Uncertainty and Explainability for Equitable Recidivism Prediction: An Analysis of Algorithmic Asymmetry

Alessandro Del Prete[1⋆][0009−0004−5768−7867], Zahida Mashaallah[1], and Flora Amato[1][0000−0002−5128−5558]

[1] Dept. of Electrical Engineering and Information Technology (DIETI), University of Naples "Federico II", Via Claudio 21, 80125, Naples, Italy, {alessandro.delprete3, zahida.mashaallah, flora.amato}@unina.it

Abstract. This study presents an Uncertainty-Aware Explainable AI (XAI) pipeline for auditing the COMPAS recidivism dataset. By integrating a Tabular Transformer architecture with Monte Carlo dropout, the approach quantifies epistemic uncertainty and its influence on algorithmic fairness. Empirical analysis reveals a systematic uncertainty gap, with minority profiles and younger defendants exhibiting significantly higher epistemic variance. Additionally, the Transformer-based architecture captures more nuanced uncertainty patterns than the linear baseline, such as Logistic Regression, thereby supporting its increased complexity. Gradient-based counterfactuals are used to differentiate between actionable and structural features, establishing a computationally grounded framework for auditing judicial Automated Decision Systems (ADS). A novel uncertainty sensitivity analysis identifies key features, including age and juvenile records, as primary drivers of model ignorance, underscoring critical areas for data governance and bias mitigation. These findings underscore the importance of incorporating uncertainty quantification into judicial AI to promote accountability and reduce systemic bias.

Keywords: Explainable AI (XAI) · Uncertainty Quantification · Algorithmic Fairness · Recidivism Prediction.

1 Introduction

The convergence of deep learning and judicial administration has initiated a paradigm shift toward "*algorithmic governance*" [1–4]. In criminal justice, Automated Decision Systems (ADS) such as COMPAS are widely used to estimate recidivism risk [5]. These automated risk assessments influence key judicial decisions, including pretrial bail, parole, and sentencing, thereby positioning computational algorithms as central gatekeepers of individual liberty [6–8].

Delegating judicial authority to opaque models presents significant challenges to constitutional principles, including due process and equality before the law [9,

⋆ Corresponding author.

10]. Classical machine learning models often violate these principles by operating as opaque black boxes that learn complex, non-linear patterns from historical data, thereby perpetuating and amplifying structural racism and socio-economic disparities [11, 12]. While extensive algorithmic fairness research has established group equity criteria such as demographic parity and equalized odds [13–15], conventional auditing frameworks exhibit systemic limitations that undermine their democratic legitimacy [16–18].

Traditional classifiers frequently exhibit two systemic flaws: (i) deterministic delusion, in which models provide point-estimate probabilities without conveying knowledge boundaries [20–22]; and (ii) over-generalization bias, where individual virtuous behavior is suppressed in favor of group-level historical statistics [23–25]. To address these issues, a multi-tiered uncertainty-aware explainable artificial intelligence (XAI) pipeline is developed. This approach employs a tabular Transformer [26, 27] integrated with Monte Carlo (MC) dropout [28, 19, 29] to extract epistemic variance for each defendant. The pipeline is further enhanced with SHAP values [30] and gradient-based counterfactuals [31] to identify precise feature contributions and actionable defense strategies. Unlike previous studies that evaluate predictive performance, fairness metrics, or explainability separately, our objective is to integrate uncertainty estimation, feature attribution, and counterfactual auditing into a unified post-hoc evaluation framework.

The investigation reveals a pronounced "*uncertainty gap*": minority profiles are associated with higher epistemic uncertainty, indicating that the most severe errors occur where the model exhibits the greatest ignorance. The approach is further validated against a Logistic Regression baseline, and an uncertainty sensitivity analysis is introduced to identify the drivers of algorithmic ignorance.

2 Related Work

The application of predictive algorithms in judicial decision-making has led to extensive interdisciplinary research. Following ProPublica's investigation of COMPAS [5], early studies demonstrated that multiple fairness criteria—demographic parity, equalized odds, and predictive rate parity—cannot all be satisfied when base recidivism rates differ across groups [32]. Subsequent research introduced bias-mitigation strategies, including pre-processing [33], in-processing regularization [34], and post-processing calibration [14]. However, these methods generally assume static, deterministic models and often neglect uncertainty in model confidence. Concurrently, Explainable AI (XAI) has produced advanced methods for interpreting black-box models. LIME [27] and SHAP [30] offer local feature attributions, while counterfactual explanations [31] provide actionable feedback by identifying minimal changes required to alter a prediction. Simultaneously, deep learning for tabular data has evolved with architectures such as the FT-Transformer [26], which introduce self-attention mechanisms to categorical and numerical datasets that were traditionally managed by gradient-boosted trees.

However, the intersection of XAI, deep tabular models, and uncertainty quantification remains largely unexplored within legal contexts. Bayesian neural net-

works and approximations such as MC dropout [19] are widely adopted in high-stakes domains, including autonomous driving and medical imaging [29], yet they are seldom employed to audit structural societal bias. Theoretical studies indicate that epistemic uncertainty, defined as model ignorance resulting from insufficient representative data, disproportionately impacts marginalized groups [22]. This research addresses this gap by integrating tabular Transformers, MC dropout, and gradient-based counterfactuals into a unified, uncertainty-aware fairness auditing framework.

3 Methodology

Algorithmic asymmetry in the COMPAS dataset is rigorously characterized using an end-to-end analytical framework that integrates deep representation learning, probabilistic inference, and post-hoc adversarial explainability.

Figure 1 presents the complete architecture. The pipeline consists of four stages: (1) ingestion and preprocessing of tabular judicial data; (2) probabilistic risk modeling using a tabular Transformer and Monte Carlo dropout; (3) forensic auditing through cooperative game-theoretic attributions (SHAP) and gradient-based counterfactuals; and (4) synthesis of systemic biases, including error asymmetry, uncertainty gaps, and over-generalization.

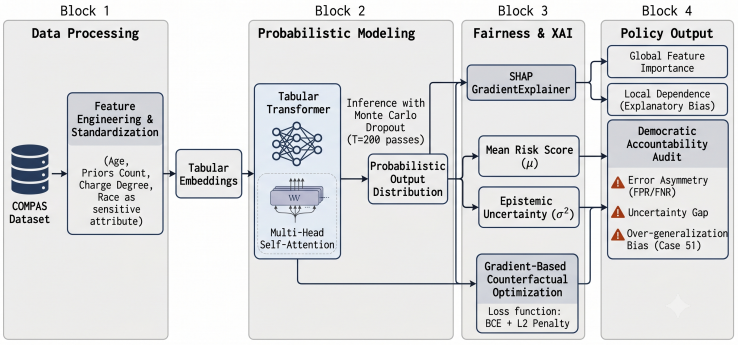


Fig. 1: Architectural flowchart of the proposed uncertainty-aware XAI pipeline. The diagram shows the data flow from COMPAS pre-processing through the Tabular Transformer, which extracts the predictive risk score (μ) and epistemic uncertainty (σ^2) via Monte Carlo dropout (T=200 passes), and concludes with post-hoc fairness audits (SHAP and counterfactuals).

3.1 Tabular Transformer Architecture

In contrast to traditional decision trees for tabular data, the predictive baseline employs a custom tabular Transformer with Multi-Head Self-Attention [26]. The

input $x \in \mathbb{R}^d$ denotes a tabular vector representing a defendant's profile, such as age, prior count, and charge degree. Each scalar feature $x^{(j)}$ is linearly projected into a shared d_{model}-dimensional embedding space:

$$e^{(j)} = W^{(j)}x^{(j)} + b^{(j)} \tag{1}$$

where, $W^{(j)}$ and $b^{(j)}$ are learned, feature-specific parameters. The resulting embeddings constitute the token sequence $E = [e^{(1)}, e^{(2)}, \ldots, e^{(d)}]$, which is processed by L Transformer encoder layers (with $L = 2$ in this study). Each layer consists of Multi-Head Attention and a Feed-Forward Network, incorporating layer normalization and residual connections. The final representation is pooled and fed into a dense classification head with a sigmoid activation to yield the predicted recidivism probability $\hat{y} \in [0, 1]$.

3.2 Epistemic Uncertainty via MC dropout

A standard deep neural network produces a single deterministic estimate, providing no mathematical characterization of uncertainty. To model epistemic uncertainty arising from model ignorance or an unrepresentative training submanifold, MC dropout is employed [19]. MC dropout maintains dropout layers active (rate $p = 0.25$) during inference. For each input x, $T = 200$ stochastic forward passes are performed, yielding an empirical predictive distribution. Let $\hat{y}_t$ denote the output probability from the t-th pass. The final expected probability of recidivism $\mu(x)$ is computed as the empirical mean:

$$\mu(x) = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_t \tag{2}$$

Epistemic uncertainty $\sigma^2(x)$ is quantified as the predictive variance across the T samples:

$$\sigma^2(x) = \frac{1}{T}\sum_{t=1}^{T}\left(\hat{y}_t - \mu(x)\right)^2 \tag{3}$$

Mapping the standard deviation $\sigma(x)$ across demographic subgroups enables dynamic auditing to assess whether the algorithm exhibits greater uncertainty when evaluating African-American defendants than Caucasian counterparts.

3.3 Global and Group-Level SHAP Attributions

To interpret the Transformer's complex, non-linear interactions, SHAP (SHapley Additive exPlanations) [30] is applied. The `GradientExplainer` method backpropagates gradients across attention heads. SHAP approximates game-theoretic Shapley values, allocating output variance to input features through the local explanation model g for a prediction $f(x)$:

$$g(z') = \phi_0 + \sum_{j=1}^{d}\phi_j z'_j \tag{4}$$

In this context, ϕ_0 denotes the global expected value, $z'_j \in \{0, 1\}$ indicates whether feature j is present in a coalition, and ϕ_j represents its marginal contribution. Aggregating local values and partitioning them by the sensitive attribute (Race) reveals explanatory biases, whereby the model disproportionately assigns larger feature contributions to factors such as prior crimes based solely on a defendant's demographic group.

3.4 Gradient-Based Counterfactual Optimization

Whereas SHAP explains why a decision is made, counterfactuals reveal how to reverse it by identifying the minimal profile changes required for a low-risk classification [31]. This process is framed as an adversarial optimization, in which a high-risk vector x_{orig} ($y_{orig} = 1$) is perturbed to obtain a vector x_{cf} that yields the target class ($y_{target} = 0$) by minimizing:

$$\mathcal{L}_{cf}(x_{cf}) = \mathcal{L}_{BCE}\left(f(x_{cf}), y_{target}\right) + \lambda \|x_{cf} - x_{orig}\|_2^2 \tag{5}$$

The binary cross-entropy (BCE) term directs the prediction toward the low-risk threshold, while the L2 penalty ($\lambda = 0.15$) maintains proximity to the original profile. This dual-objective approach quantifies the behavioral shifts required to overcome algorithmic stigma.

4 Experiments and Results

4.1 Experimental Setup and Pre-processing

The proposed pipeline was evaluated on the COMPAS dataset [5], with standard preprocessing filters applied to ensure data quality and statistical relevance. The analysis centers on the two most-represented demographics: African-American and Caucasian. The tabular Transformer was optimized using AdamW ($lr = 1e-3$) with a Cosine Annealing scheduler over 120 epochs. Epistemic uncertainty was captured using Monte Carlo dropout with a dropout probability of 0.25 and 200 stochastic passes. In response to concerns regarding model complexity, a Logistic Regression baseline was established for comparative performance analysis.

4.2 Empirical Analysis of Algorithmic Asymmetry

The pipeline identifies significant violations of judicial neutrality across predictive disparity, error asymmetry, and epistemic ignorance. As shown in Table 1, the Transformer achieves predictive performance comparable to the Logistic Regression baseline, yielding a marginally higher AUC (0.7190 vs. 0.7131) despite a slight reduction in overall accuracy. However, the results also reveal persistent observed biases. Importantly, the "*uncertainty gap*" is not simply a scalar difference. Aggregated analysis using distribution plots demonstrates that the African-American cohort exhibits a significantly higher density of high-variance predictions. Additionally, epistemic uncertainty is positively correlated with model

error, particularly in False Positive cases for minority defendants. These findings suggest that a higher false-positive burden is frequently driven by model ignorance rather than representative evidence.

Table 1: Fairness and Uncertainty Metrics across Racial Groups (COMPAS Dataset)

Metric	**Caucasian**	**African-American**	**Abs. Gap**
Positive Rate (Dem. Parity)	0.3025	0.5389	**0.2364**
False Positive Rate (FPR)	0.2242	0.3561	**0.1319**
False Negative Rate (FNR)	0.5846	0.2946	**0.2900**
Overall Accuracy	0.6282	0.6761	0.0479
Mean Uncertainty (σ)	0.0431	0.0461	0.0031
High Uncertainty Rate	0.1954	0.2871	**0.0918**
*Baseline: Logistic Regression (Accuracy: **0.6720**, AUC: **0.7131**)*			
*Model: Tabular Transformer (Accuracy: **0.6560**, AUC: **0.7190**)*			

Structural Disparities in Error Rates. The model exhibits a substantial demographic parity gap (23.64%) between African-American and Caucasian defendants, indicating marked disparities in positive risk predictions across demographic groups (Fig. 2).

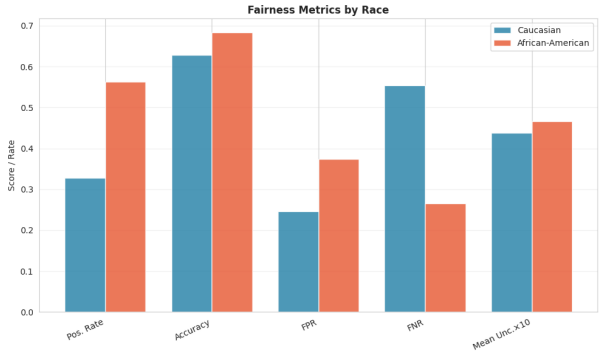


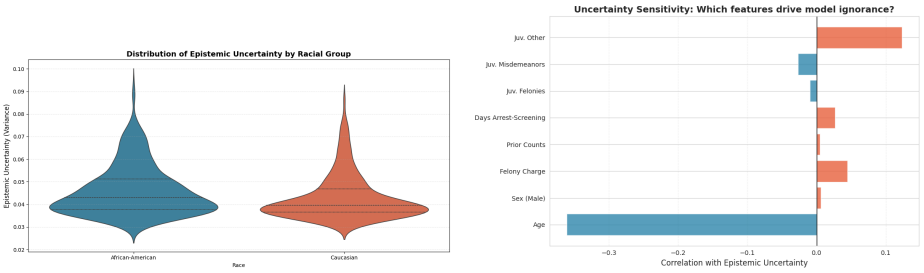
Fig. 2: Comparison of classical error rates by racial group. The analysis reveals a substantial False Negative Rate (FNR) gap (29.00%), indicating a systematic leniency toward Caucasian defendants. Conversely, the False Positive Rate (FPR) gap (13.19%) highlights the disproportionate over-punishment of African-American individuals.

A dual-error mechanism is identified as a driver of systemic bias:

1. *FNR Disparity (Differential Leniency)*: The false negative rate (FNR) is 58.46% for Caucasians compared to 29.46% for African-Americans, representing a 29.00% gap. This disparity indicates that the model is significantly more likely to incorrectly assign low risk to Caucasian defendants who subsequently recidivate.
2. *FPR Disparity (Over-punishment)*: The false positive rate (FPR) is 35.61% for African-Americans and 22.42% for Caucasians, resulting in a 13.19% gap. This suggests a higher likelihood of misclassifying African-Americans as high-risk, which may contribute to the imposition of harsher bail conditions.

Although these metrics provide valuable insights, they do not characterize the model's confidence boundaries. The following section presents a more detailed forensic audit by examining epistemic uncertainty and its relationship to these error patterns.

Quantifying the Epistemic Uncertainty Gap. In addition to point estimates, the integration of Monte Carlo (MC) dropout enables a formal mapping of the model's confidence bounds. As illustrated in Fig. 3, epistemic uncertainty (σ^2) is not uniformly distributed across demographic groups.



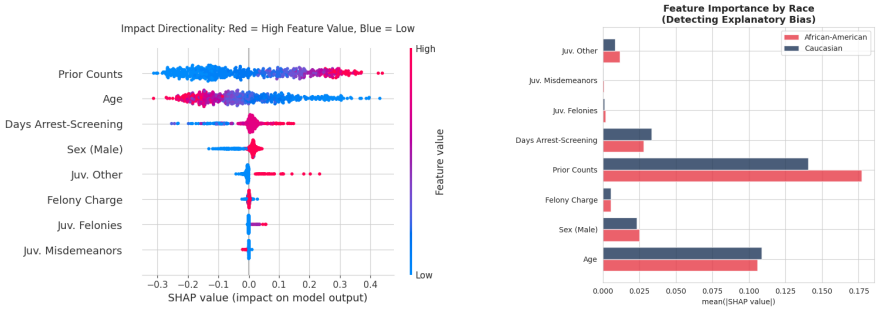
(a) Uncertainty Distribution (Violin Plot) (b) Uncertainty Sensitivity Analysis

Fig. 3: Analysis of Epistemic Ignorance. (a) The extended upper quartile for the African-American cohort reveals a systematic uncertainty gap (9.18% higher rate of high-variance predictions). (b) Correlation analysis identifies Age ($r \approx -0.35$) and juvenile records as the primary drivers of model ignorance.

Quantitatively, the proportion of predictions exceeding the 75th-percentile uncertainty threshold is 9.18% higher for African-American defendants (28.71%) than for Caucasian defendants (19.54%). This "*uncertainty gap*" is associated with insufficient representative data for specific sub-populations. As shown in Fig. 3b, age is identified as the primary negative driver of uncertainty, indicating that the model exhibits lower confidence when evaluating younger defendants. Given the dataset's demographic composition, this epistemic uncertainty is associated with the observed fairness disparities and may partially explain the heterogeneous error patterns across demographic groups.

4.3 Global XAI and Intersectional Bias

SHAP attributions were employed to interpret the internal logic of the Tabular Transformer [30]. The global summary plot (Fig. 4a) identifies Age and Priors Count as the dominant features influencing risk scores. To address the intersectional nature of algorithmic bias, the SHAP dependence of Priors Count on Race was further examined (Fig. 4).



(a) Global Feature Importance (b) Feature Importance by Race

Fig. 4: Global XAI Analysis via SHAP. (a) Ranking of predictive vectors, where Age and Priors Count dominate. (b) The plot compares feature attribution across racial groups, highlighting that the model relies more heavily on criminal history when evaluating African-American defendants than for Caucasians.

The results indicate systematic differences in feature attribution: for identical levels of prior offenses (e.g., 3 to 5 priors), the Transformer assigns larger positive SHAP contributions to African-American defendants than to Caucasian defendants. This non-linear interaction suggests that criminal history is weighted differently across demographic contexts. The ability to capture these context-dependent interactions provides complementary auditing insights beyond those obtainable from a linear baseline.

4.4 Qualitative and Systematic Auditing via Counterfactual Probes

To demonstrate methods for investigating algorithmic rigidity, gradient-based counterfactual generation is employed. Case 51 (an African-American false positive) is presented as a qualitative probe to illustrate the model's localized behavior (Fig. 5, Left). The optimizer calculates the necessary shifts to achieve a low-risk classification. For this defendant, the model requires significant alterations to non-actionable (immutable) features, such as artificially increasing age and erasing historical prior counts.

To ensure this observation is not an isolated anomaly due to selection bias, counterfactual generation is extended to a comprehensive sample of 52 African-American False Positive cases in the test set (Fig. 5, Right). The aggregate

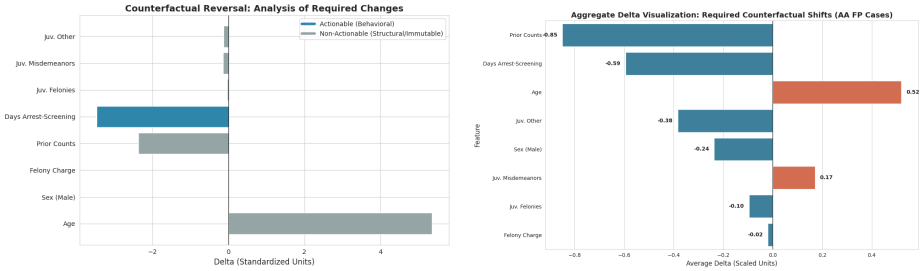


Fig. 5: Counterfactual Analysis. (Left) Delta Plot for Case 51, showing changes in non-actionable features (Age, Prior Counts) needed to reverse the high-risk prediction. (Right) Aggregate Delta Visualization for 52 African-American false positives. The large shifts in immutable features (-0.85 for Priors, $+0.52$ for Age) indicate a structurally limited algorithmic recourse beyond individual cases.

visualization confirms a consistent pattern: on average, the model requires substantial modifications to immutable attributes to grant recourse. For example, an average scaled shift of -0.85 for Prior Counts and 0.52 for Age. Although age and prior records are non-actionable in practice, their inclusion in the counterfactual optimization is highly revealing. This is not a methodological flaw but rather empirical evidence of structural constraints: the minority population may have limited actionable recourse under the current decision boundary.

5 Conclusion

This study presents an uncertainty-aware XAI pipeline utilizing deep tabular Transformers and Monte Carlo dropout to assess the COMPAS dataset. Results show that standard metrics may mask empirical disparities: a 29.00% False Negative Rate gap reflects systematic leniency toward Caucasian defendants, while a 13.19% False Positive Rate gap disproportionately impacts African-American individuals. In this experimental setting, the Transformer captures uncertainty patterns more effectively than the considered linear baseline. Age emerges as the main driver of model uncertainty, and counterfactual analysis highlights algorithmic constraints that prevent risk reclassification. These results underscore the reliance of algorithmic fairness on data quality and representativeness. This study has several limitations. The analysis is restricted to the COMPAS benchmark, epistemic uncertainty is estimated exclusively through MC dropout, and the counterfactual explanations intentionally include immutable attributes for auditing purposes rather than actionable recourse. Consequently, the findings should be interpreted as evidence within this experimental setting and validated on additional judicial datasets. Future work should address uncertainty disparities by improving loss functions and promoting transparency through robust data governance.

Acknowledgments. This work was partially supported by PNRR MUR Project PE0000013-FAIR. The FAIR project is committed to advancing an ethical, responsible, and sustainable vision of Artificial Intelligence, driving research and development in this crucial field, and consistently keeping ethical, legal, and sustainability considerations in mind.

References

1. Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. In The black box society. Harvard University Press.
2. Zuboff, S. (2024). The age of surveillance capitalism: The fight for a human future at the new frontier of power. Journal of Information Ethics, 33(1), 84-85.
3. Longworth, J. (2021). Benjamin Ruha (2019) Race After Technology: Abolitionist Tools for the New Jim Code. Medford: Polity Press. 172 pages. eISBN: 9781509526437. Science & Technology Studies, 34(2), 92-94.
4. Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. In Algorithms of oppression. New York university press.
5. Holm, S. H., & Lippert-Rasmussen, K. (2023). Discrimination, fairness, and the use of algorithms. Res Publica, 29(2), 177-183.
6. Cinnamon, J. (2020). Data inequalities and why they matter for development. Information Technology for Development, 26(2), 214-233.
7. Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science advances, 4(1), eaao5580.
8. Flores, A. W., Bechtel, K., & Lowenkamp, C. T. (2016). False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. Fed. Probation, 80, 38.
9. Binns, R. (2018, January). Fairness in machine learning: Lessons from political philosophy. In Conference on fairness, accountability and transparency (pp. 149-159). PMLR.
10. De Laat, P. B. (2018). Algorithmic decision-making based on machine learning from big data: can transparency restore accountability?. Philosophy & technology, 31(4), 525-541.
11. Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. Calif. L. Rev., 104, 671.
12. Luusua, A. (2023). Katherine Crawford: Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence: Yale University Press (April 6, 2021). Hardcover, 336 pages, ISBN: 9780300209570.
13. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference (pp. 214-226).
14. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in neural information processing systems, 29.
15. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6), 1-35.
16. Verma, S., & Rubin, J. (2018). Fairness definitions explained. In Proceedings of the international workshop on software fairness (pp. 1-7).

17. Farayola, M. M., Bendechache, M., Saber, T., Connolly, R., & Tal, I. (2026). Beyond aggregate fairness: intersectional auditing across the AI fairness pipeline. AI and Ethics, 6(1), 102.
18. Calders, T., & Verwer, S. (2010). Three naive bayes approaches for discrimination-free classification. Data mining and knowledge discovery, 21(2), 277-292.
19. Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In international conference on machine learning (pp. 1050-1059). PMLR.
20. Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. Nature, 521(7553), 452-459.
21. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems, 30.
22. Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Zhang, Y. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943.
23. Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research, 50(1), 3-44.
24. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91). PMLR.
25. Roy, A., Iosifidis, V., & Ntoutsi, E. (2022). Multi-fairness under class-imbalance. In International Conference on Discovery Science (pp. 286-301). Cham: Springer Nature Switzerland.
26. Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2021). Revisiting deep learning models for tabular data. Advances in neural information processing systems, 34, 18932-18943.
27. Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K. R. (Eds.). (2019). Explainable AI: interpreting, explaining and visualizing deep learning. Springer Nature.
28. Del Prete, A., & Mashaallah, Z. (2025). Towards Reliable Prognostics: RUL Uncertainty Estimation with Transformers and Monte Carlo Dropout. In International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (pp. 253-262). Cham: Springer Nature Switzerland.
29. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision?. Advances in neural information processing systems, 30.
30. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.
31. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv. JL & Tech., 31, 841.
32. Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data, 5(2), 153-163.
33. Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and information systems, 33(1), 1-33.
34. Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In Proceedings of the 26th international conference on world wide web (pp. 1171-1180).