ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Fairness in machine learning: A Compact Survey

Jérémy de Bodt[1], Dounia Mulders[2], Cyril de Bodt[3],
John A. Lee[2], and Marco Saerens[1,2] *

1- UCLouvain - LouRIM - Belgium
2- UCLouvain - ICTEAM - Belgium
3- UNamur - Dept. of Mathematics & naXys institute - Belgium

Abstract. Considering, assessing, and ensuring fairness is key when relying on machine learning (ML) for sensitive decision making. Yet, despite growing attention, multiple fairness definitions are currently adopted and consensual guidelines across learning paradigms are still lacking. Therefore, this review first describes sources of bias and focuses on outcome fairness, comparing popular criterion such as independence, separation, and sufficiency. Fairness interventions are then examined at pre-, in-, and post-processing stages and across supervised, semi-supervised, unsupervised, and self-supervised settings.

1 Introduction

Fairness has become a central concern as machine learning (ML) increasingly supports decisions in sensitive domains such as hiring, credit decisions, healthcare provision, or criminal case investigations. While ML promises efficiency and scalability, it may also reproduce or amplify existing biases when trained on unbalanced data or when deployed in settings where model predictions influence future data. Ensuring fair outcomes hence requires understanding how unfairness arises and how to mitigate it during learning [1, 2].

Fairness is usually considered from two perspectives [3]: **procedural fairness**, which evaluates how decisions are produced, and **outcome fairness**, which assesses whether the resulting decisions are equitable across groups or individuals. Because a seemingly neutral procedure can still yield discriminatory outcomes and fair outcomes are the central issue, this review focuses on outcome fairness, formalized by statistical principles of independence, separation, and sufficiency [1]. Throughout the paper, we rely on a stylized credit approval example to illustrate how fairness criteria and mitigation strategies operate within the ML pipeline.

Prior surveys have examined fairness concepts or mitigation techniques in isolation [2, 1]. In contrast, this work offers a concise, unified overview linking sources of bias, key fairness criteria, and implementation strategies at the pre-, in-, and post- processing stages. We also analyze how fairness connotations vary across supervised, semi-supervised, unsupervised, and self-supervised learning paradigms, in which label availability shapes both definitions and interventions. By studying these elements within a single framework, this survey elucidates trade-offs between model performance and equitable outcomes. This helps to clarify how fairness is defined and enforced in modern ML pipelines.

*JAL is a Research Director with the Belgian F.R.S.-FNRS.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

2 Typology of Biases

Fairness biases can enter an ML system at several points along the pipeline [2]. **Data-to-Algorithm bias** occurs when data distortions (under-representations, sampling or measurement errors, ...) are transmitted or amplified by the learning process. **Algorithm-to-User bias** arises when model outputs shape user behavior changing what users see, trust, or choose. **User-to-Data bias** captures how user behavior and societal preferences feed back into newly collected data, so that human biases become part of future training sets e.g,denied credit applicants being underrepresented in future repayment data. Because of this feedback loop, all types of bias can influence the decision process fairness. There is therefore no general mapping between bias types and the fairness metrics of Sect. 3. However, these metrics will allow us to address implementation issues in Sect. 4, both from the algorithmic perspective and across learning paradigms.

3 Fairness criteria

Fairness criteria assess whether ML algorithms lead to equitable decisions. They differ by the level at which fairness is evaluated (groups, subgroups, or individuals) [2] and by the nature of the statistical links between the model predictions $\hat{Y}$, the true labels Y, and the *sensitive* attribute A (e.g., race or gender). In what follows, random variables and observed values are respectively denoted by uppercase and lowercase letters; we assume that $Y \in \{0, 1\}$ for simplicity, A is categorical (and thus defines groups), and X are the *non-sensitive* features.

Group fairness ensures *balanced outcomes* across protected groups, while subgroup fairness requires the fairness condition to hold across other subpopulations, beyond the usual protected groups or their intersections (see [2] for examples), and individual fairness demands similar treatment for similar individuals. The notion of *balanced outcomes* itself is formalized through 3 core principles [2, 1]:

Independence requires predictions $\hat{Y}$ to be statistically independent of A, ensuring equal or close acceptance rates across groups:

$$\mathbb{P}(\hat{Y} = 1|A = a) = \mathbb{P}(\hat{Y} = 1), \text{ for all } a.$$

Separation holds when predictions are conditionally independent of A given Y, which implies equal true and false positive rates between groups:

$$\mathbb{P}(\hat{Y} = 1 \mid Y = y, A = a) = \mathbb{P}(\hat{Y} = 1 \mid Y = y), \text{ for all } y, a.$$

Sufficiency requires Y to be conditionally independent of A given $\hat{Y}$, ensuring equal predictive values across groups (calibration):

$$\mathbb{P}(Y = 1 \mid \hat{Y} = \hat{y}, A = a) = \mathbb{P}(Y = 1 \mid \hat{Y} = \hat{y}), \text{ for all } \hat{y}, a.$$

Under independence, each group should receive positive predictions at the same rate, regardless of their characteristics (e.g., same loan approval rate for all groups). Separation implies that, given the true outcome, predictions do not depend on group characteristics (e.g., among those who will repay, the model

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

#	Name	Level	Princip.	Required across groups	Formula, $\forall\ a$
1	Demographic Parity (DP)	Group	Indep.	Equal positive prediction rates (PPR)	$p_{\hat{Y}\mid A}(1\mid a) = p_{\hat{Y}}(1)$
2	Equal Opportunity (EOP)	Group	Sep.	Equal true positive rates (TPR)	$p_{\hat{Y}\mid Y,A}(1\mid 1,a) = p_{\hat{Y}\mid Y}(1\mid 1)$
3	Predictive Equality (PE)	Group	Sep.	Equal false positive rates (FPR)	$p_{\hat{Y}\mid Y,A}(1\mid 0,a) = p_{\hat{Y}\mid Y}(1\mid 0)$
4	Predictive Parity (PP)	Group	Suff.	Equal probability of correct positive predictions across groups.	$p_{Y\mid \hat{Y},A}(1\mid 1,a) = p_{Y\mid \hat{Y}}(1\mid 1)$
5	False Omission Rate Parity (FORP)	Group	Suff.	Equal negative predictive values (NPP)	$p_{Y\mid \hat{Y},A}(0\mid 0,a) = p_{Y\mid \hat{Y}}(0\mid 0)$
6	Overall Accuracy Equality (OaE)	Group	Indep.	Equal overall accuracy	$p_{\hat{Y}\mid A}(Y\mid a) = p_{\hat{Y}}(Y)$
7	Treatment Equality (Tr.Eq.)	Group	/	Equal ratio of false negatives to false positives	$\frac{p_{\hat{Y}\mid Y,A}(0\mid 1,a)}{p_{\hat{Y}\mid Y,A}(1\mid 0,a)} = \frac{p_{\hat{Y}\mid Y}(0\mid 1)}{p_{\hat{Y}\mid Y}(1\mid 0)}$
8	Conditional Statistical Parity (CSP)	Group	Indep.	Equal PPR for individuals sharing the same legitimate attribute L	$p_{\hat{Y}\mid L,A}(1\mid l,a) = p_{\hat{Y}\mid L}(1\mid l)$
9	Equalized Odds (EO)	Group	Sepa.	Equal TPR and FPR ; combines EOP and PE	$p_{\hat{Y}\mid Y,A}(1\mid y,a) = p_{\hat{Y}\mid Y}(1\mid y), y\in\{0,1\}$
10	Test Fairness (Calibration)	Group	Suff.	Same likelihood of the positive class for any $\hat{Y}=s$	$p_{Y\mid \hat{Y},A}(1\mid s,a) = p_{Y\mid \hat{Y}}(1\mid s)\ \forall s$
11	Counterfactual Fairness (CF)	Indiv.	/	Same prediction for an individual with features $X=x$ when A is counterfactually set to a'	$p_{\hat{Y}_{A\leftarrow a}\mid X,A}(y\mid x,a) = p_{\hat{Y}_{A\leftarrow a'}\mid X,A}(y\mid x,a)\ \forall y\in\{0,1\}, a'$
12	Equalized Correlation (EC)	Group	Sepa.	$\hat{Y}$ is uncorrelated with A conditional on Y	$\sigma_{\hat{Y}A}\,\sigma_Y^2 = \sigma_{\hat{Y}Y}\,\sigma_{YA}$, where $\sigma_{\alpha\beta} = \mathbb{E}[(\alpha-\mathbb{E}[\alpha])(\beta-\mathbb{E}[\beta])]$
13	Darlington Criterion 1 (DC1) (4 in total)	Group	Suff.	Partial correlation of A and Y controlling for $\hat{Y}$ is zero	$r_{A\hat{Y}} = \frac{r_{AY}}{r_{Y\hat{Y}}}$, with $r_{\alpha\beta} = \sigma_{\alpha\beta}/\sigma_\alpha\sigma_\beta$
14	Conditional Procedure Accuracy (CPA)[4]	Group	Sepa.	Equal conditional accuracy (TPR & TNR): same correct prediction rates	$p_{\hat{Y}\mid Y,A}(y\mid y,a) = p_{\hat{Y}\mid Y,A}(y\mid y,b), y\in\{0,1\}$
15	Avoiding Disparate Mistreatment (ADM)	Group	Sepa.	Equal false positive and false negative rates	$p_{\hat{Y}\mid Y,A}(y'\mid y,a) = p_{\hat{Y}\mid Y,A}(y'\mid y,b)\ \forall y\neq y'$
16	Fairness Through Unawareness (FTU)	Indiv.	/	Attributes A are not explicitly used in decisions	/
17	Fairness Through Awareness (FTA)	Indiv.	/	Similar individuals (under a task-specific metric) receive similar predictions	/
18	Balance for Negative Class (BNC)[5]	Group	Sepa.	Equal average predicted scores for truly negative individuals ($Y=0$)	$\mathbb{E}[\hat{Y}\mid Y=0,A=a] = \mathbb{E}[\hat{Y}\mid Y=0,A=b]$
19	Balance for Positive Class (BPC)[5]	Group	Sepa.	Equal average predicted scores for truly positive individuals ($Y=1$)	$\mathbb{E}[\hat{Y}\mid Y=1,A=a] = \mathbb{E}[\hat{Y}\mid Y=1,A=b]$

Table 1: Definitions and formulations of common group and individual fairness notions in the case of binary predictions, with $p_{\hat{Y}\mid A}(y\mid a) := \mathbb{P}(\hat{Y}=y\mid A=a)$.

predicts repayment across groups equally). Finally, under sufficiency, given $\hat{Y}$, Y does not depend on group characteristics anymore (e.g., among those who are predicted to repay, the true repayment rate should be the same across groups). Table 1 details the main fairness criteria, their levels, principles, and formulas.

4 Fairness implementation

Having identified fairness criteria and balancing principles, we now turn to implementation, translating ethical principles into practical mechanisms within the ML pipeline. Fairness can be implemented at different training stages and across

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

learning paradigms, from supervised to self-supervised learning.

4.1 Algorithmic perspective of fairness implementation

Fairness implementation in ML pipelines can take place in 3 stages: pre-, in-, and post-processing [2, 6].

Pre-processing mitigates data bias prior to training through operations such as suppression (FTU: removing sensitive or correlated attributes), relabeling (correcting biased labels), and reweighting (adjusting sample importance to increase influence of underrepresented groups). Decorrelation techniques further restrain the information from sensitive attributes among the features, ensuring that protected information does not leak into the model [1]. These methods are model-agnostic and suitable when algorithms cannot be modified[6].

In-processing techniques try to adapt learning to remove discrimination during training [2]. The *Prejudice Remover* [7] adds a regularization term to reduce the dependency between $\hat{Y}$ and A; it therefore primarily targets independence-based criteria such as DP. Besides, *Adversarial Debiasing* [1, 8] introduces a predictor adversary framework where the model minimizes error while an adversary tries to infer A, so that $\hat{Y}$ is accurate yet minimally informative about A. The framework is instantiated for three fairness definitions (DP, EOP, and EO).

Post-processing adjusts model outputs to satisfy fairness constraints without accessing training data, making it applicable at deployment [2] and is model agnostic. It typically adjusts predictions for selected samples to meet a group fairness criterion; for example, Reject Option Classification [9] alters decisions in an uncertainty band, while Equalized Odds post-processing randomizes predictions to equalize group error rates [10]. Although post-processing is sometimes criticized for requiring sensitive attributes and for weaker accuracy fairness trade-offs [11], recent work shows it can still outperform other strategies [12].

4.2 Fairness across learning paradigms

The notion of fairness in ML depends on the learning paradigm, since the availability of labels shapes how fairness is defined and enforced, and each paradigm poses distinct challenges for achieving equitable or unbiased outcomes.

In *supervised learning*, fairness can be enforced via training regularization/constraints, post-hoc thresholding or feature selection, and criteria like DP or EO are directly measurable through group-wise rates. A representative in-processing method is *adversarial debiasing*, as defined in Section 4.1. Another example is [13], an *in-processing* approach using logistic regression and enforces fairness by constraining the covariance between $\hat{Y}$ and A, a relaxed proxy for DP.

In *semi-supervised learning*, fairness is usually first enforced on a small labeled subset, then extended to the larger unlabeled population through pseudo-labeling. Methods such as Fair-SSL [14] build on state-of-the-art semi-supervised techniques to generate reliable pseudo-labels for the unlabeled data. Combined with Fair-SMOTE rebalancing, these pseudo-labeling methods transfer fairness constraints beyond the labeled subset, helping maintain group parity.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

In *unsupervised learning*, fairness shifts from outcome parity to representational neutrality. In dimension reduction (DR), methods such as Fair-NeRV and Fair-t-NeRV [15] yield data representations that provide no information about protected attributes while preserving local geometry, by adding a fairness term to the DR objective. Their fairness criterion is essentially a *local* analogue of DP. Another major line of work addresses *fair clustering*, which imposes balance constraints so that clusters remain demographically representative. Fairlet or fair coreset are often considered, consisting in data subgroups in which fairness is imposed as a pre-processing step [16]. The authors in [17] complements this view by defining a fair radius for k-clustering, imposing individual fairness during in-processing. They design a local search algorithm, coined bicriteria approximation, that simultaneously targets two objectives: (i) minimizing the clustering cost, and (ii) enforcing the fairness constraint up to a constant factor.

In *self-supervised learning* (SSL), models learn representations from unlabeled data via pretext tasks (e.g., contrastive or masked prediction). SSL fairness aims to limit the encoding of A during pre-training and its amplification in downstream fine-tuning. The study [18] shows, using a multi-stage evaluation with gradual unfreezing, that SSL often improves group fairness with minor accuracy lost.

4.3 Comparison across learning paradigms

To clarify the implementation challenges, we now focus on the distinction between supervised and semi-supervised methods on the one hand, and unsupervised learning and SSL on the other hand, building on the studies of Sect. 4.2.

In *supervised learning*, fairness can be evaluated and enforced directly thanks to labeled data (X, A, Y). Fairness in *semi-supervised learning* extends supervised approaches to treat unlabeled data.

In *unsupervised learning*, fairness is enforced by constraining learned representations rather than predictive accuracy [15, 16, 17]. SSL is typically employed to enforce fairness constraints before a subsequent supervised task. These methods can therefore also be casted as pre-processing steps for supervised algorithms.

To sum up, supervised and semi-supervised fairness are task dependent and expressed in terms of prediction outcomes since labels make disparities measurable. Methods act on the features to enforce criteria like DP while trying to preserve accuracy [13]. In contrast, unsupervised and self-supervised fairness targets the learned representations, constraining embeddings or clusters to avoid encoding protected attributes [15, 16, 17, 18]. Hence the former seeks equitable predictions, while the latter seeks neutral representations.

5 Conclusion

Fairness has become a central challenge for responsible ML. This brief review focuses on outcome fairness, examines how biases emerge across data, algorithmic, and societal levels, and summarize important statistical notions. Fairness interventions vary according to processing stages and learning paradigms, in which

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

equity takes different meanings with respect to accuracy and interpretability. The credit example showed fairness depends on paradigm and intervention stage.

Future work will aim to (1) better delineate accuracy-fairness trade-offs across implementation stages and learning paradigms, and (2) connect computational approaches with philosophical fairness theories, including egalitarianism, utilitarianism, and procedural justice [19], by complementing currently adopted criteria with fairness notions from the humanities, to promote more ethically grounded algorithmic decision-making.

References

[1] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

[2] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[3] M. Defrance and T. De Bie. Maximal combinations of fairness definitions. *Journal of Artificial Intelligence Research*, 82:1495–1579, 2025.

[4] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments: The state of the art. *Socio. Meth. & Research*, 50(1):3–44, 2021.

[5] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. In *ITCS*, pages 43–1. Schloss Dagstuhl, 2017.

[6] A. Nielsen. *Practical fairness*. O'Reilly Media, 2020.

[7] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *ECML-PKDD*, pages 35–50. Springer, 2012.

[8] B.H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[9] P. K. Lohia, K. N. Ramamurthy, M. Bhide, D. Saha, K. R. Varshney, and R. Puri. Bias mitigation post-processing for individual and group fairness. In *ICASSP*, 2019.

[10] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *NeurIPS*, 29, 2016.

[11] B. Woodworth, S. Gunasekar, M.I. Ohannessian, and N. Srebro. Learning non-discriminatory predictors. In *Conference on learning theory*, pages 1920–1953, 2017.

[12] A. F Cruz and M. Hardt. Unprocessing seven years of algorithmic fairness. *ICLR*, 2023.

[13] F.V. Vromman, S. Courtain, P. Leleux, C. de Schaetzen, E. Beghein, A. Kneip, and M. Saerens. Maximum entropy logistic regression for demographic parity in supervised classification. In *Benelearn*, pages 189–208. Springer, 2023.

[14] J. Chakraborty, H. Tu, S. Majumder, and T. Menzies. Can we achieve fairness using semi-supervised learning? *arXiv preprint arXiv:2111.02038*, 2022.

[15] J. Peltonen, W. Xu, T. Nummenmaa, and J. Nummenmaa. Fair neighbor embedding. In *ICML*, pages 27564–27584. PMLR, 2023.

[16] A. Chhabra, K. Masalkovaitė, and P. Mohapatra. An overview of fairness in clustering. *IEEE Access*, 9:130698–130720, 2021.

[17] S. Mahabadi and A. Vakilian. Individual fairness for k-clustering. In *ICML*, pages 6586–6596. PMLR, 2020.

[18] S. Yfantidou, D. Spathis, M. Constantinides, A. Vakali, D. Quercia, and F. Kawsar. Using self-supervised learning can improve model fairness. In *ACM SIGKDD*, 2024.

[19] J. Ochmann, L. Michels, V. Tiefenbeck, C. Maier, and S. Laumer. Perceived algorithmic fairness: An empirical study of transparency and anthropomorphism in algorithmic recruiting. *Information Systems Journal*, 34(2):384–414, 2024.