International Conference on Multidisciplinary Perspectives in Advanced Computing and Technology (IMPACT 2026)
G. B. Pant University of Agriculture and Technology, Uttarakhand, India. Jan. 10-11, 2026

Algorithmic Bias and Fairness in Machine Learning Systems: A Review

Tooba Fatima, Kashifa Khanam, Abhishek Jaiswal, Khanak Saxena, Madhuresh Yadav, Mohammad Jeelani and Abhishek Saxena

Department of Computer Application, Future University, Bareilly, U.P. India

ftooba831@gmail.com, abhishekjs0230@gmail.com, khankashifa9068@gmail.com, jeelani.0018@gmail.com

Abstract

Machine learning (ML) systems are increasingly deployed in sensitive domains such as healthcare, finance, and criminal justice. Despite their benefits, these systems often exhibit algorithmic bias, raising concerns about fairness, accountability, and trust. Bias may emerge from historical inequalities in training data, model design, or deployment practices, resulting in disparate impacts on marginalized groups. Over the years, researchers have proposed multiple fairness definitions and mitigation strategies, ranging from data preprocessing and in-processing adversarial learning to post-processing adjustments. Toolkits like AI Fairness 360 and Fairlearn support practical implementation, while regulatory frameworks such as the EU AI Act emphasize governance and accountability. This survey consolidates theoretical foundations, technical approaches, domain-specific applications, and policy perspectives to provide a comprehensive understanding of algorithmic bias and fairness in ML, highlighting open challenges and future research directions.

Keywords: *Algorithmic bias, fairness, machine learning, fairness metrics, bias mitigation, AI ethics, governance.*

1. Introduction

Machine learning (ML) systems are increasingly embedded in decision-making processes across domains such as criminal justice, finance, healthcare, and employment. While these systems promise efficiency and scalability, they also raise critical concerns regarding algorithmic bias and fairness. Bias in ML can arise from historical inequalities embedded in training data, model design choices, or deployment contexts, potentially amplifying existing social disparities [1], [8].

The study of fairness in ML has gained momentum over the past decade, with foundational works highlighting both the sources of bias and the conceptualization of fairness. Barocas and Selbst [1] provided one of the earliest critical examinations of disparate impact in big data applications, showing how seemingly neutral algorithms can result in systemic discrimination. Dwork et al. [2] introduced the notion of "fairness through awareness," emphasizing individual-level fairness guarantees, while Hardt et al. [3], [14] proposed "equality of opportunity," ensuring equal treatment across groups in supervised learning tasks. However, Kleinberg et al. [4] demonstrated inherent trade-offs between fairness criteria, underscoring the impossibility of simultaneously satisfying multiple fairness definitions in risk assessment models. To address these challenges, a variety of technical and policy-driven approaches have been developed. Preprocessing techniques that modify training data to remove discriminatory patterns [13] and adversarial learning methods to mitigate unwanted bias during model training [15] represent key technical strategies. Comprehensive surveys [5], [16] have catalogued fairness definitions, measurement techniques, and mitigation methods, providing a roadmap for both researchers and practitioners. Toolkits such as IBM's AI Fairness 360 [6] and Microsoft's Fairlearn [7] operationalize these methods, enabling systematic bias detection and correction in deployed ML systems.

Beyond technical interventions, empirical studies and public debates have highlighted the social consequences of algorithmic bias. ProPublica's analysis of the COMPAS risk assessment tool revealed significant racial disparities [8], sparking widespread discussion on accountability in AI. Recent works also contextualize fairness in localized settings. For example, Bhatt et al. [9] and Girhepuje et al. [10] explore fairness in natural language processing (NLP) and legal AI systems within India, while Sahoo et al. [11] introduce IndiBias, a benchmark dataset tailored to the Indian socio-cultural context. These works highlight the importance of situating fairness debates within specific geopolitical and cultural frameworks. At the governance level, legal and policy frameworks such as the European Union's AI Act [12] have emphasized the need for transparency, accountability, and fairness in high-risk AI systems. These regulatory developments stress that fairness cannot be addressed solely through technical fixes but must also be embedded within organizational processes, ethical considerations, and societal oversight. In this paper, we build on these interdisciplinary insights to critically examine algorithmic bias and fairness in ML systems. We synthesize technical approaches, domain-specific challenges, and policy frameworks to highlight the tensions, trade-offs, and opportunities in creating more equitable machine learning systems.

2. Literature Review

Over the past few decades, there has been a significant issue of algorithmic bias and fairness in machine learning, leading to a diverse and evolving body of research. foundational work has laid the groundwork for understanding how machine learning systems can commemorate societal inequalities. Barocas and Selbst [1] argue that big data and algorithmic decision-making affect many groups and current anti-discrimination laws, as they focus on intentional discrimination, whereas algorithmic discrimination is mostly unintentional. On this basis, Dwork et al. [2] introduced the concept of" fairness through awareness", presenting that an algorithm should treat similar individuals similarly, while clearly focusing on sensitive attributes to verify fairness. Hardt et al. [3] focus on Equality of Opportunity, commending for equal true positive rates across various groups as a fairness metric. Meanwhile, Kleinberg et al. [4] mathematically display the incompatibility of some fairness criteria-for example, calibration and equal rates - focusing on the inherent trade-offs that must be weighed in fair algorithm design. Mehrabi et al. [5,16] provide a clear summary of bias in machine learning, including historical, representational, and measurement biases. They mainly focus on key fairness concepts - group, individual, and erroneous fairness - and summarise bias mitigation concepts applied before, during, and after model training. To promote fairness in real-world systems, researchers and many companies have developed tools such as AIFairness 360 by IBM [6] and fairlearn [7,18]. These include many tools, such as ready-made fairness measures and concepts to reduce bias, making it easier for developers to build fairer machine learning models. Observational studies state that real-world AI bias. ProPublica [8] found that the COMPAS system biasedly predicted higher backsliding for Black defendants. Dastin [19] showed Amazon's hiring tool favoured male resumes. Buolamwini and Gebru [20] discovered facial recognition system had the highest error rates for black women(racism), underlining the need for intersectional fairness. Past western settings, recent research shows the need for fairness frameworks that fit local cultures. Bhatt et al. [9], Girhepuje et al. [10], and Sahoo et al. [11] focus on fairness measures developed in the West that often ignore key social factors in India, such as language, caste, and gender norms. They display that models trained on Western Data can be biased in the Indian context. To show this, they create local datasets and tools, for example, IndiBias [11], to help build fairer AI systems in India. Similarly, the UCI

Indian Liver Patient Dataset [17], often used in fairness studies, may contain unseen socioeconomic biases that can affect model outcomes. On the policy side, the EU AI Act [12] proposes rules to manage AI risk, pointing to more transparency, fairness, and human oversight-mainly in high-risk systems. This shows a global move to make AI more accountable and ethical. On the technical side, researchers have developed ways to reduce bias in AI. Kamiran and Calders [13] developed methods to remove bias from data before training. Zhang et al. [15] used adversarial learning to hide sensitive information so that it doesn't affect any predictions. Ribeiro et al. [14] introduced LIME, which is a tool that explains individual model decisions, helping spot and fix biased behaviour. Together, this growing body of research shows that machine learning is highly complex and layered. Whereas theoretical foundations give useful definitions. More studies now stress that fairness depends on local context, as bias is shaped by culture, history, and society. Moreover, AI fairness will require both technical solutions and teamwork across fields, along with policies that address local needs, so that AI benefits everyone fairly.

3. Concrete Research Gaps

- **Formal frameworks for contextual fairness selection**: Methods to map social/ethical requirements and legal norms onto formal fairness objectives; decision-procedures that recommend the right metric(s) per context.
- **Robust fairness under distribution shift and small subgroup samples**; Algorithms that maintain fairness guarantees under covariate or label shift, or when protected/intersectional groups are data-scarce.
- **Causal, actionable interventions:** Practical causal discovery + counterfactual methods that require less unrealistic domain knowledge and that can be validated empirically.
- **Benchmarks & evaluation suites with deployment realism**: Datasets and simulation environments that include temporal dynamics, feedback loops, economic incentives, and intersectional subgroup labels to stress-test methods.
- **Auditability & provable fairness guarantees**: Scalable verification techniques, certifications, and logging standards that make audits meaningful without exposing proprietary data
- Fairness in foundation models / LLMs / multimodal pipelines. Techniques for pretraining, instruction tuning, and fine-tuning that reduce societal biases while preserving model utility; evaluation metrics suited to generative behavior.
- **Operational governance: tools for continuous monitoring and red-teaming**: Automated monitoring pipelines, drift detection for fairness metrics, and standardized counterfactual red-teaming protocols.
- **Human-centred methods & participatory design**: Co-design methods where affected communities participate in defining harms, metrics, and acceptable trade-offs.

4. Key Definitions and Formalization

Algorithmic bias in machine learning refers to systematic and unfair discrimination that results when model predictions disproportionately disadvantage certain individuals or groups [1], [8]. Bias may stem from historical inequities in training data, model assumptions, or deployment contexts. Fairness has therefore been formalized through multiple definitions that capture ethical, social, and statistical perspectives.

Individual fairness requires that *similar individuals receive similar predictions* [2]. Formally, if

$$d(x_i, x_j)$$

is a distance metric between individuals, then predictions should satisfy:

$$|h(x_i) - h(x_j)| \leq \epsilon \quad \text{whenever} \quad d(x_i, x_j) \text{ is small.}$$

Group fairness ensures parity across demographic groups. Two widely used measures are:

- **Demographic Parity (Statistical Parity):**

$$P(\hat{Y} = 1 \mid A = 0) = P(\hat{Y} = 1 \mid A = 1),$$

where $\hat{Y}$is the predicted label and **A** is a protected attribute (e.g., gender, race).

- **Equality of Opportunity** [3], [14]:

$$P(\hat{Y} = 1 \mid Y = 1, A = 0) = P(\hat{Y} = 1 \mid Y = 1, A = 1),$$

ensuring equal true positive rates across groups.

Kleinberg et al. [4] proved that achieving multiple fairness constraints (e.g., calibration, balance, and equalized odds) simultaneously is often impossible, underscoring inherent trade-offs.

Mitigation approaches are categorized as preprocessing (bias removal from data) [13], in-processing (adversarial debiasing [15]), and post-processing (output adjustments). Toolkits such as AI Fairness 360 [6] and Fairlearn [7] operationalize these definitions, while surveys [5], [16] consolidate the theoretical landscape as shown in Table 1.

Table 1: Measurement: Metrics and Evaluation

Author(s)	**Topic**	**Key Findings**	**Domain**	**Protected Attribute (s)**	**Fairness metric(s)**
Bhatt et al. [9]	Re-contextualizing Fairness in NLP	Sociolinguistic facts must be considered in account for fairness in India. Indian diversity can be disregarded by Western metrics. Caste and gender biases can be embedded in regional linguistic variance.	Natural Language Processing (NLP)	Region and Religion	Demographic Parity and Contextual Fairness (linguistic and cultural)
Girhepuje et al. [10]	Are Models Trained on Indian Legal Data Fair?	There is a significant bias against female participants in legal models	Legal(Bail Prediction)	Gender and legal role	Equalized Odds and Error Rate Disparity

		trained on Indian court data. Systemic bias is present in the legal language now in use.			
Sahoo et al.[11]	IndiBias Benchmark	According to benchmarks, the vast majority of models trained on Indian datasets reinforce social biases. Provide recent datasets to improve the accuracy of fairness evaluations.	Language Models(Benchmarking)	Gender, Region, Caste, Religion	Counterfactual Fairness , Disparate Impact and Intersectional fairness metrics
Healthcare study: ILPD (Dataset, 2022) [17]	Liver Disease Prediction	Preliminary models indicated an increased false-negative rate among female patients. With the implementation of fairness-aware sampling and thresholding, inequities diminished. Explainability showed that age was the preeminent predictor, with bilirubin level coming next	Healthcare (Liver Disease Prediction	Gender, Age Groups	Equal Opportunity, TPR/FPR Difference, Disparate Impact and SHAP-based subgroup analysis

5. Algorithmic Bias Mitigation Strategies

In the model development pipeline, machine learning bias can be mitigated at different stages. These are commonly categories into three types

1. Pre-processing method
2. In the processing method
3. Post-processing method

A. Pre-processing Strategies

It involves modifying the data before training to ensure fairness and reduce bias.

- Data Re-sampling: oversampling minority groups or undersampling the majority group.
- Re-weighting data: assign sample importance to reduce data.
- Features modification: removing sensitive attributes.[13]

B. In-processing Strategies

In-processing strategies focus on adjusting the learning algorithm itself during training to reduce bias.

- Fair constraint: Adding a regularisation to penalise outcomes.
- Adversarial Debiasing: remove correlation with sensitive attributes
- Fairness-Aware Algorithms: Some algorithms are designed with fairness built in, such as:
- Fair decision trees.[15]

C. Post-processing Strategies

Applied after a model has been trained, to adjust its output to ensure fairness.

- Fair Threshold: involves adjusting the decision thresholds for different groups to ensure equalize.
- Re-ranking: modify output for fairness in search. [3]

D. Explainability and Transparency

Use interpretable models and explanation tools (SHAP, LIME) to identify bias.Helps improve accountability and trust.[14]

6. Tools And Practical Resources

- **AI Fairness 360 (AIF360):** Created by IBM Research, AI Fairness 360 (AIF360) is a comprehensive Python library for evaluating and mitigating algorithmic bias[6]. It includes 70+ fairness metrics, supports pre-, in-, and post-processing techniques, and provides benchmark datasets (e.g., Adult Income, COMPAS), making it a widely used tool for fairness research and reproducibility.

- **Fairlearn:** Fairlearn, developed by Microsoft [18], is a Python toolkit designed to assess and improve fairness in machine learning models. It includes tools like fairness evaluation dashboards and algorithms such as Grid Search and Exponentiated Gradient for bias mitigation. Fairlearn helps practitioners balance predictive accuracy with fairness by focusing on user-centered model selection.

7. Representative Case Studies

COMPAS Algorithm in Criminal Justice

The COMPAS tool, used by U.S. courts to predict if someone might reoffend, was found to be racially biased. A 2016 ProPublica investigation showed that Black defendants were nearly twice as likely as White defendants to be wrongly labelled as high-risk. This revealed challenges in balancing different fairness measures and the dangers of relying on black-box algorithms in important decisions.

7.1. Gender Bias in Automated Hiring

Between 2014 and 2017, Amazon developed an AI hiring tool that showed bias against resumes containing terms like “women’s” or referencing women’s colleges Dastin, [19]. The bias stemmed from historical data reflecting past hiring discrimination. The project was shut down before launch, highlighting the risks of biased training data in automated hiring systems.

7.2. Bias in Facial Recognition Systems

Buolamwini and Gebru[20] found that commercial facial recognition systems had much higher error rates for darker-skinned women (34%) compared to lighter-skinned men (less than 1%). The bias was due to the lack of diversity in the training data. Their findings sparked public concern and led to temporary bans on facial recognition by some U.S. law enforcement agencies.

8. Legal, Ethical and Policy Context

Algorithmic fairness is increasingly central in both academic and regulatory discourse. Dwork et al. [2] introduced foundational fairness concepts, such as individual fairness, which holds that similar individuals should be treated similarly. This ethical framing underpins much of subsequent work, such as equality of opportunity by Hardt et al. [3], which formalizes fairness using conditional independence between model predictions and protected attributes. Kleinberg et al. [4] advanced this by showing inherent trade-offs among fairness definitions, emphasizing that not all criteria can be simultaneously satisfied. Ethically, Barocas and Selbst [1] highlighted how big data systems can unintentionally reproduce structural discrimination, even when explicit intent is absent. This echoes the concerns raised by ProPublica’s (2016) exposé [8] on COMPAS, which revealed racial bias in recidivism-prediction tools, sparking global scrutiny of algorithmic fairness. As shown in Table 2.

From a legal and policy perspective, the EU AI Act [12] represents a landmark regulatory effort to classify AI systems by risk and to mandate transparency, accountability, and bias mitigation, particularly in high-risk areas such as healthcare and justice. In India, while no comprehensive AI regulation exists yet, researchers like Bhatt et al. [9] and Girhepuje et al. [10] have contextualized fairness within Indian socio-cultural realities, stressing the need for caste, regional, and gender-sensitive metrics. The IndiBias benchmark [11] is a step toward building datasets that can surface and measure these contextual biases. Technological toolkits like AI Fairness 360 by IBM [6] and Microsoft's Fairlearn [7] have operationalized fairness metrics into usable libraries, empowering practitioners to diagnose and mitigate bias. Complementing this, Mehrabi et al. [5] provide a comprehensive survey that outlines bias sources, mitigation techniques, and fairness metrics,forming a bridge between ethics and implementation.

Table 2: Comparison Table

Reference	**Type**	**Focus / Contribution**	**Region / Context**	**Relevance to ML Fairness**
Barocas & Selbst (2016)[1]	Legal / Ethical	Disparate impact in big data; anti-discrimination law	U.S./Global	Highlights legal risk of unintentional bias in data-driven decisions
Dwork et al. (2012)[2]	Ethical / Theoretical	Defines individual fairness ("treat similar people similarly")	Theoretical / Global	Foundational fairness principle
Hardt et al. (2016)[3]	Ethical / Algorithmic	Defines **Equality of Opportunity** in supervised learning	Global	Introduces group fairness aligned with equal TPR
Kleinberg et al. (2016)[4]	Ethical / Theoretical	Shows incompatibility between fairness definitions	Global	Highlights trade-offs between fairness metrics
Mehrabi et al. (2019)[5]	Survey / Ethical	Taxonomy of bias types, sources, and mitigation strategies	Global	Consolidates technical and ethical literature
Bellamy et al. (2018)[6]	Technical Toolkit	IBM's AI Fairness 360: implements various fairness metrics & debiasing methods	Global	Practical tool for bias detection and mitigation
Microsoft Fairlearn (JMLR)[7]	Technical Toolkit	Fairlearn toolkit for group fairness via constraints & post-processing	Global	Supports developers in fair model development
ProPublica (2016)[8]	Investigative Journalism	Revealed racial bias in COMPAS recidivism tool	U.S.	Triggered public debate on AI bias
Bhatt et al. (2022)[9]	Ethical / Regional NLP	Proposes fairness framework contextualized to Indian sociolinguistic landscape	India	Shows Western metrics fail in multilingual Indian context
Girhepuje et al. (2023)[10]	Legal AI / Empirical	Audits fairness of legal NLP models trained on Indian court data	India	Finds systemic gender bias in Indian legal predictions
Sahoo et al. (2024)[11]	Dataset / Benchmark	Introduces IndiBias: fairness benchmark	India	Enables fairness auditing across Indian datasets

		with Indian demographic attributes		
EU AI Act (2023) [12]	Legal / Regulatory	AI risk classification; mandates transparency, accountability, fairness	European Union	Enforces fairness in high-risk AI (e.g., healthcare, justice)

9. Challenges on Algorithm Bias and Fairness in Machine Learning Systems

i. Conflicting fairness definitions & unavoidable trade-offs

Different formal fairness criteria (demographic parity, equalized odds, predictive parity, individual fairness, counterfactual fairness, etc.) are often mutually incompatible in realistic settings; choosing one forces trade-offs with others and with utility. This makes selecting the "right" fairness formalism application-dependent and philosophically hard.

ii. Data limitations: bias sources, label noise, and distribution shift

Many fairness failures originate in data (historical discrimination, sampling bias, annotation bias) and are amplified during training and deployment. Real-world data are noisy, incomplete, or non-stationary, leading to bias that reappears or worsens after deployment (bias amplification, feedback loops).

iii. **Evaluation gap,** benchmarks vs real-world complexity

Existing fairness benchmarks and metrics often fail to capture intersectional harms, long-term societal effects, or domain-specific constraints (e.g., healthcare, criminal justice, finance). Benchmarks may give a false sense of safety because they omit deployment context, changing populations, or economic incentives.

iv. Explainability, accountability and actionable transparency

Explainability methods (local explanations, feature importance) face fidelity/usability trade-offs: explanations that are simple for stakeholders are often unfaithful to the model; faithful explanations are often too complex. This reduces utility of XAI for auditing bias or satisfying legal notice requirements. Also, corporate secrecy and IP concerns limit audit access.

v. Scalability & effectiveness of mitigation methods in modern large models (LLMs / multimodal systems)

Many mitigation techniques were developed for tabular/classification settings. Applying them to large pretrained models, LLMs, or graph models is nontrivial: interventions may degrade performance, be incompatible with fine-tuning/transfer learning, or lead to unexpected failure modes.

vi. Intersectionality & subgroup fairness

Most methods optimize for coarse protected attributes (e.g., male/female). But harms often occur at intersections (race × gender × age). Detecting and protecting fine-grained subgroups without exploding sample complexity remains a major challenge.

vii. Causality vs correlation: need for causal methods

Statistical fairness criteria are correlation-based and can miss causal pathways that create harm. Causal approaches offer promise (e.g., counterfactual fairness) but require strong, often untestable assumptions and good causal models hard in practice.

viii. Regulation, standards, and operational compliance gaps Emerging regulations

(e.g., EU AI Act) raise requirements for risk assessments, documentation, and non-discrimination, but operationalizing legal concepts into technical checks is unresolved. Enforcement and global alignment (different jurisdictions) are open problems.

ix. Long-term, socio-technical & economic impacts

Fairness research often focuses on immediate prediction errors. Long-term dynamic effects (e.g., labor market shifts, segregation, feedback loops in policing or lending) are under-studied; methods for modelling and measuring these dynamics are immature.

x. Human factors and deployment: HCI, trust, and governance

How stakeholders (end users, impacted communities, auditors) interpret fairness claims, explanations, and remedial steps matters. Designing human-in-the-loop systems, usable audit tools, and equitable governance mechanisms remains an open research area.

10. Conclusion

This survey looked at algorithmic bias and fairness in machine learning. It covered key fairness principles, such as treating groups equally, ensuring similar outcomes for different groups, and being fair to individuals. We explored where bias comes from in data and models, and how to reduce it using different methods—before, during, or after model training. Tools like AI Fairness 360 and Fairlearn help put these methods into practice. Real-world examples, such as the COMPAS system, show that bias can still be a significant problem. Even with progress, there are trade-offs between fairness goals and challenges that depend on the specific context. New laws like the EU AI Act and local efforts like IndiBias show that fairness in AI needs a mix of technical solutions, ethical thinking, and smart policies.

REFERENCES

[1] Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, *104*, 671.

[2] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214-226).

[3] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, *29*.

[4] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

[5] 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.

[6] 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.

[7] Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., ... & Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI.

[8] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica.

[9] Bhatt, S., Dev, S., Talukdar, P., Dave, S., & Prabhakaran, V. (2022, November). Re-contextualizing fairness in NLP: The case of India. In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 727-740).

[10] Girhepuje, S., Goel, A., Krishnan, G. S., Goyal, S., Pandey, S., Kumaraguru, P., & Ravindran, B. (2023). Are models trained on Indian legal data fair?. arXiv preprint arXiv:2303.07247.

[11] Sahoo, N., Kulkarni, P., Ahmad, A., Goyal, T., Asad, N., Garimella, A., & Bhattacharyya, P. (2024, June). IndiBias: A benchmark dataset to measure social biases in language models for Indian context. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 8786-8806).

[12] Pehlivan, C. N. (2024). Report: The EU Artificial Intelligence (AI) Act: An Introduction. Global Privacy Law Review, 5(1)," 2023.

[13] Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and information systems, 33(1), 1-33.

[14] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " 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 (pp. 1135-1144)..

[15] Zhang, B. H., Lemoine, B., & Mitchell, M. (2018, December). Mitigating unwanted biases with adversarial learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 335-340).

[16] 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.

[17] Ramana, B., & Venkateswarlu, N. (2012). ILPD (Indian liver patient dataset). UCI Machine Learning Repository, 10, C5D02C.

[18] Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V.,& Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI.

[19] Dastin, J. (2022). Amazon scraps secret AI recruiting tool that showed bias against women. In Ethics of data and analytics (pp. 296-299). Auerbach Publications.

[20] Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91). PMLR.