous changes in both error and selection behavior.

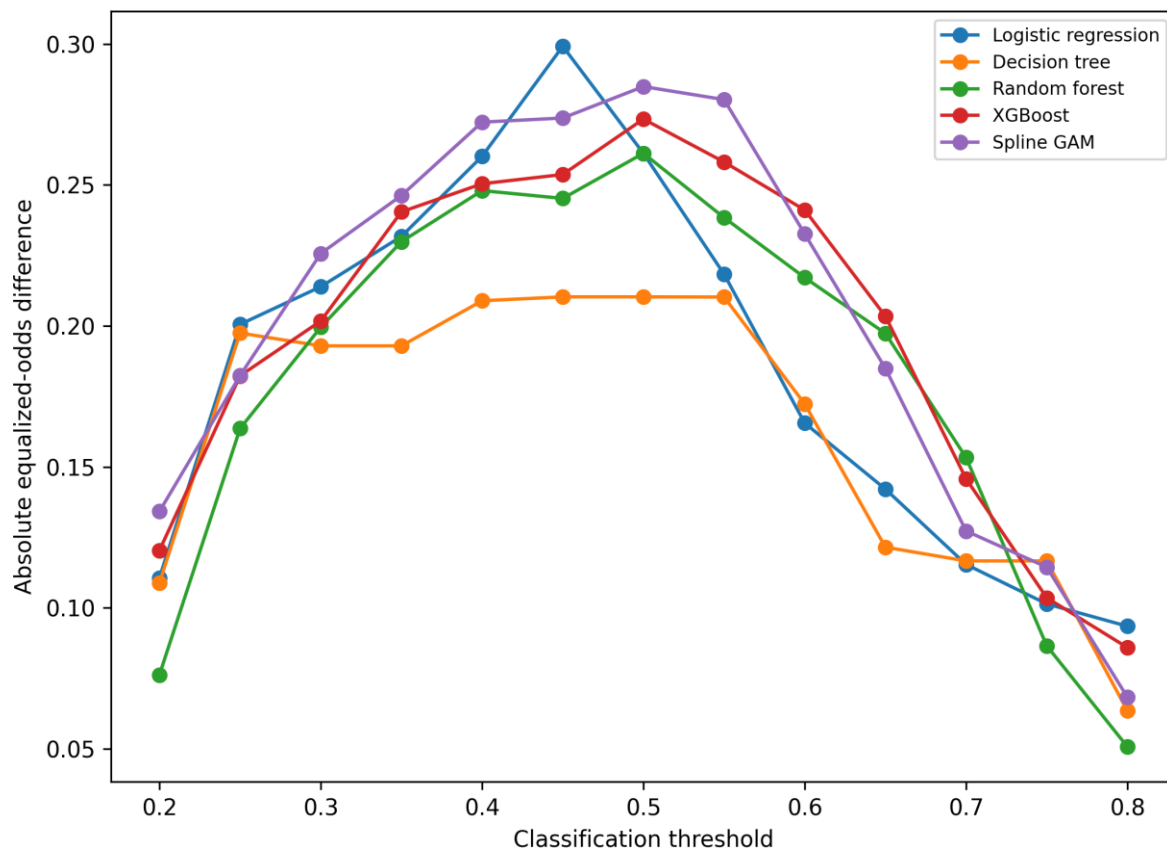


Figure 3. Threshold sensitivity of the absolute African-American-Caucasian equalized-odds difference from thresholds 0.20 to 0.80.

Table 7. Numerical summary of threshold sensitivity for race-based equalized-odds disparity.

| Model               | Minimum  EqOdds | Threshold minimum | at | Maximum  EqOdds | Threshold maximum | at |
|---------------------|-----------------|-------------------|----|-----------------|-------------------|----|
| Logistic regression | 0.093           | 0.800             |    | 0.299           | 0.450             |    |
| Decision tree       | 0.063           | 0.800             |    | 0.210           | 0.450             |    |

| Model         | Minimum<br> EqOdds | Threshold<br>minimum | at | Maximum<br> EqOdds | Threshold<br>maximum | at |
|---------------|--------------------|----------------------|----|--------------------|----------------------|----|
| Random forest | 0.051              | 0.800                |    | 0.261              | 0.500                |    |
| XGBoost       | 0.086              | 0.800                |    | 0.273              | 0.500                |    |
| Spline GAM    | 0.068              | 0.800                |    | 0.285              | 0.500                |    |

Across the 0.20-0.80 threshold range, equalized-odds disparity varied substantially rather than remaining a fixed model property. For XGBoost, the absolute disparity ranged from

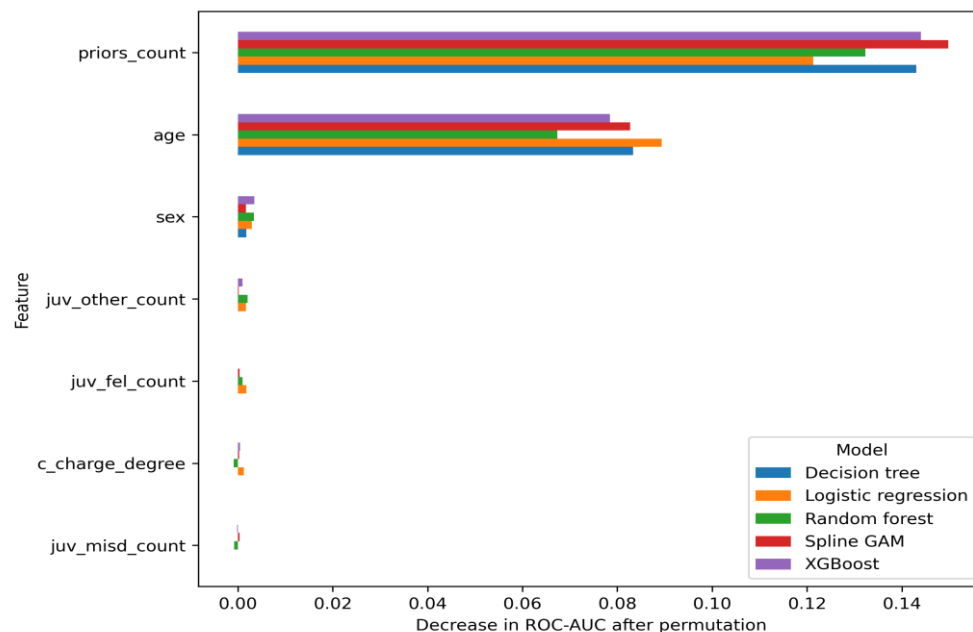
approximately 0.086 to 0.273; for the spline GAM it ranged from approximately 0.068 to 0.285. This reinforces that fairness conclusions depend on the operational threshold.

#### 4.4 Sex-based fairness sensitivity analysis

**Table 8. Secondary sex-based fairness sensitivity analysis at threshold 0.50. Signed differences are Male minus Female; equalized-odds difference is the maximum absolute TPR/FPR gap.**

| Model               | DP difference<br>(Male - Female) | EO difference<br>(Male - Female) | FPR difference<br>(Male - Female) | Equalized-odds<br>difference |
|---------------------|----------------------------------|----------------------------------|-----------------------------------|------------------------------|
| Logistic regression | 0.296                            | 0.335                            | 0.215                             | 0.335                        |
| Decision tree       | 0.214                            | 0.212                            | 0.157                             | 0.212                        |
| Random forest       | 0.219                            | 0.197                            | 0.172                             | 0.197                        |
| XGBoost             | 0.233                            | 0.231                            | 0.173                             | 0.231                        |
| Spline GAM          | 0.243                            | 0.218                            | 0.195                             | 0.218                        |

#### 4.5 Model interpretability and explanation stability



**Figure 4. Model-agnostic permutation importance across all five models, measured as decrease in held-out ROC-AUC after feature permutation.**

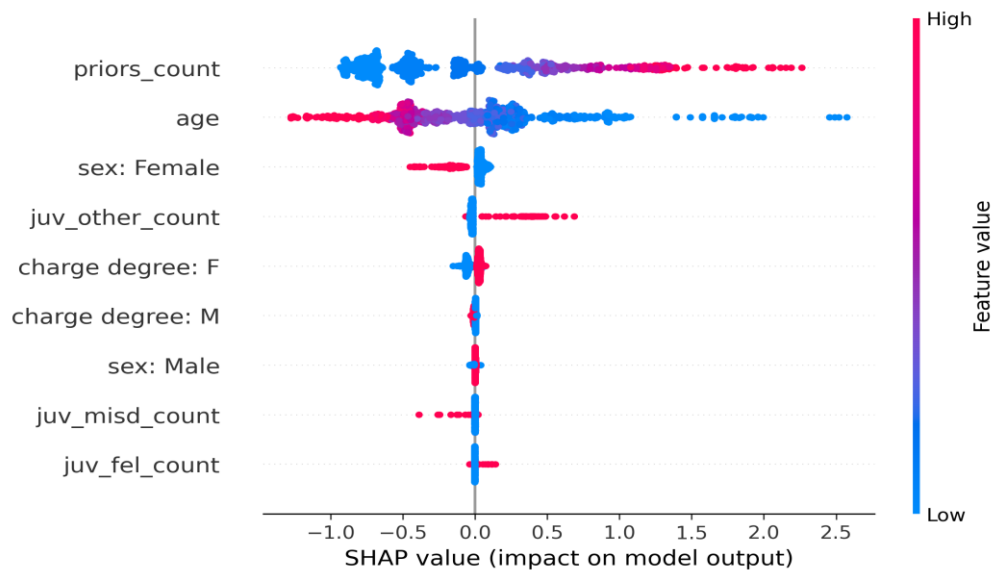


Figure 5. XGBoost SHAP beeswarm with actual transformed predictor names, showing the direction and magnitude of feature attributions.

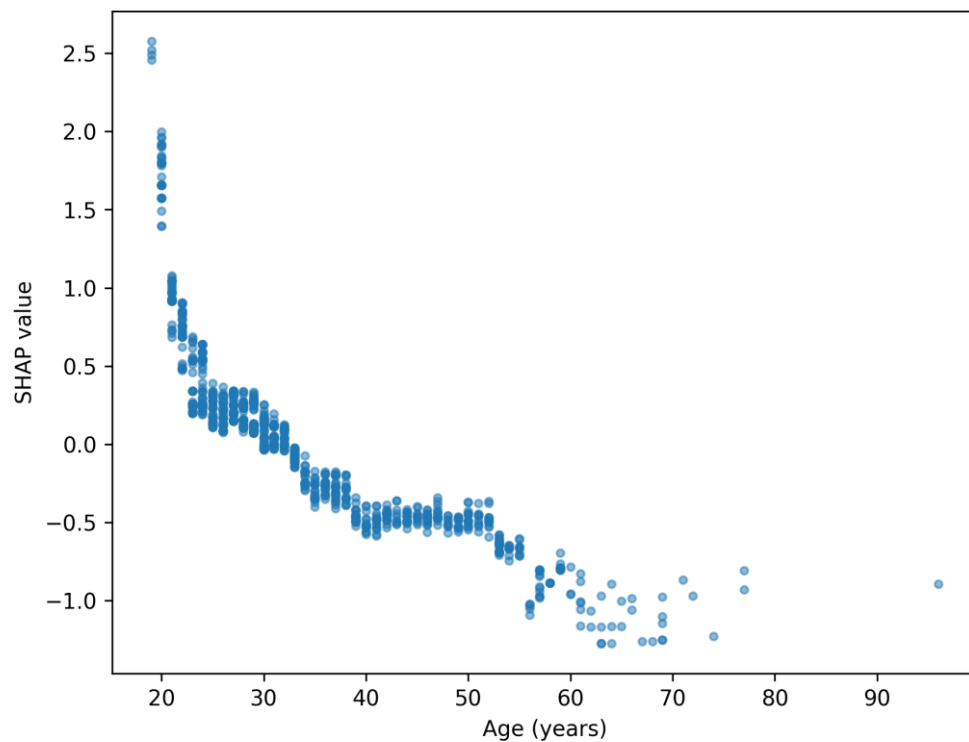
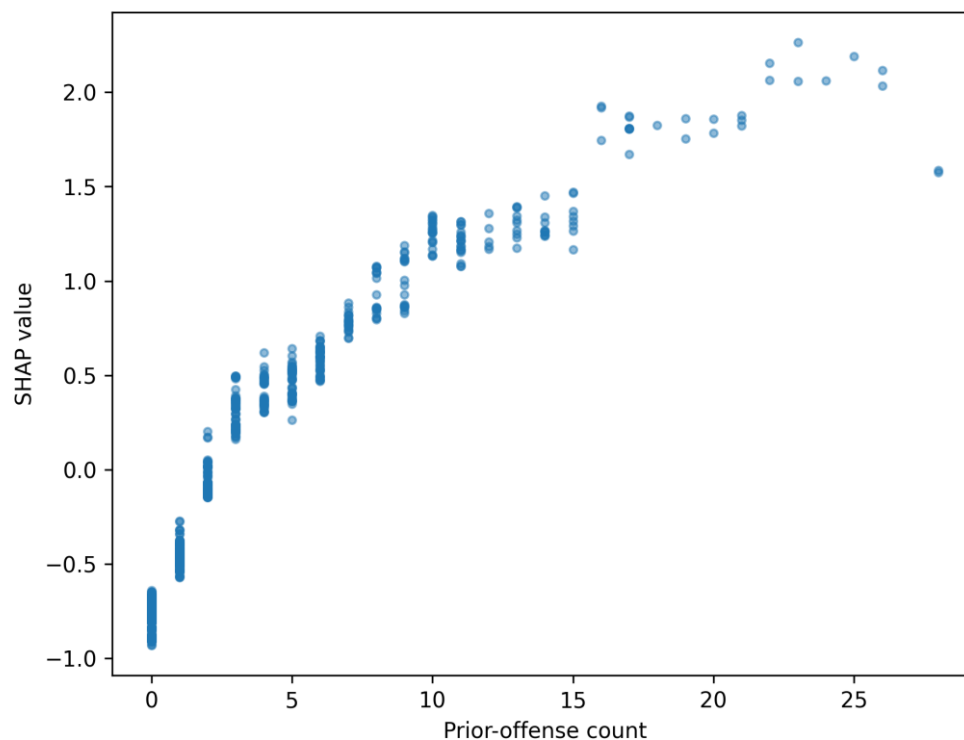


Figure 6. XGBoost SHAP dependence pattern for age, displayed in original years.



*Figure 7. XGBoost SHAP dependence pattern for priors\_count, displayed as the original prior-offense count.*

Permutation importance identified priors\_count and age as the dominant predictors across models. The corrected SHAP beeswarm likewise showed these features as the principal XGBoost contributors. Across five bootstrap refits, mean pairwise Spearman correlation of global SHAP importance rankings was 0.973 (SD 0.015), and mean top-five overlap was 1.00. The leading global attribution pattern was therefore highly stable under the resampling procedure used here, while still not implying causal or normative importance.

## 5. Discussion

This study found that substantially different modeling strategies produced similar discrimination on the COMPAS benchmark. In repeated validation, XGBoost and the spline GAM both achieved mean AUC of approximately 0.731, random forest reached 0.729, and logistic regression remained close at 0.724. This pattern is consistent with arguments that interpretable or semi-interpretable models can remain competitive on structured tabular high-stakes problems and that complexity should

be justified by measurable benefit rather than assumed superiority. [6,17,21]

The fairness findings are consistent with the central lesson of prior COMPAS research: aggregate predictive performance can coexist with materially different group error profiles. [2,8,25] Prior work has demonstrated tensions among fairness criteria when base rates differ and related incompatibilities among calibration and balance conditions. [8,9] In the present analysis, all five classifiers produced higher selection rates and FPRs for African-American than Caucasian observations at threshold 0.50. Importantly, the threshold analysis showed that the magnitude of equalized-odds disparity changed substantially as the operating point moved. The result therefore supports reporting the full operating context rather than labeling a model fair or unfair from one threshold and one metric.

The secondary sex analysis broadens the audit beyond race and shows why subgroup conclusions should not be generalized automatically. Contemporary fairness literature in health and other high-stakes settings likewise

recommends stratified evaluation because performance and error burdens can differ across populations and deployment contexts. [19,20,31]

The principal contribution of this study is the integration of several audit dimensions that are often examined separately. Rather than treating predictive discrimination, calibration, subgroup fairness, operating-threshold choice, and explanation quality as independent checks, the analysis evaluates them within one reproducible pipeline and adds bootstrap assessment of global explanation stability. This design makes it possible to distinguish a small gain in predictive discrimination from broader evidence about calibration, subgroup error burdens, threshold dependence, and the reproducibility of model explanations. [13,17,18,26,30]

The interpretability analysis extends beyond a single feature-ranking display. Model-agnostic permutation importance, a corrected SHAP beeswarm, dependence plots in original units, and bootstrap stability analysis were evaluated together. This responds to recent work arguing that explanation evaluation should include stability, fidelity, usability, and context rather than treating an attribution plot as sufficient evidence of trustworthy explanation. [13,18,23,24,26,27,28,29,30,32,33,34,35] The high SHAP ranking stability observed here is encouraging for reproducibility of the fitted explanation, but it does not establish causal validity, procedural justice, or human comprehensibility.

Calibration results further narrow the practical differences among models. Brier scores, calibration slopes, and ECE were broadly similar. The combined evidence therefore does not support a claim that the more opaque ensemble has a decisive predictive advantage. Instead, model selection in a high-stakes setting should consider discrimination, calibration, group-specific error costs, explanation properties, operational thresholds, and governance requirements together. [1,21,22,36]

## 6. Implications for High-Stakes AI Governance

A practical audit should proceed in layers: establish out-of-sample validity with repeated validation; evaluate calibration as well as discrimination; disaggregate errors across

relevant groups; examine threshold sensitivity; inspect intrinsic structure or post-hoc explanations and test explanation stability; and document preprocessing, hyperparameters, human review, appeal, monitoring, and retraining procedures. XAI is one component of a broader risk-management process rather than a standalone solution. [1,13,17,21,22,36]

## 7. Limitations

The COMPAS data describe a specific historical criminal-justice setting in Broward County, Florida and should not be treated as a current or universally representative population. The public variables also do not reproduce the proprietary COMPAS model. The outcome `two_year_recid` is observational and may reflect detection and enforcement processes as well as underlying behavior; the analysis therefore estimates predictive associations, not causal effects.

Although threshold sensitivity was examined from 0.20 to 0.80, no threshold is treated as normatively privileged. External validation on another high-stakes dataset remains necessary before generalizing the numerical findings. The fold-level statistical comparisons are exploratory because repeated cross-validation folds are dependent despite Holm adjustment for multiplicity. SHAP explains the fitted XGBoost model rather than the social process generating the data, and correlated or omitted predictors can affect attribution. [23,24]

The spline GAM was selected as an intrinsically interpretable nonlinear comparator because it allows smooth nonlinear effects of age and prior history while preserving an additive structure. It should not be interpreted as the proprietary COMPAS model or as an Explainable Boosting Machine (EBM). Future work could compare this specification with EBM and other interpretable boosting approaches under the same validation and fairness protocol.

## 8. Conclusion

Across repeated validation, mean AUC was 0.724 for logistic regression, 0.718 for the decision tree, 0.729 for random forest, and approximately 0.731 for both XGBoost and the spline GAM. Calibration differences were modest, demographic disparities persisted across

models, and their magnitude changed with threshold choice. XGBoost explanations were dominated by prior-offense count and age and showed high global ranking stability across bootstrap refits. These findings support a multidimensional approach to high-stakes AI evaluation in which predictive performance, calibration, subgroup error burdens, threshold behavior, and explanation quality are reported together and reproduced transparently before deployment decisions are made.

## REFERENCES

- Tabassi E. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. National Institute of Standards and Technology; 2023. doi:10.6028/NIST.AI.100-1.
- Angwin J, Larson J, Mattu S, Kirchner L. Machine Bias. ProPublica. 2016.
- ProPublica. COMPAS Analysis: Data and analysis for Machine Bias. Public repository.
- Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*. 2017;30.
- Ribeiro MT, Singh S, Guestrin C. Why Should I Trust You? Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016;1135-1144. doi:10.1145/2939672.2939778.
- Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*. 2019;1:206-215. doi:10.1038/s42256-019-0048-x.
- Hardt M, Price E, Srebro N. Equality of Opportunity in Supervised Learning. *NeurIPS*. 2016;29.
- Chouldechova A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*. 2017;5(2):153-163.
- Kleinberg J, Mullainathan S, Raghavan M. Inherent Trade-Offs in the Fair Determination of Risk Scores. *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. 2017;67:43:1-43:23. doi:10.4230/LIPIcs.ITCS.2017.43.
- Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Computing Surveys*. 2021;54(6):Article 115, 1-35. doi:10.1145/3457607.
- Mitchell M, Wu S, Zaldivar A, et al. Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 2019:220-229. doi:10.1145/3287560.3287596.
- Geburu T, Morgenstern J, Vecchione B, et al. Datasheets for Datasets. *Communications of the ACM*. 2021;64(12):86-92. doi:10.1145/3458723.
- Mohamed A, Abdelqader K, Shaalan K. Explainable Artificial Intelligence: A systematic review of progress and challenges. *Intelligent Systems with Applications*. 2025;28:200595. doi:10.1016/j.iswa.2025.200595.
- Černevičienė J, Kabašinskas A. Explainable artificial intelligence (XAI) in finance: a systematic literature review. *Artificial Intelligence Review*. 2024;57:216. doi:10.1007/s10462-024-10854-8.
- Mersha M, Lam K, Wood J, AlShami A, Kalita J. Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. *Neurocomputing*. 2024;599:128111. doi:10.1016/j.neucom.2024.128111.
- Cabitza F, Campagner A, Malgieri G, et al. Quod erat demonstrandum? Towards a typology of the concept of explanation for the design of explainable AI. *Expert Systems with Applications*. 2023;213:118888. doi:10.1016/j.eswa.2022.118888.

- De Bock KW, Coussement K, De Caigny A, et al. Explainable AI for Operational Research: A defining framework, methods, applications, and a research agenda. *European Journal of Operational Research*. 2024;317(2):249-272. doi:10.1016/j.ejor.2023.09.026.
- Velmurugan M, Ouyang C, Xu Y, Sindhgatta R, Wickramanayake B, Moreira C. Developing guidelines for functionally-grounded evaluation of explainable artificial intelligence using tabular data. *Engineering Applications of Artificial Intelligence*. 2025;141:109772. doi:10.1016/j.engappai.2024.109772.
- Zhao Y, Wang Y, Derr T. Fairness and Explainability: Bridging the Gap towards Fair Model Explanations. *AAAI*. 2023;37(9):11363-11371. doi:10.1609/aaai.v37i9.26344.
- Deck L, Schoeffler J, De-Arteaga M, Kühl N. A Critical Survey on Fairness Benefits of Explainable AI. *ACM FAccT*. 2024. doi:10.1145/3630106.3658990.
- Breidbach CF. Responsible algorithmic decision-making. *Organizational Dynamics*. 2024;53(2):101031. doi:10.1016/j.orgdyn.2024.101031.
- European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*. 2024;L 2024/1689, 12 July 2024.
- Huang X, Marques-Silva J. On the failings of Shapley values for explainability. *International Journal of Approximate Reasoning*. 2024;171:109112. doi:10.1016/j.ijar.2023.109112.
- Kelodjou G, Rozé L, Masson V, et al. Shaping Up SHAP: Enhancing Stability through Layer-Wise Neighbor Selection. *AAAI*. 2024;38(12):13094-13103. doi:10.1609/aaai.v38i12.29208.
- Lagioia F, Rovatti R, Sartor G. Algorithmic fairness through group parities? The case of COMPAS-SAPMOC. *AI & Society*. 2023;38:459-478. doi:10.1007/s00146-022-01441-y.
- Kaplan S, Uusitalo H, Lensu L. A unified and practical user-centric framework for explainable artificial intelligence. *Knowledge-Based Systems*. 2024;283:111107. doi:10.1016/j.knosys.2023.111107.
- Sovrano F, Vitali F. An objective metric for Explainable AI: How and why to estimate the degree of explainability. *Knowledge-Based Systems*. 2023;278:110866. doi:10.1016/j.knosys.2023.110866.
- Antoñanzas JM, Jia Y, Frank E, Bifet A, Pfahringer B. teex: A toolbox for the evaluation of explanations. *Neurocomputing*. 2023;555:126642. doi:10.1016/j.neucom.2023.126642.
- Jung J, Lee H, Jung H, Kim H. Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. *Heliyon*. 2023;9(5):e16110. doi:10.1016/j.heliyon.2023.e16110.
- Salih AM, Boscolo Galazzo I, Gkontra P, Rauseo E, Lee AM, Lekadir K, Radeva P, Petersen SE, Menegaz G. A review of evaluation approaches for explainable AI with applications in cardiology. *Artificial Intelligence Review*. 2024;57:240. doi:10.1007/s10462-024-10852-w.
- Chen RJ, Wang JJ, Williamson DFK, Chen TY, Lipkova J, Lu MY, Sahai S, Mahmood F. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering*. 2023;7:719-742. doi:10.1038/s41551-023-01056-8.
- Tiukhova E, Vemuri P, López Flores N, Island AS, Óskarsdóttir M, Poelmans S, Baesens B, Snoeck M. Explainable Learning Analytics: Assessing the stability of student success prediction models by means of explainable AI. *Decision Support Systems*. 2024;182:114229. doi:10.1016/j.dss.2024.114229.

Dieber J, Kirrane S. A novel model usability evaluation framework (MUsE) for explainable artificial intelligence. *Information Fusion*. 2022;81:143-153. doi:10.1016/j.inffus.2021.11.017.

Jahromi MNS, Muddamsetty SM, Jarlner ASS, Høgenhaug AM, Gammeltoft-Hansen T, Moeslund TB. SIDU-TXT: An XAI algorithm for NLP with a holistic assessment approach. *Natural Language Processing Journal*. 2024;7:100078. doi:10.1016/j.nlp.2024.100078.

Drapal P, Prudêncio RBC, Silva Filho TM. Towards explainable evaluation: Explaining predicted performance using local performance regions. *Applied Soft Computing*. 2024;167:112351. doi:10.1016/j.asoc.2024.112351.

Autio C, Schwartz R, Dunietz J, et al. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. National Institute of Standards and Technology; 2024. doi:10.6028/NIST.AI.600-1.